闪存数据库技术研究进展报告

——The Forth Workshop on Flash-based Database Systems

1 引言

闪存产品的性价比在不断地提高，使得闪存的应用越来越广泛，尤其是作为大数据集的替代磁盘的存储设备的应用越来越受关注。对于数据库研究者来说，建立高效的基于闪存的数据库管理系统成为数据库研究者亟待解决的问题。

以孟小峰教授为负责人的课题组（获国家自然科学基金重点项目 “闪存数据库技术研究”的资助，项目号：60833005）于 2010 年 7 月 28 号在中国北京中国人民大学召开了第四届专题研讨会。会议主要讨论了基于闪存存储器的数据库的存储管理、缓冲区管理、事务处理、查询处理、闪存存储板设计、闪存存储器在集群平台上的应用等关键问题。与会人员有来自于中国人民大学的孟小峰教授、杨楠副教授、单智勇讲师、中国科技大学的岳丽华教授、金培权副教授、以及香港浸会大学三所高校相关的硕士博士研究生，同时还有幸邀请到了百度刘斌等高级工程师和北京大学崔斌教授和他的学生们。会议包含以下几个报告：操作系统和数据库缓冲区管理算法、数据库外排序算法、闪存存储板和闪存芯片测试、数据库事务处理和 TPCC 测试结果。这些报告展示了最新的研究进展和技术成果，为基于闪存存储器的数据库的进一步研究与应用奠定基础，为基于闪存存储器的数据库理论和技术的进一步发展提供新思路。

2 闪存数据库技术研究

课题组负责人、中国人民大学孟小峰教授做了题为“闪存数据库技术研究”的报告，介绍了基于闪存存储器的数据库研究进展和新存储介质 **SCM**。闪存数据库系统研究主要以 NAND 型闪存为基础展开研究，目前已经在 EDBT、MDM、ACM SIGSPATIAL GIS、CIKM、WIAM、WISA、SAC、NDBC 和一些期刊上发表了数十篇文章，主要内容包括缓冲区中数据调度的问题研究、查询处理中的外排问题研究、闪存存储板的设计和实现、基于闪存的索引的改进和实现、基于闪存的数据库恢复方法的探索、闪存模拟器的设计与实现以及基于闪存、闪存模拟器和 SSD 的一些评测结果。通过对实验室结果、国际研究进展和现有系统的分析，找到适合闪存的数据库技术并实施在闪存数据库系统中，提高闪存数据库系统的性能。

SCM 存储器是一类存储器，性能介于内存和磁盘之间，全电设备无机械延迟，非易失性存储介质，具有很快的访问速度，并且成本很低。PCM 就是一种 SCM 存储器，相比于闪存和内存具有很强大的优势，所以目前越来越多的人关注这种新的存储介质，对于 PCM 的关注包括，把 PCM 放在整个体系结构中什么位置——使用 PCM 作为内存、或者扩展内存、或者页设备、或者取代磁盘，因为 PCM 的成本越来越低并且 PCM 性能比较好，所以这几种情况都有可能存在。

3 缓冲区管理

硬盘调度算法对数据库应用系统的整体性能具有关键性的影响。本次研讨会关于缓冲区管理策略主要有以下四个报告组成。

传统的硬盘调度算法是基于磁盘设计的，也就是电梯调度算法。这种算法的主要目的是尽量减少磁盘的磁头移动，因为磁头移动是机械运动，耗时很长。但是，面对 SSD 固态硬盘时，显得难以发挥 SSD 的内部优势，就好比是雇用一位电梯司机去指挥一个复杂的管弦乐团。在这种情况下，中国人民大学的单智勇做了题为“**A SSD I/O Scheduler for Database Systems**”的报告，该报告提出了一种新硬盘调度算法 **GP-Deadline**。设计 SSD 调度算法的七条原则：并发，无饥饿，写聚合，读聚合，读写分离，读优先和对齐块边界。在这七条原则的基础上，提出一种新的 SSD 调度算法。该算法引入 deadline 机制，以适应数据库应用系统的专门需求。该算法充分利用 SSD 的特性来发挥调度策略的作用。在后面的研究中，我们将进一步验证这七个原则的有效性。然后，在 Linux 内核实现这个新的调度算法，并且在数据库应用的负载下进行测试。

