国家重点基础研究发展规划（973）项目

面向复杂应用环境的数据存储系统理论与技术基础研究

（项目编号：2011CB302300）

简　　报

[2013] 01 号　总第 8 期

项目办公室　编　　**2013 年 07 月 05 日**

项目研究进展

国家 973 计划项目“面向复杂应用环境的数据存储系统理论与技术基础研究”自 2012 年 8 月份中期检查以来，各课题根据项目总体任务、目标和调整方案，围绕面向服务的异构融合存储体系和复杂应用环境下的泛在存储服务这两个科学问题推进研究工作。截止 2013 年 6 月，项目研究成果在包括 INFOCOM、CIKM、MSST 等本领域国际顶级会议以及 IEEE TC、IEEE TPDS、ACM ToS 等国际权威期刊上共计发表论文 83 篇，申请发明专利共 45 项，获授权专利 21 项，申请软件著作权 8 项，培养博士 25 人、硕士 88 人、博士后 4 人，1 人次获得 2013 年度“中青年科技创新领军人才”称号。

课题一：融合存储体系结构与服务架构研究

课题从融合存储系统的组织、面向复杂应用环境的按需服务以及构建示范三个层面开展了融合存储系统与面向复杂应用的按需服务等关键技术的研究和开发工作。发表了 23 篇论文，其中在中国计算机学会推荐的 A 类期刊和会议上发表论文 5 篇（IEEE TC 3 篇、TPDS 1 篇、INFOCOM 1 篇），B 类论文 3 篇（ACM TACO 1 篇、MSST 1 篇、CIKM 1 篇）。申请发明专

利 14 项，授权发明专利 4 项；培养博士 5 人，硕士生 22 人，1 人次获得 2013 年度“中青年科技创新领军人才”称号。

对于融合存储体系结构及关键技术研究，除了按照原有任务计划继续推进融合存储软硬件平台的设计实现工作外，根据中期检查专家有关加强新型存储器件关注和新型存储系统结构研究的建议，安排部署了基于 DRAM 和新型 NVRAM 的内存扩展研究以及基于 PCM 的文件系统研究工作。对于复杂应用下的存储服务研究，根据中期检查专家关于加强对于大数据应用产生存储需求的分析，来开展融合存储系统按需服务的研究建议，课题针对大数据应用场景下对于安全性的需求，开展数据安全删除、溯源重建等研究。主要研究进展包括：

(1)构建完成能够支持 Flash/PCM 等多种异构存储介质的融合存储平台。实现了 FMC（FPGA Mezzanine Card）接口和 DDR 接口两套子卡，完成了 NAND Flash 控制器核 PCM 控制器功能性验证工作；实现设备端的关键软件，包括支持 NVMe 协议的固件层软件、FTL、中断、DMA 等底层驱动软件。后续将安排进行软硬件整体联调测试工作。

(2) 基于元数据关联特征的数据组织与近似查询技术。提出 NEST 新结构，进行数据的语义特征和行为模式的挖掘、获取和快速分类，使用增强 LSH 来存储相关数据在一个哈希桶，并利用 cuckoo 散列来实现负载平衡，显著降低重新哈希的概率，提供易于使用的、高性价比的近似查询服务。相比传统 O(n)复杂度的垂直寻址结构（例如链表 LSH），NEST 复杂度 O(1)。提出了 MERCURY 数据分析方法，在时间、空间和内容三个维度上挖掘数据的相似性特征，进行数据的聚集。为解决存储空间利用率低和同构数据放置问题，设计了一种新型的面向多核处理器的 MC-LSH 方法，以准确获取数据的相似性特征，这种数据相关关系用于缓存优化和数据布局，最大限度地减少缓存污染和数据迁移。利用分析所得数据相关关系，可进一步优化面向数据立方体的海量数据分析，提出一种可扩展的分布式结构 ANTELOPE，通过分析用户访问模式，判定相关数据集合，对预计算的数据立方体进行部分物化，显著降低预计算在时间和空间上的开销。

