闪存数据库技术研究进展

——2008 Workshop on Flash-based Database System

1、引言

闪速存储器（简称闪存）是一种高速、低功耗、抗震、小巧轻便的存储介质。闪存的高速、抗震等特性使得它成为替代磁盘从而突破上述局限性的首选存储介质。无论是在笔记本市场还是在服务器领域，都已经有闪存的身影。随着闪存容量的不断增长，越来越多的电子设备直接用其来存储大量的数据，由此带来的新问题是“如何有效地组织、存储、管理和使用闪存中的数据”。目前，数据管理技术的主流是数据库技术。因此，采用数据库技术来存储和管理闪存中的数据是目前首选的途径，即建立“闪存数据库”。

为了解决闪存数据日益严峻的管理问题，以中国人民大学孟小峰教授为组长的课题组项目于2008 年 9 月获得了国家自然科学基金重点项目（Key Project of NSFC：60833005）“闪存数据库技术研究”的资助，展开了对闪存数据库的研究。在此基础之上，课题组于 2008 年 11 月 8 号在中国合肥中国科技大学召开了“2008 Workshop on Flash-based Database”的会议，与会人员有来自于中国人民大学的孟小峰教授（项目组负责人）、杨楠副教授、中国科技大学的岳丽华教授、金培权副教授、香港浸会大学的徐建良副教授及三所高校相关的研究生。会议内容涉及到闪存数据库存储管理、缓冲区管理、索引、查询处理及优化、事务处理等关键问题，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。这次会议主要讨论了存储介质问题、数据库搭建问题和模拟仿真环境问题，涵盖了从 FTL 到数据库查询处理的闪存数据库技术研究的方方面面，展示了最新的研究进展和技术成果。

2、闪存数据库技术研究

课题负责人、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，介绍了闪存数据库研究现状及课题的总体情况。闪存一般分为 NOR 型和 NAND 型。数据库系统的研究主要以 NAND 型闪存为基础展开研究。相对于磁盘，闪存具有很多优势，比如说无机械延迟和读写特性等，但是闪存与传统磁盘的主要差异在于其读写操作具有明显的不对称性。由于闪存具有的特性和磁盘有较大差异，基于磁盘的数据库技术直接移植到闪存上存在很多问题，所以需要结合闪存自身的特性，研究针对闪存数据库的新理论和新方法。

结合理论分析、仿真试验和系统验证的研究方法，自上而下解决课题研究中的瓶颈问题。首先根据理论分析提出适合闪存的查询处理、事务处理等新机制与新方法；然后采用内存模拟仿真闪存的方式验证上面所提出的机制和方法；最后搭建闪存数据存储平台，采用大规模数据集进行性能和功能验证。

2.1、存储管理

存储管理的功能是将闪存的硬件细节进行屏蔽，为上层应用提供硬件抽象。目前闪存数据库的存储管理技术大致可分为两类：一类是采用闪存转换层（Flash Translation Layer，FTL）来实现闪存存储管理，另一类则是采用全新的闪存数据库存储管理器来实现此目的。全新的闪存数据库存储管理器所面临的问题有：空间分配策略和垃圾回收策略和磨损平衡等等。

中国科学技术大学刘沾沾做了题为“基于 Flash DBMS 的存储支持”的报告，介绍了一种 NOR 和 NAND 的混合存储结构。由于 NOR 闪存具有细粒度的操作，可以按位读/写；而 NAND 闪存具有粗粒度的操作，按页进行读/写，所以提出了一种基于 NOR 和 NAND 的混合式闪存存储结构，充分发挥这两种闪存的特性，找到合适的操作粒度，系统中 NOR 用来存储元数据和日志等等，NAND 用来存储数据和元数据快照。

中国人民大学单智勇提交了题为“基于置换页的 NAND 闪存转换层”的报告，主要介绍了一种改进的 FTL，降低垃圾回收的代价。改进的 FTL 模型将混合映射方式进行了改进，使得读写粒度发生了改变，同时根据页标识进行垃圾回收。另外，该模型提出了一种新的缓冲区中页的置换策略。