验证基于闪存的缓冲区置换算法性能的一种有效的测试手段，是在实际的 DBMS 中实现该算法，并使用测试工具（例如各种 benchmark）进行测试。PostgreSQL 是当前世界上最先进，功能最强大的自由数据库管理系统，具有良好的扩展性，适合在其中实现缓冲区置换算法，测试算法性能。为完成这项工作，需要对 PostgreSQL 缓冲区管理部分进行研究。在这种情况下，中国科学技术大学的陈恺萌做了题为“**Extending a Flash-aware Buffer Replacement Algorithm on PostgreSQL**”的报告，该报告介绍了在 **PostgreSQL** 中替换缓冲区置换算法的工作方案。在了解 PostgreSQL 的缓冲区管理后，以 CCF-LRU 算法为例介绍如何替换缓冲区置换算法。要在 PostgreSQL 中实现 CCFLRU 算法，主要是解决两个问题：现有的数据结构无法支持双 LRU 队列；在数据库工作过程中需要对两条 LRU 链表进行维护。在算法实现后，本文利用 TPC-C 测试工具 Sysbench 以及利用 ODBC 自行编写测试程序这两种测试手段，对改造前后的 PostgreSQL 进行了对比测试。实验结果表明，使用该改造方案实现的 CCFLRU 算法能够在 PostgreSQL 中正常工作，但想要证实两种算法实际运行性能的优劣，还需要进一步的测试工作。

现代计算机系统常常借助缓冲区来提升整体性能，缓冲区管理方法也成为计算机领域的热点研究问题之一。随着闪存得到越来越广泛的应用，闪存新特性对传统的缓冲区管理方法提出了新的挑战。传统针对磁盘的缓冲区管理办法专注于提升缓冲区的命中率，没有考虑到闪存独特的读写不均衡的特性，致使在装备了闪存的系统上不能发挥出缓冲区的最大性能。为了更好地在闪存系统上发挥出缓冲区的性能，北京大学的吕雁飞做了题为“**Operation Aware Buffer Management in Flash based Systems**”的报告，该报告讨论了可以感知操作的缓冲区管理方法。ACAR 的替换策略，即自适应代价感知缓冲区替换策略。针对不同的操作命中，ACAR 调整方式不同，这样就考虑了操作的差异的信息，更好地实现缓冲区管理。ACAR 的实验结果，表明 ACAR 比现有的方法更能有效地在闪存上的系统上管理缓冲区。

为了进一步利用操作的特性，报告还提出了两种估价操作频繁度的指标，分别是操作的近度(recency)和间隔距离。两种估价的指标各有侧重。报告结合这两种指标提出权重公式来决定替换出缓冲区的页面。但是不足之处是这种方法时间复杂度过高，报告最后提出如何进一步提高这种方法的速度是将来可能的方向之一。

在已有的基于闪存的缓冲区管理算法中，一般假设闪存的读操作的代价是远远小于写操作的代价，因此，已有的算法一般采用减少写操作的方法来提高系统的性能，但是对于不同的 SSD 来说，其读写性能的不对称性有着很大的差异，如果仅考虑减少写操作的次数，将仅对读写代价差异大的 SSD 适用。在这种情况下，中国人民大学的汤显作了“**ACR: an Adaptive Cost-Aware Buffer Replacement Algorithm for Flash Storage Devices**”的报告，该报告提出了一种适用于各种闪存的基于代价的缓冲区管理策略—**ACR**。提出了一种新的自适应缓冲区置换策略ð ACR，此策略使用的是基于代价进行置换的策略，因此能够适用于不同的 SSD；再者，此策略将权值与整个队列相关联，使得内存操作的代价将为 O(1)；最后，此策略能够比较好的适应长序列和循环访问模式。为了验证 ACR 的性能，我们在仿真平台上进行了实验，对比了 LRU、CF-LRU 以及 CFDC 的性能。最后的实验结果显示，ACR 算法显著地提高了缓冲区置换算法的性能。

4 查询处理

外排序是 DBMS 中最基本的算法之一。数据库中很多操作都是以其为基础，因此，外排序对数据库的整体性能有很大影响。传统的外排序算法是基于磁盘而设计的，没有结合闪存的读写不均衡特性，直接移植到闪存上时，性能未能得到优化。

目前，闪存上有关外排序算法的优化研究比较少。有鉴于此，香港浸会大学的高岫做了题为“**SSDSort: A New flash-aware external sorting algorithm**”的报告，该报告提出了一种新的适于闪存特点的外排序算法：**SSDSort**。传统的外排序算法是根据归并排序而设计的。所有需要排序的数据，需要经过排序与合并的过程。在闪存环境下，写的代价相对较高。如果对于所有的页面都进行排序与合并，将会使性能大大下降。尤其对于将近排好序或者含有数据分布较集中页面的数据，写出此类页面对性能有很大影响，没能发挥闪存快速读取的特性。针对以上问题及闪存的物理特性，本文提出了一种新的外排序算法:SSDSort。为了验证 SSDSort 的性能，论文在仿真平台中进行了实验，对比了 SSDSort 与传统归并排序所需要写出的数据量。最后的实验结果显示，SSDSort 能显著的减少中间结果的写出量，提高数据库的排序性能。