(3) 语义感知的存储系统命名机制。数据量快速增长和数据处理高复杂性影响文件系统运行性能。当前文件系统分层目录树结构的命名空间容易产生严重性能瓶颈，不能提供实时响应，不能支持复杂数据查询，需要新命名空间管理来提供易用性和高数据访问效率。提出了一种语义感知命名空间 SANE，为超大规模存储系统提供动态和自适应命名空间和存储服务。SANE 语义感知每个文件的命名空间，利用语义概念和文件间相互关系，动态地关联相关的文件到规模较小且水平的组内，以实现快速、准确的查询服务。SANE 作为中间件，可用在常规文件系统和垂直分层目录树下。SANE 的语义关联和文件组也可用于文件预取和数据去重等，实现存储系统的性能优化。

(4) 存储服务的性能隔离与性能保障技术。为云计算中心提供云存储服务的数据中心需要提供一种简单且鲁棒的性能隔离和保障服务，即对用户租赁的虚拟机提供简单的存储性能定制接口和强大的存储性能保障。基于对云存储环境下多级 IO 调度的研究，提出一种分布式调度方法避免集中式调度所面临的整体性公平和客户性能保障两难问题，提升存储性能隔离的水平；同时采用细粒度的性能调节技术有效控制存储带宽的波动，减轻并行虚拟机间的性能干扰。高度聚合的存储系统导致多个应用间存在对存储资源的竞争。提出一种可调预期算法 ASPM，按应用对缓存动态分区方法，不同应用的缓存区互相隔离；并针对负载动态变化，通过分析预取长度对每一应用的影响效果，动态分配应用的预取长度，与 LRU 和 AMP 相比，提升了系统整体性能。

(5) 复杂应用环境下的数据服务安全保障技术。提出一种基于 FLASH SSD 的用户数据安全删除方法，使用分区识别功能区分不同硬盘分区，使用户只对存储敏感文件的硬盘分区进行安全删除，让用户像使用磁介质硬盘一样，方便地对需要的文件进行安全删除，不影响计算机正常使用。同时采用 RAID5 数据冗余算法，对所有数据进行保护，保证在单个物理页、物理块、芯片损坏时，相关数据可恢复。提出一种基于溯源信息的数据重建方法，系统故障时，溯源重建的性能显著优于日志恢复的性能，与快照恢

复性能相当，但溯源重建能精确的将损坏或丢失的文件恢复至文件完好时的最新版本。

(6) 上线部署了“晒点”系统（http://paper.ustor.cn/）数据共享服务平台、广域网存储服务应用 U-Stor（www.ustor.cn（电信网）和 ustor.hust.edu.cn（教育网））以及 HUSTBackup 数据备份服务系统 。

课题二：海量数据组织与资源共享的方法研究

2013 年上半年研究主要集中在海量非结构化数据的处理编程模型及其海量数据的挖掘分析与计算方法等方面的关键技术研究。经过半年的研究，取得了一些成果。在软件学报、物理学报、计算机科学、COMPSAC 2012 和 SSS 2012 等国际期刊与会议上录用或发表了 EI 国际检索论文 14 篇，SCI 国际检索论文 2 篇；申请发明专利 2 项；进一步完善了海量结构化数据处理平台华鼎-C，进一步完善了海量非结构化数据存储管理平台华鼎-U。

(1) 设计并实现了基于行列混合存储的关系数据库。针对传统关系数据库存在的问题，结合列数据库的技术特点，提出一种基于数据块的行列混合存储模型，在一个存储引擎中同时应用行存储和列存储两种存储模型，并利用工作负荷的变化对行列混合存储进行推荐优化和自适应转换，实现一个整体的、一致的、全面的数据库管理系统，使其能够更好地同时支持多种类型的应用需求，降低企业的数据管理和维护成本。基于开源关系数据库 Axion，通过修改存储引擎，实现了一个基于行列混合存储的关系数据库 AxionMixDB。最后利用 CBTR 的测试数据集对行列混合存储的关系数据库 AxionMixDB 与开源关系数据库 Axion 进行对比实验，主要包括功能检测和性能评估。

