DOI:10.16262/j.cnki.1000-8217.2012.03.003

·学科进展与展望·

闪存数据库研究进展及发展趋势

孟小峰[1] 金培权[2] 曹巍[1] 岳丽华[2]

(1 中国人民大学信息学院，北京 100872；2 中国科学技术大学计算机科学与技术学院，合肥 230027)

[摘 要] 本文简要介绍了国家自然科学基金重点项目“闪存数据库技术研究”在闪存存储管理、闪存数据库索引、闪存数据库缓冲区管理、闪存数据库查询处理、闪存数据库事务管理等方面取得的重要创新性研究成果，并对主要研究热点和发展趋势进行了分析。

[关键词] 闪存数据库，重点项目，研究成果，研究热点，发展趋势

在过去的几十年里，磁盘一直是最常用的存储介质。然而，随着嵌入式系统、航空航天等应用领域对数据存储提出了远超出磁盘能力的需求。例如，航空航天技术中探测数据十分重要，而目前由于磁盘存在抗震存储管理能力差、存取性能低、体积功耗大等缺点而使其不能直接用作航天器的存储平台，使我国航空航天技术的发展受到了极大的限制。此外，磁盘存储的机械局限性也日益突出。由于磁盘的机械寻道特性，磁盘的 I/O 速度很难继续提高，低速的磁盘与高速的 CPU、总线之间的不协调问题已越来越严重(在过去的 20 年间，CPU 处理速度增加 570 倍，而磁盘的速度却只增加了 20 倍)，从而极大地影响了计算机技术在国民经济发展中的进一步应用。因此，研究新型的数据存储介质已成为目前我国国民经济和社会发展中的重大需求。

闪存存储作为新一代的存储介质，存取速度快、耗电量小、存储内容非易失、小巧轻便等是取代磁盘作为计算机系统主要数据存储介质的首选。但现有的数据管理软件都是在传统的磁盘数据存储上针对磁盘的物理特性进行设计和优化的，直接在闪存存储上应用这些软件无法充分发挥闪存存储的优越性、满足数据管理的需求。因此有必要研究在闪存存储上的数据管理技术，包括：数据存储、索引设计、缓冲区管理、查询处理、事务管理等数据库管理系统的核心技术；通过在数据库管理系统软件层面研究充分发挥利用底层存储硬件优势的技术，探索新存储器件等新硬件的发展对现代计算机软硬件体系的推动和影响。

1 项目概况

国家自然科学基金委员会(以下简称自然科学基金委)从国家经济和科技发展的需要出发，以新兴的闪存技术为背景，针对基于闪存的数据管理问题，于 2009 年 1 月启动了国家自然科学基金重点项目“闪存数据库技术研究”(项目批准号：60833005)。该项目的目的是从闪存的器件特性入手，针对闪存应用的数据存取特点，研究闪存数据库的系统性理论和设计方法，解决闪存数据库的存储管理、缓冲区管理、查询处理、索引等关键问题，建立闪存数据库的基本理论和方法体系，为闪存数据库的进一步研究与应用奠定基础，为数据库理论和技术的进一步发展提供新思路。

该重点项目负责人由中国人民大学信息学院的孟小峰教授和中国科学技术大学计算机科学与技术学院的金培权副教授担任。项目实施过程中注重学科交叉、优势互补、开放合作，实现了课题之间的资源与信息共享。在项目执行期间，项目组探讨了一种新的组织方式，即定期举办高质量的专题研讨会，并吸收国内外学术界和工业界的知名学者参加，从而拓宽了合作领域，扩大了项目的国内外影响，也提高了项目研究成果的水平。迄今为止，项目组已经召开了 5 次研讨会。最近的一次采取了国际惯例，借 DASFAA(数据库系统高级应用国际会议) 2011 之机，组织了正式的国际学术研讨会“Workshop on

本文于 2011 年 10 月 31 日收到.