5 闪存存储

闪存存储板是闪存数据库的物理基础，它的性能对闪存数据库系统的整体性能有着很大的影响。目前闪存相关的算法研究发展很快，包括 DBMS 层面的算法和 SSD 内部的控制算法。因为商用 SSD 的内部算法被厂家固化，研究者难以在商用 SSD 上进行相关研究，所以开发一个可定置的闪存存储板很有必要。先前开发的闪存存储板在测试中的表现和理论计算

误差很大，所以必须找出第一块闪存存储板的问题并作出改变。在这种情况下，中国科学技术大学的杨濮源同学做了题为**“An Improved Flash Storage Board”**的报告，该报告指出了第一块闪存存储板设计中的问题并提出了改进方案。第一块闪存存储板的一个页的读写速度均比理论值低，而且该板的写速度比读速度快，这是和闪存的 IO 特性相违背的。导致这种现象的原因是和 PC 平台 PCI 接口相关的。在这种情况下，采用 DMA 的接口工作方式成为很好的选择。在比较了各种 DMA 模式的接口芯片的优劣后，改进的闪存存储板采用了 PEX8311 作为接口芯片。同时，在改进板的设计中，地址扩展方式由原来的位扩展改成了字扩展，目的是为了在改进板上测试多通道设计的效果，以便为以后进一步的设计积累资料。目前改进板已经制作完成，正处在调试阶段，调试工作完成后，可以很快获得测试数据。

作为一种新型的存储设备，闪存具有与传统磁盘完全不同的物理特性和更加复杂的工作原理，因此为了充分利用闪存自身的特性构建更加高效的系统，对这些特性的了解和把握是必不可少的。但实际上各个闪存芯片的生产厂家出于各种商业原因，对其所生产的芯片特性要么守口如瓶要么只是给出一些保守模糊的数据，这无疑给基于闪存之上的系统设计加上了桎梏，因此针对闪存芯片自身特性的测试就变得至关重要。在这种情况下，中国人民大学的梁智超做了题为**“Hush…tell you something novel about flash memory !”**的报告，该报告结合加州大学圣地亚哥分校非易失性系统实验室所作的工作介绍了一些针对闪存芯片的最新测试结果。在这项工作中，他们对来自五个不同生产厂家的闪存芯片进行了测试，这些闪存芯片的容量不尽相同，制作工艺包括从 50nm 到 70nm 的 SLC 型和 MLC 型。测试目的主要是从性能、耗电量以及可靠性三个方面对闪存的已知特性进行验证，对未知的特性进行探索和揭示。通过对测试结果的分析，一些有趣的规律得到了呈现。针对这些测试结果得来的特性，该实验室的研究人员提出了一种新的 FTL 算法和针对闪存的数据编码方式，实验结果表明两者在性能、耗电量和闪存使用寿命上等方面能够起到积极的作用，这也从一个侧面说明了针对闪存特性的测试工作的重要性。

6 事务处理

事务处理作为 DBMS 中不可或缺的重要组成部分，它的性能对于数据库的整体性能有着重大的影响，而由于 SSD 与磁盘不同的读写特性，使得基于闪存的数据库系统的事务管理部分变得更加复杂，因此，如何快速有效的维护闪存数据库系统 ACID 特性就变得越来越重要。在这种情况下，中国人民大学的卢泽萍、范玉雷做了题为“**Session on Transaction of Flash-DB**”的报告，该报告介绍了本项目组最近的关于事务方面的几项工作。

PTLog 采用对页表记日志的方法，来有效的减少由于 WAL（先写日志）规则所带来的大量的闪存空间浪费，并利用闪存快速的随机读特性，通过改变日志结构，提供行之有效的恢复策略。HV-recovery 对闪存中天然存在的数据的历史版本使用新的日志结构加以管理和利用，提供高效的恢复。通过周期性设立检查点，减小无效日志记录的长度，节约闪存空间。引入混合式存储系统，将日志记录单独存放在磁盘上，以便对闪存数据库的恢复性能进一步提高。同时也保证了算法具有在数据库正常运行时有较小的开支、算法有比较强的可靠性、