(2)提出一种新的网络传播中最有影响力节点发现方法。在复杂网络的传播模型研究中，如何发现最具影响力的传播节点在理论和现实应用中都有重大的意义。目前的研究一般使用节点的度数、紧密度、介数和 K-shell 等中心化指标来评价影响力，这种方法虽然简单，但是由于它们仅利用了节点自身的内部属性，因而在评价影响力时精确度并不高。为了解决这个

问题，提出了 KSC(K-shell and Community Centrality)指标模型。此模型不但考虑了节点的内部属性，而且还综合考虑了节点的外部属性，例如节点所属的社区等。

(3)研究并实现了基于 Paxos 算法的低延迟容错复制技术：针对具有低延迟、高可用性要求的联机事务处理应用，首先提出了一个通用的低延迟容错复制框架。该框架基于状态机复制技术，并融合了联机事务处理系统中的数据驻留内存、单线程处理、冗余备份提供高可用性等设计趋势，框架的核心是一个基于 Paxos 算法的低延迟容错复制协议。为实现该复制协议，首先给出了一个消除协议状态持久化的 Paxos 算法变种。该变种在普通 Paxos 算法的基础上进行了两个主要优化以保证消除持久化后算法的正确性：一是通过引入向量时钟机制以判断消息发送与结点崩溃之间的时序关系，二是增加一个 Pre-Prepare 阶段将提案编号发送到超过半数的结点上以保证崩溃结点的正确恢复。该变种由于无需持久化而能够大大降低延迟。在该算法变种基础上，我们扩展实现了无协议状态持久化(NPSP)的 Multi-Paxos 容错复制协议并对其容错性进行了分析，协议中对于主结点选举、进度追赶、崩溃结点的恢复、更新操作的异步持久化等实现问题给出了解决方案。

(4)研究了文档全生命周期管理中的区间持久性 top-k 查询技术，提出了一种新的算法即双哈希表关联算法（简称 DHA）。DHA 算法利用哈希表快速定位的特性及双哈希表键值对的相互关联，采用先定位 top-k 结果后划分时间子区间，再标记重叠部分记录数不小于 k 的时间区间为“finished”，最后统计时间长度的方式来输出区间持久性 top-k 查询结果。

(5)研究了面向行列混合数据库的布局敏感技术。行列混合型存储模型（row-column hybrid storage model）是同时高效在线事务处理和在线分析处理的一种解决方案，这些方法都是在物理存储层面采用了行列混合的技术，可以取得一定程度的性能提升。行列混合的技术不仅可以应用于物理存储，也可以用在中间结果、结果集以及网络包。提出布局敏感技术的概念，并

指出其对行列混合型存储模的重要意义，提出布局敏感的数据库网络传输优化方法，设计了行列混合的数据库传输协议。

课题三：高效能存储系统组建方法

根据研究计划开展了相关研究工作。本年度的工作整体分为三个部分，第一部分继续研究存储节点、服务器和数据中心负载能耗测量及其验证方法，在能耗测试基础之上设计并开发数据中心能耗仿真系统；第二部分是设计并开发低能耗存储节点，并研究基于新型存储介质的高能效计算机系统结构；第三部分是研究大规模纠删码存储系统的高能效方法及技术。

设计并开发数据中心下分布式实时负载能耗测量系统。能够支持数据中心三级电力实时监控方式，研究数据中心整体负载和电流流动模式及其供电优化方式。在此基础之上，研究大规模存储系统新型数据分布组织策略；研究多维多目标动态平衡理论及其算法；设计并实现数据中心能耗仿真系统，针对典型存储应用和过程建立高效能算法和调度机制来提高存储整体能效。

设计并实现低能耗存储结点。采用嵌入式处理器和高能效处理架构，控制器满工作能耗小于 10 瓦；支持 8 个以上 SATA 接口，单节点容量超过 10TB，整体功耗小于 100 瓦，I/O 存取性能超过 100MB/s。设计并实现面向虚拟机镜像的重复数据删除；提出基于多级 Cache 算法中数据隐示 Hint-K 算法；提出优化重构前台 I/O 性能 VDF 算法；提出混合的瓦记录系统 HWSR。研究基于新型固态存储器件的高效存储系统结构，设计并实现针对闪存、相变存储器的仿真器，包括块设备和内存仿真；研究 NAND 闪存的寿命模型和写放大特性；设计基于 PCI-E 接口全固态存储加速卡，设计相应的软硬件结构，针对性地优化传统存储栈；对于传统文件系统进行优化发挥固态存储器件的性能优势。