2.2、闪存和磁盘的混合式系统

磁盘是目前应用最广泛的存储介质，但是其性能提升的空间很小，而 flash 作为一种新兴的存储介质，以其独有的特性得到了迅速的发展，但是其读写差异以及价格的昂贵，使得它在近期还不能完全取代磁盘。将 flash 和磁盘组合成混合式系统，可利用两种存储介质的优点，克服其缺点，从而可在很大程度上提高系统的性能。混合式系统目前有多种，flash 在其中扮演不同的角色。混合式系统目前也面临许多问题，主要有数据放置问题，数据页的替换策略等等。

中国人民大学汤显作了题为“混合式系统”的报告，介绍了目前现有的一些混合式系统，并且详细介绍了 NAND 和 Disk 混合的模型和模型下的数据放置问题。Flash 快速随机 IO 的特性可用来增强磁盘的性能，因此，可使用 flash 和磁盘组合来提高整个系统的性能。Flash 按其在系统中所扮演的不同角色，可分为作为 cache 的 flash 和作为辅存的 flash。而处于不同混合式系统中的 flash 又有各自的特点和不同的作用。Flash 和磁盘的混合式系统是目前研究的热点，其上有着许多值得研究的问题，比方说地址映射，buffer 管理，页放置等等。

2.3、索引

索引管理是有效组织数据、提高数据库查询性能的关键技术之一。研究高效的闪存数据库索引管理方法对提高系统查询性能至关重要。在基于闪存的索引管理方面，国内外已有的大量的研究工作。在此基础上中国人民大学周大做了题为“基于 Flash 的索引”的报告，报告总结了整个闪存上的索引的发展脉络，并且结合自己的研究提出了一些新问题。首先介绍了 FTL 的发展历程，FTL 是对磁盘的运行方式的模拟，可以直接运行在 Flash 之上。由于主存的消耗、冷热数据的差别和地址映射的粒度的差别，先后又出现了 NFTL、AFTL、Super block-based FTL、FAST、Hybrid FTL、LAST、STAFF 和 LGeDBMS。索引是数据库管理系统的一个重要组成部分。基于 FTL 的索引包含动态哈希、B-树索引和 R-树索引。纯闪存索引有 hash 类索引、树型索引和特殊索引结构。

2.4、查询处理

查询处理与优化是数据库系统的一项主要功能，其中的外连接算法是查询中最重要的算法之一。闪存读写不对称性，需要尽量减少写增加读来提高外连接的速度。有效的利用闪存随机读取的性能和减少写的次数来提高闪存上外连接的效率将是闪存数据库一个重要的研究方面。针对闪存数据库，在查询优化策略方面，优化的原则是在减少 I/O 次数的时候，重点考虑减少写的次数，在适当的情况下甚至需要增加读的次数来减少写的次数。

针对闪存数据库中查询处理中的研究问题，香港浸会大学 Sai Tung On 做了题为“基于 Flash 优化连接”的报告，提出了一种新基于 Flash 的连接算法。对数据库的查询过程中，连接是个非常重要的操作。传统的非索引连接算法有嵌套循环、排序合并和哈希连接算法。在访问二级存储过程中，排序或者哈希是最主要的连接代价。若需对数据库中的表做查询操作，需要先将两个表读到内存中进行排序，如果表的大小大于内存的大小，还需要多趟读取操作才能完成两个表的 join 操作。本报告算法提出了两阶段连接算法。

2.5、事务处理

日志和恢复是保证事务的原子性、一致性和持久性，提高数据库可靠性的重要技术之一。在闪存数据库中，由于闪存以页为读写单位且一般采用换位更新策略，频繁写入日志项会引起大量多余的数据写入，影响系统性能；将日志存放在单独区域可能造成存储器中的热区，导致各块的擦除频率不均，影响闪存的使用寿命。因此，闪存数据库的日志管理和恢复需要研究新的方法。目前，在闪存数据管理领域，相关研究已经提出了一些日志记录方法，主要是日志的存储方式，较少考虑事务处理，难以达到数据库系统对事务原子性和一致性的要求。