Flash-Based Databases (FlashDB 2011)”，吸引了国内外近30位研究人员参加，并由 Springer Lecture Notes in Computer Science (LNCS)出版了正式论文集。在项目执行期间，项目组不断凝炼科学问题，突出重点，在显示各自特色的同时，重视理论创新与系统验证。到目前为止，在闪存存储管理、闪存数据库索引、闪存缓冲区管理等方面均取得了较好的研究结果。

2　项目取得的主要研究进展

2.1　闪存存储管理

闪存存储管理是指针对闪存的特性提出的新的存储管理策略和访问策略，这些策略按照研究对象可以划分为针对闪存存储芯片和针对闪存设备的优化策略。

(1) 针对闪存存储芯片的优化存储管理策略。闪存存储管理有两个目标：保证高效的空间利用率和高效的数据访问性能。目前在闪存空间管理方面，闪存存储芯片级的空间管理有基于块的空间管理机制和基于页的空间管理机制。基于页的空间管理策略在空间使用、地址转换和垃圾回收上具有很好的性能，但是主存消耗比较大，并且这种消耗会随着闪存容量增大而增大。基于块的管理策略能够明显降低主存消耗，但是无法在空间利用和垃圾回收方面同时提供很好的性能。从闪存存储的访问模式而言，闪存存储芯片的物理特性意味着闪存存储更适合于随机读和顺序读写，小规模的随机写操作在闪存存储上效率很低。

项目组的工作从分析这些问题的根源入手开展了研究工作。这部分工作的典型代表有：(i) 在基于块的闪存空间管理基础上，提出了基于块集的存储管理策略，探索了在这种新的存储管理策略下，相应的页更新算法以及垃圾空间回收和地址映射等机制，进一步优化了闪存存储的空间分配、垃圾回收效率和空间利用率；(ii) 在基于页的闪存空间管理方面，为进一步提高闪存空间利用率和数据更新操作性能这一对相关的参数，提出了基于页的自适应存储管理 AFS 方法，在提高日志页空间利用率的同时，提高数据更新性能，并能适应不同的负载自动调整数据页的模式；(iii) 项目组提出的基于分离日志的闪存数据更新方法 OPL 是一种块页结合的存储管理策略，并改进缓冲区的管理，提高闪存的空间利用率和数据更新性能；(iv) 另外项目组利用闪存存储的顺序访问的优势，提出了一种高效的闪存分区存储架构，既通过分区存储提高空间利用率，又利用了顺序存取的优势保证了数据维护的效率，并通过辅助数据结构提高数据查询效率。

(2) 针对闪存设备的优化管理策略。闪存设备的特点是闪存存储芯片、内部的软硬件设计包括 FTL 等均封装起来，只能通过块设备的接口进行访问，因此针对闪存设备的存储管理策略无法通过直接操纵修改内部的存储设计实现优化的存储管理。项目组针对存储设备的特性，从改进闪存设备的昂贵的操作——写操作和擦除操作入手，提出优化的存储和操作策略。

项目组提出的 RS-Wrapper 方法的出发点是直接避免昂贵低效的随机写操作的方法，Stable Buffer 则利用更有效的闪存写操作实现优化写性能。这两项研究工作均能有效管理闪存设备的存储和访问。

2.2　闪存数据库索引

闪存数据库索引设计面临的主要问题有：如何在闪存设备上实现索引的基本功能——支持对闪存数据的快速查找和定位？如何在闪存上实现索引的更新维护功能？为了支持数据的查询，索引在数据发生改变时需要进行更新维护，这是使用索引的额外代价；由于闪存的物理访问特性，在闪存上对索引进行更新维护会带来更显著的性能代价(如异地更新，频繁擦除操作等)。因此如何针对闪存物理访问特性，设计出针对闪存优化的索引结构，更有效地支持对闪存数据的查询和更新操作性能，是项目组要解决的问题。

在闪存数据库索引设计方面，项目组的研究工作针对闪存设备设计了优化的顺序索引结构，支持嵌入式应用环境下闪存设备上的高速数据访问；在通用的闪存数据库应用环境中，项目组集中攻关了闪存索引的更新效率差的问题，采用延迟的策略进行解决，一方面是延迟更新操作，提出了基于闪存的延迟更新 B+树算法，不惜牺牲读操作的效率也要尽量减少写来提高更新索引的性能；另一方面是在多事务的环境下，采用延迟提交结点更新的策略，降低平均响应时间，提高并发度；另一个研究思路是在解决闪存存储上的索引更新性能差问题的同时，保持高效的索引查询功能，从而采用混合型索引的方法。

2.3　闪存数据库缓冲区管理

缓冲区管理是提高闪存数据库访问性能的有效手段之一。在闪存数据库缓冲区管理研究中，存在两类问题：DBMS 缓冲区管理问题和 SSD 缓冲区管

理问题。DBMS 缓冲区管理所面临的问题是：在读写代价不对称的闪存设备上，如何根据不同的负载选择优化的缓冲区页置换算法，尽可能地减少向闪存设备中的写操作次数，并提高缓冲区的命中率。在 SSD 缓冲区管理中，除了 DBMS 缓冲区管理所面临的问题之外，还需要考虑另一个问题——即不同的闪存设备之间的读写代价差异很不一致。因此通用的闪存数据库系统的缓冲区管理策略应该能够适应不同的闪存设备，在不同程度的读写代价差异的闪存设备上均能够实现优化的命中率，提高系统性能。在闪存固态硬盘内部的写缓冲区管理方面，需要解决的问题是写缓冲区的刷出(flush)操作会带来闪存页上昂贵的合并操作，因此需要优化的缓冲区置换算法来减少这一可能开销，并且在置换的过程中尽量避免耗时的写操作和擦除操作。