研究基于纠删码的大规模存储系统。设计新型纠删码编码，包括高容错（容错度≥3）编码及其实现方式 PSG 等；研究大规模分组编码 Code-M+ 等；研究重点发挥存储设备局部计算能力的新型编码结构 LDF 等。提出新

型纠删码集群存储节能机制 ECS2 等，利用活动数据节点的可用内存来缓存冗余分块，以关闭冗余节点，达到存储节能的目的。

自 2012 年底至今，在国外学术期刊和会议上发表论文 10 篇，并且 2 篇被 IEEE Transaction on Parallel and Distributed Systems 录用，1 篇被 ACM Transaction on Storage 录用。已申请和授权发明专利各 5 项；获软件著作权 4 个；在人才培养方面，培养博士后 2 人，博士研究生 3 人，硕士研究生 24 人。

课题四：存储服务关键支撑技术

中国科学院计算技术研究所存储中心承担课题 4“存储服务关键支撑技术”的研究任务。自 2012 年下半年中期考核顺利通过之后，本课题组在文件系统元数据集群、数据高可用方面取得了一定进展。

在文件系统和数据分布方面，继续深入研究机群文件系统基于卷的元数据集群技术及其负载均衡技术、机群文件系统客户端低延迟访问技术，并实现了相应的原型系统，使得文件存储系统能够提供面向大规模文件数据的高可扩展的访问能力，并对小文件访问等科学计算、互联网典型访问模式提供针对性的优化。元数据集群技术测试显示，系统提供统一的名字空间，并且文件访问可在 MDS 之间透明的在线迁移，在多目录并发访问下，两台 MDS 相对于一台 MDS，负载分布算法提升了 100%的文件创建性能，60%的文件 stat 性能，233%的文件删除性能。客户端低延迟访问技术的测试结果显示，客户端缓存文件创建技术可以使小文件首次写性能最高提升 128-400 倍，文件删除速度提升 40-60 倍。

在高可靠阵列级存储系统方面，主要继续深入对网络 RAID 阵列系统的关键技术开展研究：面向高效去冗的异步版本化技术以及面向高维冗余的网络 RAID 关键技术，使得高可靠阵列系统在阵列节点扩展、可靠性提升等方面得到深化。此外，课题组已经根据资源分配、快照、缓存、迁移等一系列阵列技术的堆叠，构建了较为稳定的网络 RAID 存储系统平台，为相关技术的深入研究打下了良好基础。目前异步版本去冗技术的实验表

明在顺序写应用中的数据版本确定比例为 100%。高维冗余的事务化技术初步验证了事务与 RAID 存储过程解耦的效果,目前事务 RAID 读性能与原始 RAID 相同，经 VFS 层缓存的顺序写入性能达到原始 RAID 的 95%以上。

课题组在知识产权方面共完成 4 项发明专利的申请，完成博士论文 1 篇，已投稿和正在撰写的论文 3 篇。

课题五：云存储服务和保障机制研究

根据任务要求，课题组本年度在外延云存储服务质量保障机制方面开展了如下工作，主要包括：

(1)提出一个新的指标 TUE（traffic usage effectiveness）来衡量云存储数据同步的节流效率，并全面发掘影响 TUE 的主要因素，可以帮助云存储服务提供者设计更为经济、节流的数据同步机制。研究发现以 Dropbox 为代表的众多云存储系统在内容分发过程中普遍存在“同步流量滥用问题”,从而给云端、客户端以及传输网络都带来沉重却并非必要的流量负担。为解决这一问题,提出并实现了“高效批同步算法”解决该问题,在不明显影响用户体验的前提下、有效避免了流量滥用。发现云的价格策略的不同，有些云对流出带宽的收费较高而存储的价格较低，有些云的收费策略则相反。因此，我们设计了一种基于多云的数据存储和迁移策略，依据用户数据的读写特性把数据存放到不同的云中，并依据数据特性的变化实时迁移数据的位置，同时还可以指导用户选择适合自身需求的云存储服务。