针对闪存数据库中事务处理中的研究问题，中国人民大学向锂作了题为“基于 Flash 存储的恢复机制”的报告，阐述了日志同步 I/O 问题，并提出了解决策略。同步 I/O 普遍存在于基于日志的恢复办法中， 成为了 DBMS 的系统瓶颈。虽然有很多致力于解决这个问题的工作，但是同步 I/O 的问题都没有真正得到解决。 基于日志的恢复算法在闪存中遇到了许多新的问题。通过全面研究了传统基于磁盘的恢复算法的优缺点，以及他们在闪存中应用的各种问题，提出了一种全新的面向闪存的恢复算法——采用了通过保持事务的Propagation的原子性来达到保持事务操作的原子性的目的，在提高恢复效率的同时很好的符合了闪存的物理特性。

2.6、Flash DB 技术评价的仿真框架

中国科学技术大学苏轩作了题为“Flash DB 技术评价的仿真框架”的报告，介绍了自主实现开发的一种仿真平台，并介绍了一些仿真平台上测试结果。由于闪存的特殊物理特性，传统的数据库管理系统模式不能够直接应用在闪存数据库上。因此，需要该改进传统的数据库模式，使其适合新的硬件特性。为了评价所研究的模式，提出了一种新的仿真框架。这个仿真框架灵活、可扩展、可重用，可以减少冗余代码，支持存储管理、缓冲管理和索引的仿真。闪存仿真框架通过面向对象里面的多态来支持底层虚拟设备的选择，可以是NOR虚拟闪存设备，也可以是NAND虚拟闪存设备。针对不同的虚拟闪存设备可以选择FTL或者NFTL的算法，例如地址转换算法、空间回收算法和垃圾回收算法，可以通过选择合适的接口把待测试的代码实施到该仿真框架上。

3、总结

闪存数据库的研究目前已经取得了一定的进展，比如数存储管理、缓冲区管理、索引、事务恢复和查询的研究，对于项目以后的研究会有很大的推动作用，但是根据目前的研究，闪存的高速访问特性还没能完全发挥出来，闪存的访问性能要比磁盘的高出近百倍，但是目前从软件上提高的性能却远远达不到这种程度，这就需要更多的研究工作和研究投入。

目前对于闪存数据库的研究基本上可以分为存储管理、缓冲区管理、索引、查询处理以及优化、事务处理这几个主要部分。对于存储管理，倾向于采用闪存和磁盘的混合式存储方式，这种方式可以发挥这两种存储介质的各自优点。对于缓冲区管理，采取的措施是把磁盘上的缓冲区管理与闪存的读写特性直接结合。在闪存上建立的索引，可以分为两大类，一类是建立在 FTL 之上的索引，一类是直接建立闪存之上的索引，目前采用的索引基本上都是从磁盘上借鉴过来，而不是针对闪存而专门提出的。查询操作的两个最主要操作就是排序和连接，研究排序和连接的工作目前已经取得有一定的工作，但是这方面的工作相对有限，因为查询以及查询优化受到底层的限制。事务处理部分目前的研究也已经取得一部分进展。

总之，目前的闪存数据库并没有充分发挥闪存的高速读写特性，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。

以下是孟小峰教授在“2008 Workshop on Flash-based Database”研讨会上所作的报告。

闪存数据库技术研究进展

负　责　人：孟小峰
依 托 单 位：中国人民大学
合 作 单 位：中国科学技术大学

研究意义

- 传统的硬盘性能升级速度太慢，无法跟上随摩尔定律进化的微处理器发展速度
  - 过去10年里微处理器速度提高了约30倍，而硬盘的性能只提高了1.3倍
- 闪存具有比磁盘更好的性能
  - 无机械延迟，存取速度快
  - 低功耗、便携、抗震、低噪音
  - 应用广泛，前景广阔，已呈取代磁盘之势