系统失败后恢复速度快和日志文件的空间需求较小等优势。通过针对 TPCC 的分析及和开源数据库 Oracle Berkeley DB 的对比实验看出，HV-recovery 比传统数据库的恢复时的写操作数可以减少接近一半，其恢复时间与传统数据库相比，能缩短到原来的大约 1/8，与在 SSD 上的传统数据库相比，也可以缩短 40%，充分显示了本算法的优越性。

闪存昂贵、异位更新和有限擦除次数，但是闪存有很好的读写性能，所以目前采用固态硬盘和磁盘进行混合存储是比较好的解决方案。但是目前 OLTP 系统要求越来越高的并发度，在混合系统之上提高事务的并发成为一个关键问题。由于网络 DBMS 和用户双方对某些信息的一致性要求并不是很高，可以引入弱一致性来提高 DBMS 的并发。现有的 DBMS 事务子系统采用的并发控制协议基于串行化理论，主要分为两大类：单版本的和多版本的。单版本的并发控制协议主要有两阶段加锁协议、乐观协议、时间戳协议和有效性确认等等。多版本的并发控制协议主要有多版本加锁协议、多版本乐观协议、多版本时间戳协议和多版本有效性确认等等。这些并发控制协议使得事务执行都具有较高的一致性，这种一致性限制了事务并发度的提高。由于现在对于一致性的要求不是那么严格，数据存在天然的多版本特性，所以采用引入弱一致性的多版本的并发控制协议会增加事务的并发度。固态硬盘内部封装了多块闪存芯片，对外通过软件模拟成为块设备，但是这些闪存芯片之间存在着可以并发操作的现象，可以利用闪存芯片之间的并发机制来提高事物的并发度。利用混合式系统存储数据——读多的数据放在固态硬盘上，写多的数据放在磁盘上——大部分读操作发送到固态硬盘上执行，大部分写操作发送到磁盘上执行。对于固态硬盘的读操作可以不进行一致性控制，让更多的读操作尽量并发执行，使用闪存芯片之间的并发控制。由于不进行严格一致性控制，使得数据存在多版本的特性，所以提出了“基于弱一致性、芯片级并发、多版本并发控制协议和混合式存储的数据管理系统”。

PG 的 **TPCC** 测试平台：借助开源代码程序 BenchmarkSQL 对 PostgreSQL 数据库进行 TPCC 测试，BenchmarkSQL 源代码采用 Java 语言编写，这就使得该测试程序可以很容易移植到其它操作系统之上并能很好的运行。BenchmarkSQL 采用 DOS 命令的方式测试 PostgreSQL 的数据库，操作简洁方便。

针对磁盘单一存储系统、固态硬盘单一存储系统和混合式存储系统，采用 BenchmarkSQL 测试程序进行 PostgreSQL 的 TPCC 测试，测试结果显示固态硬盘具有良好的性能，TPCC 测试标准下的 Tpmc 值在固态硬盘上比在磁盘上提高约 10 倍，大大显示了固态硬盘的优势。

7 总结

闪存数据库中数据存储管理、索引、缓冲区管理、查询处理和事务处理等模块已经取得了很大的研究进展，对于项目以后的研究会有很大的推动作用。缓冲区管理对于提高整个闪存数据库系统的性能起着至关重要的作用，研讨会对于缓冲区管理机制进行深刻的研究，从不同角度设计基于闪存缓冲区管理策略，并把其实现到开源数据库中来检测基于闪存的缓冲区管理策略的性能。但是根据目前的研究，闪存数据库中各个模块仍然有很大的改动空间，

尤其是闪存数据库的存储管理和事务处理应该具有较大的改动空间使其更好的适应闪存的特性，使得闪存数据库的整体性能大幅度提高，这主要体现为查询速度提高、事务处理速度快和事务并发度提高等等。

同时，研讨会还就将来硬件的发展对数据库系统带来的变化进行了专门的讨论，闪存相对于磁盘仍然还具有很大的优势，但是新硬件的产生给闪存提出了巨大的挑战。

研讨会还就闪存的发展现状、发展趋势，以及对基于闪存的数据库系统研究中存在的新的挑战进行了热烈的讨论，闪存的寿命问题受到越来越多的关注，比如百度就在致力解决闪存在服务器上的使用寿命，以提高闪存在现实应用中的性价比，这对项目以后的研究会有很大的推动作用，但是仍有许多未知的领域需要我们去探讨和研究，这就需要我们投入更多的研究工作和热情。

总之，目前的闪存数据库并没有充分发挥闪存的特性，需要根据闪存的物理特性进行重新设计，以发挥闪存的优越物理性能。