(2)设计并实现了一个基于 P2P 存储的低开销内容预约系统 Beehive，用来缓存和分发用户的预约内容。我们提出了一系列技术（包括网络编码，任务外包等）实现高效且可保障的预约服务质量。考虑到内容预约服务中 P2P 节点间频繁的数据通信，提出组织原 P2P 系统中节点构建端在线传递网络以中转常见的失败或受阻的数据连接。研究在该网络中的流量管理问题，提出最优化模型，联合的降低延迟和减少跨自治域系统流量。

(3)设计了一种由组员节点、超级节点和中心存储组成的混合云存储系统，数据访问采用分级访问模式。 基于固态盘内部越来越丰富的内在并行

性,提出一种在 LINUX 内核块层主动利用并行性来提高性能的 I/O 调度器。其主要思想是按 SSD 逻辑地址空间划分为一定大小的区域，然后以这些区域为调度单位。实验表明 PASS 的性能要高于 Linux 系统自带的四种调度器，并且能够减少 SSD 的擦写操作，延长 SSD 的使用寿命。

课题六：面向数字城市的实时跨媒体信息存储与公众服务

课题六针对课题定位和任务计划安排，围绕理论探索和应用原型系统搭建进行研究，所取得的研究进展主要体现在以下几个方面：

(1)针对传统 GIS 静态数据模型与时态 GIS 时空数据模型缺乏对动态地理现象与其变化机制的表达，并且难以支持实时接入数据的集成表示的问题，根据数字城市中多源空间数据及实时动态信息的特点，研究设计了时空语义集成表示的实时 GIS 数据模型。此模型以时空变化为核心，实现多维动态地理信息数据在时间、空间、语义方面的集成表达，能够准确刻画时空变化的机制与关联关系，支持地理对象、时空过程、地理事件的一致性关联，实现时空变化的多尺度描述，为时空数据的自适应组织与时空过程动态模拟提供基础。为了实证该模型的有效性，课题组提出了一种面向室内火灾动态分析的三维建筑信息模型，此模型充分结合建筑防火专题语义信息，建立面向建筑物内部空间划分与通达性分析的一体化语义表达，实现高动态环境变化下，顾及部件行为特征的语义扩展，支持火灾应急疏散的三维空间逃生分析、动态环境重建。

(2)针对数字城市网络应用中用户访问分布不均和系统各服务器处理能力差异巨大的问题，根据用户的访问特征，提出了一种基于局部访问控制和负载配置的负载均衡方法。该方法基于用户对空间数据的访问服从 Zipf 分布，根据各服务器缓存大小自适应的将用户访问分配至不同服务器以提高热点数据的访问获取命中率。引入排队理论解决数据请求最小代价问题，根据各服务器的处理能力均衡负载，由此获得最佳响应时间，依据数据内容将用户请求分散至各服务器中进行处理，防止一些热点数据带来的集中式访问。且考虑到大规模通信和高并发的用户访问特性，通过动态分配数

据访问请求来提高单位时间的请求处理速度，可以有效提高响应时间和系统吞吐量，提升数字城市中海量空间数据的利用率。

(3)根据面向数字城市的实时跨媒体信息存储与公众服务系统的要求，继续典型环境监测数据、各类高分卫星、资源环境卫星数据的搜集、加工和整理。进一步完善空间环境感知数据示范应用系统，借助时空可视化方法，以等值面图方式展现数据空间分布情况，以空间统计分析图揭示数据相关关系，以动态地图展现数据的空间特性随时间的变化情况，实现空间环境感知数据的时空联合展示。面向专业用户研制 MODIS 遥感数据服务系统制作我国逐月植被指数产品，研制自来水信息监控原型系统，包括自来水管网信息采集、自来水厂监控和水资源自动测报三个子系统，实现自来水厂全信息和水源水情的监测和管理。