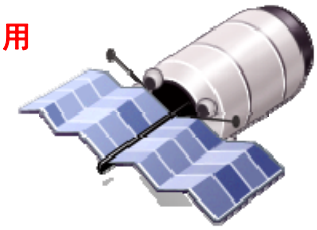
多样化的闪存应用

移动设备

个人信息管理

航天平台、空间应用

海量数据管理



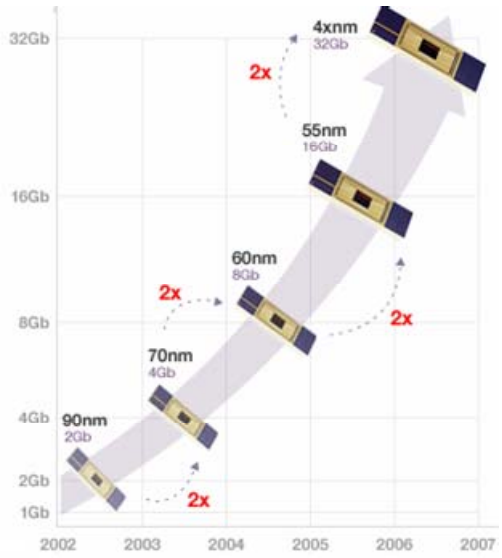
Web数据管理

野外应用

恶劣环境数据管理

研究意义

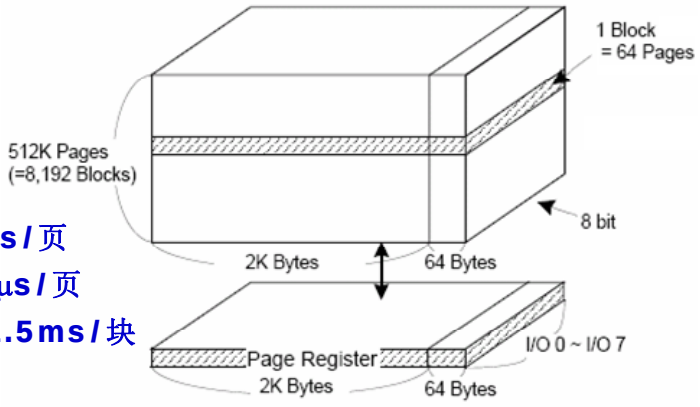
- 每年闪存芯片的容量都以2倍的速度增长
- 近年这一增长速度还在加快，目前已经存在128G的基于NAND闪存的固态硬盘
- NAND闪存 vs. NOR闪存
  - NAND闪存适合数据存储
  - NOR闪存适合程序代码存储



http://www.samsung.com/global/business/semiconductor/products/flash/Products_NANDFlash.html

闪存芯片特性

- 闪存芯片组成
  - 1 芯片 = 若干块
  - 1 块 = 64 页
- 基本操作
  - 读：按页读，25 μs/页
  - 写：按页写，200 μs/页
  - 擦除：按块擦除，1.5ms/块
- 闪存芯片硬件限制
  - 读写速度不均衡
  - 异地更新：重写页前必须先擦除其所在的块
  - 每一块允许的平均擦除次数有限



传统查询处理性能不足

- Flash存取速度提高了60～150倍 (vs. [illegible])

Media	Access time		
	Read	Write	Erase
Magnetic† Disk	12.7 ms (2 KB)	13.7 ms (2 KB)	N/A
NAND Flash‡	80 μs (2 KB)	200 μs (2 KB)	1.5 ms (128 KB)

现有数据库的查询处理不能充分发挥闪存的快速读写特性

- 但商用DBMS上的查询处理性能改进很少

Read Queries	Query processing time (sec)		Write Queries	Query processing time (sec)	
	Disk	Flash		Disk	Flash
Sequential (Q_1)	14.04	11.02	Sequential (Q_4)	34.03	26.01
Random (Q_2)	61.07	12.05	Random (Q_5)	151.92	61.76
Random (Q_3)	172.01	13.05	Random (Q_6)	340.72	369.88