项目组针对这些技术难点进行了一系列研究工作，研究了闪存数据库的事务应用环境下，改进的缓冲区管理算法，区分提交和未提交事务并采用不同的优先级驻留内存，从而降低开销，提高系统总体性能；针对闪存存储的读写代价不对称的特点，提出了基于代价的自适应的缓冲区页置换策略 ACR、基于数据访问频率的置换算法 CCF-LRU 以及基于双队列的自适应置换策略 AD-LRU，解决了存在不同存取方式的负载情况下如何在高命中率和低闪存写代价之间的均衡问题；针对 SSD 设备间读写代价不对称及其巨大差异，提出了面向 SSD 的自适应缓冲区管理算法 FClock，提高闪存数据库的自适应缓冲区管理算法在不同 SSD 设备上的通用性并优化系统综合性能。在闪存固态硬盘内部的缓冲区算法方面，项目组提出 BPCLC 缓冲区置换方法，改进闪存写缓冲区的管理机制，用部分块填充(Partial Block Padding)的方法减少随机写操作，减少闪存页的完全合并次数，从而减少页置换过程中的写次数和擦出次数，提高总体性能；项目组提出的高效缓冲区置换算法 LEAC 根据预期访问开销设计页面置换策略，可以显著降低闪存访问开销，并能较好地适应不同读写比例的访问负载。

2.4 闪存数据库查询处理

查询处理与优化是数据库系统的一项主要功能。I/O 问题是基于磁盘的传统数据库查询处理和优化的基础和核心问题。由于闪存和磁盘物理特性的不同，两者具有不同 I/O 机制，所以基于磁盘 I/O 机制而设计的查询处理和优化机制不能直接应用于闪存数据库上，需要从闪存 I/O 机制来设计全新的查询处理和优化机制。

关系数据库查询执行过程中，连接操作的方法对于查询执行的效率影响很大，特别是连接操作会产生大量的中间结果，需要写到外部存储，这对于闪存来说代价非常昂贵，闪存关系数据库要避免在闪存上的大量写操作对查询执行效率的影响。项目组提出的 Digest Join 算法，运用两阶段连接的方法发挥闪存的高效随机读的优势以降低在闪存上的连接操作的代价；项目组提出的子连接算法(Sub-Join)也是一种针对闪存关系数据库上的连接操作，通过减少对数据的读取和连接中的随机写操作来提高查询执行效率。

2.5 闪存数据库事务管理

并发控制与恢复是数据库系统的核心功能。并发控制和恢复使得数据库具有了很好的处理并发用户的能力以及保障数据 ACID 特性的能力。由于闪存和磁盘的基本特性(读写特性，闪存具有写前擦特性)的不同，导致两者具有不同 I/O 机制，所以基于磁盘 I/O 机制而设计的并发控制和恢复策略不能直接应用于闪存数据库上，需要从闪存 I/O 机制来设计全新的并发控制协议和恢复策略。

项目组的 HV-recovery 研究了闪存数据库中的恢复问题，并提出了新的适用于闪存的恢复方法。这种方法提供简单有效的恢复操作，有效地减少在恢复过程中容易出现的冗余写操作，从而大幅度减少恢复时间；同时，优化了日志结构，减少过多的日志冗余，提供高效的日志文件，减少大量垃圾数据的存在，从而提高存储设备中的空间利用率。

2.6 闪存数据库实验环境的搭建

闪存数据库技术的研究需要以闪存存储和闪存设备的物理访问特性为基础，由于实验室中和市场上先进的闪存存储和闪存设备的研制和推出需要有一定的时间延迟，并且持续不断地有新的型号、更高级的访问特性的产品问世，一个比较可行的研究方案是先在闪存模拟器上验证提出的算法的有效性，然后在实际的闪存存储和闪存设备上进行验证。因此迫切需要对闪存模拟器进行开发和研究。闪存模拟器可以对典型的闪存存储和闪存设备的物理访问特性进行模拟，并可以灵活地定制不同特性的存储设备的模拟器，达到有效验证所提出的闪存数据库关键算法的目的，这是一种节约成本、加速科研周期的方法，通过共享的方式还可以有益于闪存数据库技术这一领域的研究社区。闪存模拟器对于基于闪存的研究课题有着重要的意义。