Sang-Won Lee et al. Design of Flash-Based DBMS: An In-Page Logging Approach. SIGMOD'07

事务处理性能有待提高

- 商用DBMS上并发事务处理性能仅提高了2～10倍

no. of concurrent transactions	hard disk		flash SSD	
	TPS	%CPU	TPS	%CPU
4	178	2.5	2282	28
8	358	4.5	4050	47
16	711	8.5	6274	
32	1403	20	5953	
64	2737	38	57[illegible]	

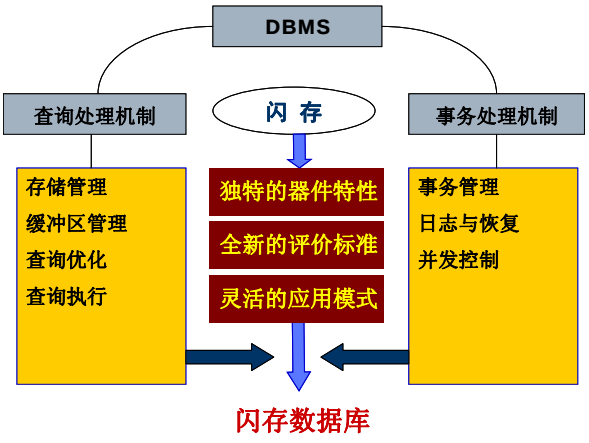
TPS: Transactions-per-Second

现有数据库的事务处理性能与闪存高速存取特性不匹配

Sang-Won Lee et al. A Case for Flash Memory SSD in Enterprise Database Applications. SIGMOD'08

课题的提出

- 针对闪存硬件特性、灵活的应用模式和传统数据库技术的不足，研究全新的闪存数据库管理技术
- 即基于闪存的数据库管理系统（Flash-Based DBMS）



闪存数据库研究现状

- 商用数据库如**Sybase iAnywhere，Oracle Database Lite 10g**等通过闪存转换层**(FTL)**模拟磁盘以支持闪存，但不能充分体现闪存的读写性能
- **FlashDB**采用闪存转换层支持闪存存储，提出了自适应索引，但同样存在FTL性能问题，也没有考虑恢复、并发等问题
  ——*Suman Nath et al. 2007*
- **LGeDBMS**直接管理**NAND**型闪存，并提供了简单的查询处理、日志和恢复功能，但仅支持移动设备，数据量小，结构和功能简单
  ——*Gye-Jeong Kim et al. 2006*
- **StonesDB**支持闪存数据的索引、查询及数据挖掘，但系统仅针对传感器网络中节点数据管理，不支持数据库事务处理等高级功能
  ——*Yanlei Diao et al. 2007*

闪存数据库研究现状

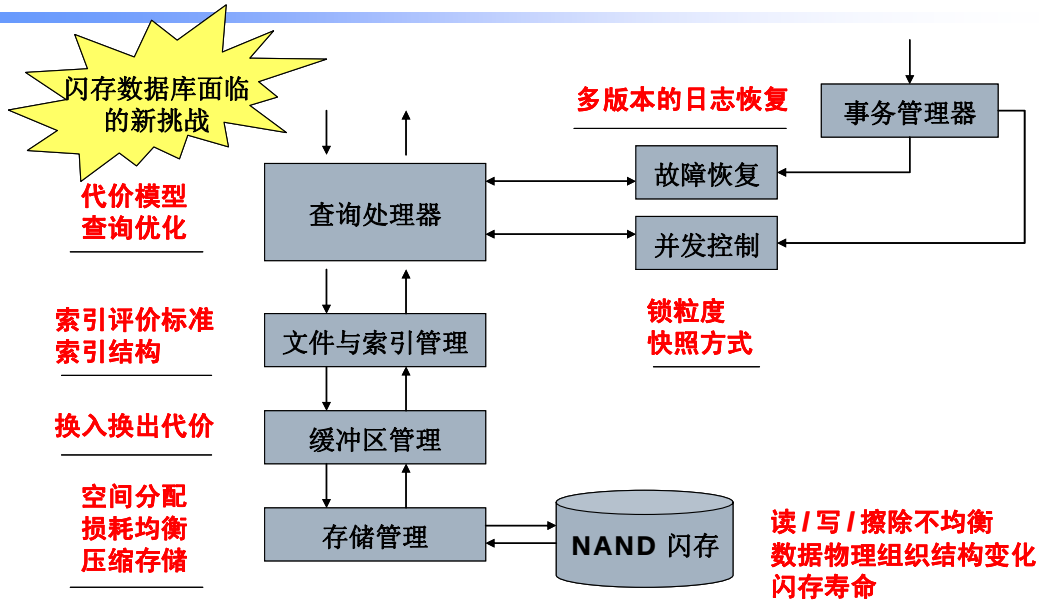
- 闪存数据管理尚停留在文件管理阶段，数据库层面的研究少
- 专用的闪存数据库系统多，通用的少
- 对于数据库管理系统如何适应闪存器件特性的问题尚未明确阐明
- 点式研究多，系统研究少，对于闪存与数据库存储管理、索引、缓冲区、查询优化等部件之间的关系尚停留在孤立分析阶段

系统性的闪存数据库研究亟待深入

项目的研究目标

- 建立适用性强、可剪裁的闪存数据库体系结构，从理论上探索闪存数据库的系统架构和标准
- 提出针对闪存特性的若干数据库新理论和新方法，解决闪存数据库中存储管理、缓冲区管理、索引、查询优化、事务处理等关键技术，为闪存数据库的大规模应用奠定基础

项目的主要研究内容



项目的主要研究内容

- 可剪裁的闪存数据库体系结构
  - 内核模块的分解与描述
  - 内核模块的封装与持久化
  - 内核剪裁模型与剪裁规则
- 闪存存储管理技术
  - 空间分配与垃圾回收策略
  - 磨损平衡策略
  - 多芯片存储结构
  - 快速启动与掉电保护机制
- 闪存数据库缓冲区管理技术
  - 顾及读写代价差异的缓冲区置换策略
  - 缓冲区控制机制