早期的评估闪存数据管理算法的有效性和性能的模拟器多是针对特殊的使用，其他人很难使用。此外，因为实现方法的不一致，各个模拟器之间的性能很难进行比较。而目前针对 DBMS 层面的算法和由于闪存 SSD 的封装性使 SSD 内部的算法很难在现有的模拟平台上进行实验对比。鉴于此，项目组开发了灵活的闪存算法模拟器 Flash-DBSim，它可以模拟各种闪存设备，方便为提出的算法在各种不同的闪存存储上进行比较，是一款灵活简单易用的闪存模拟器；项目组研制和开发的软硬件集成的闪存数据管理实验平台 FEP，实现了对闪存设备（如 SSD）内部硬件和软件的模拟，为 SSD 内部算法和 DBMS 层面算法的验证提供了支持。

2.7　项目成果总结

本项目到目前为止已经执行过半，经过项目组成员的合力攻关，在学术论文、专利、国际交流以及人才培养上都取得了较好的成果。项目组在 *IEEE Trans, On Knowledge & Data Engineering* (*TKDE*), *Data & Knowledge Engineering* (*DKE*), *SIGMOD Record* 等期刊和 SIGIR, CIKM, DASFAA 等国际会议上发表论文 30 多篇，申请专利 8 项（已授权 2 项）。项目组织合作方研究人员参加的高质量闪存数据库技术专题研讨会 4 届，组织国际学术会议级别的闪存数据库专题研讨会 1 次。项目组成员参加学术会议及邀请学者来访讲学共计 30 余次。此外，项目组与百度、华为等知名企业开展了合作，以期研究结果能解决企业实际应用中所遇到的问题。

培养高素质创新型科研人才始终是以高校为主的科研单位的首要目标。实施该项目以来，随着研究工作的推进和深入，共培养硕士研究生 15 人，其中 9 人已获得硕士学位，博士研究生 10 人，其中 5 人已获得博士学位。

3　未来的研究趋势

项目组于 2011 年 4 月在香港组织了第一届闪存数据库技术国际研讨会（FlashDB 2011）。该研讨会专门设立了一个 Session 讨论今后闪存数据库技术的研究趋势和热点。针对这一问题，与会专家展开了热烈的讨论。大家一致认为，未来闪存数据库技术研究的主要趋势为以下几个方面。

3.1　面向企业级应用的闪存数据库技术

虽然目前 SSD 已经在企业级应用中被普遍使用，但是已有的研究基本针对如何利用 SSD 提高应用性能这一问题，而对于如何提高 SSD 的生存能力、如何在应用存储架构中合理地安排 SSD、如何利用 SSD 来提高系统的能耗有效性等问题还缺乏深入研究。从项目组与国内外企业的合作交流结果来看，这些问题都是目前企业级应用所关注的重要问题，在未来研究中应加以重视。

3.2　基于闪存的混合存储技术

虽然闪存以及 SSD 已经成为一种重要的存储选择，但是从目前发展趋势来看，闪存还不能完全地取代磁盘，未来的趋势很有可能是闪存、磁盘、RAM 以及其他存储介质共存，因此研究基于闪存的混合存储技术，包括其存储架构、缓冲区的设计、查询处理算法等将是未来的研究热点。

3.3　新型存储技术对闪存数据管理的影响

近年来，相变存储器（Phase Change Memory, PCM）等新型存储技术的发展对数据存储领域带来了新的问题和启示。这些新型存储介质对于基于闪存的数据库系统有何可用之处以及限制，这一问题也是未来值得研究的一个方向。

RESEARCH ADVANCES AND FUTURE TRENDS ON FLASH-BASED DATABASE SYSTEMS

Meng Xiaofeng[1]　Jin Peiquan[2]　Cao Wei[1]　Yue Lihua[2]

(1 *School of Information, Renmin University of China, Beijing* 100872;
2 *School of Computer Science & Technology, University of Science and Technology of China, Hefei* 230027)

Abstract　This report briefly introduces the research advances of the NSFC Key Project titled "Flash-based Database Technologies". In particular, we present some innovative research achievements in the project, including storage management, indexing techniques, buffer management, query processing, and transaction management in flash-based DBMS. In addition, the future trends and hot topics in flash-based databases research are discussed.

Key words　flash-based database, key project, research achievements, hot topics, future trends