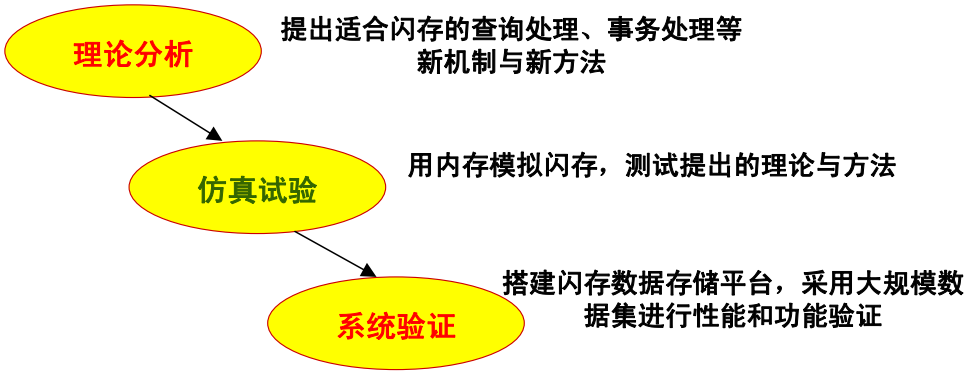
项目的主要研究内容

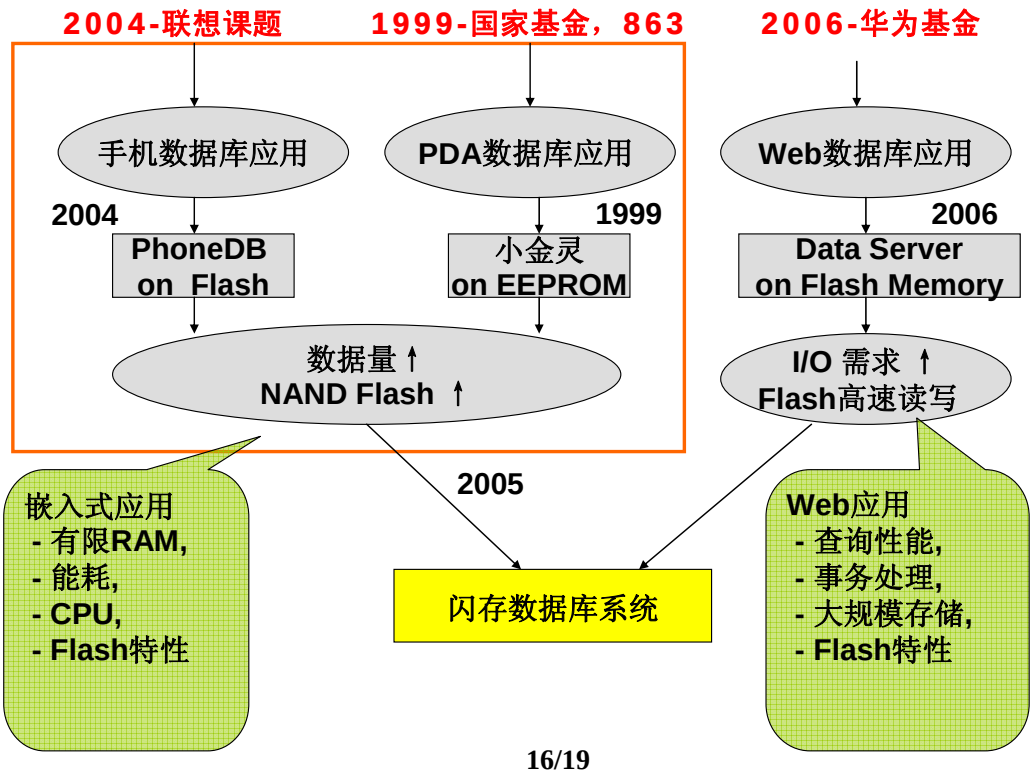
- 闪存数据库索引技术
  - 基于闪存的B+树，Hash索引
  - 基于闪存的新的索引方法
- 闪存数据库查询处理机制
  - 闪存数据库查询代价模型
  - 闪存数据库查询算法的高效实现
  - 查询优化策略
- 闪存数据库事务处理机制
  - 快照隔离的并发控制方法
  - 基于闪存的多粒度封锁机制
  - 多版本的日志恢复方法
- 闪存性能评价模型
  - I/O评价方法
  - 磨损平衡评价方法
  - 垃圾数据的估算方法

项目研究方案

研究方法

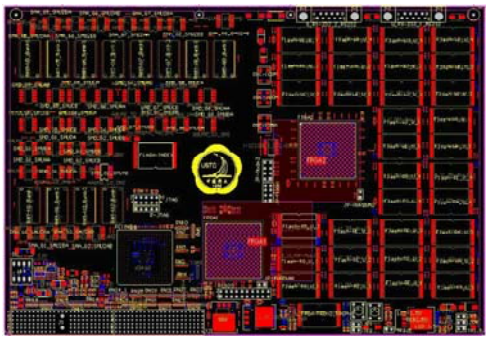
- 结合理论分析、仿真试验和系统验证的研究方法，自上而下解决课题研究中的瓶颈问题





研究工作基础

2006-高速大容量闪存存储平台



目前已经达到单板128GB，共64片，每片2GB

闪存技术的新进展和有关实验

- Nand:SAMSUNG,K9WAG08U1A
- Nor ：Spansion, S29WS-R

希望大家批评指正！

谢　谢！

Tape is Dead
Disk is Tape
Flash is Disk

Jim Gray
Microsoft
December 2006