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] 02 号　总第 9 期

项目办公室　编　　**2012 年 12 月 10 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 12 月，项目研究成果包括在 INFOCOM、ACM-MM、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 91 篇，申请国内国际发明专利共 57 项，获授权专利 19 项，申请软件著作权 4 项，培养博士 30 人、硕士 96 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 31 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 6 篇（IEEE TC 3 篇、TPDS 2 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 26

项，授权发明专利 5 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，获得 2013 年中国计算机大会最佳技术展示奖。

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，根据中期检查专家的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，根据中期检查专家的加强对大数据应用产生存储需求分析来开展融合存储系统按需服务的研究建议，课题依据大数据应用场景下对于安全性需求，开展数据安全删除、溯源重建等研究。主要研究结果包括：

(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，并完成系统控制管理功能。融合存储系统平台实现了 FMC(FPGA Mezzanine Card)接口和 DDR 接口两套硬件设计；完成了 NAND Flash 控制器和 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，NVMe 驱动已经通过了数据传输测试。

(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 算法将相关数据存储在一个哈希桶，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，以准确地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)存储服务的近似精确 SLO 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，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。针对集中式资源调度对虚拟机进行准确的存储性能保障存在的若干问题，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。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增加/减少效果，动态分配应用的预取长度，与 LRU 和 AMP 相比，ASPM 获得 41.5%-6.8%的整体性能提升。提出了基于恢复缓存感知的去碎片化方法，通过模拟恢复缓存的行为，防止了备份过程中已在恢复感知缓存中命中的数据块被重写，有效避免了去重率的过多损失，同时改进了恢复性能。

(5)复杂应用环境下的数据服务安全保障技术。提出了一种新型的采用溯源信息来分析入侵检测的方法。这种方法对和系统进行交互的进程在线收集溯源信息，并建立表现文件和进程之间依赖关系的规则数据库。每当系统中有新的进程活动时，通过和规则库中的依赖性关系规则进行比较，

就可以分析出是否存在入侵。不仅如此，该机制还可以通过构造溯源图，直接找出具体的入侵路径。这样管理员就可以对入侵链上的每个事件进行分析，从而方便快捷的找出系统漏洞。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)采用 Flash+PCM 的融合存储平台原型系统在 2013 中国计算机大会和国际超级计算机大会 SC2013 上进行了成果展示，取得了良好的效果。升级部署了基于再生码的云存储系统 Ustor（www.ustor.cn 电信网）。上线部署了“晒点”系统（http://paper.ustor.cn/）数据共享服务平台和 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 年课题二的研究主要集中在海量非结构化数据的处理编程模型及海量数据的挖掘分析与计算方法等方面的关键技术研究。经过今年的研究，取得了一些成果,在软件学报、物理学报、IJCAI 2013、WSDM2013 和 WISA2013 等国际期刊与会议上录用或发表了 EI 国际检索论文 11 篇,SCI 国际检索论文 1 篇，其中 NCIS2013 和 WISA2013 论文获得优秀论文奖；申请专利 6 项；进一步完善了海量结构化数据处理平台华鼎-C 以及海量非结构化数据存储管理平台华鼎-U，并在华鼎-K 平台上进行了部分大数据分析应用研究，开发设计了移动健康服务平台。

（1）开展了基于大数据的社会影响力分析，提出了一种基于用户行为的社会网络影响力形式化模型。本研究通过分析微博的大量用户群形成的社会网络及其用户发表微博形成的大数据进行用户行为分析并进行挖掘，得出一套基于用户行为的社会网络影响力形式化模型，其中论文《Social

Influence Locality for Modeling Retweeting Behaviors》发表在人工智能领域国际顶级学术会议 IJCAI2013 上。通过总结现有的基于大数据的社会网络影响力分析，提出了相应的模型和算法。其中有 2 篇论文发表在数据挖掘领域重要国际学术会议 WSDM2013 上。通过对海量大数据的研究，在社会网络领域提出一种新的网络传播中最有影响力节点发现方法。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，该研究成果发表在物理学报上（SCI）。

（2）研究了一种支持超大数字图书馆非结构化海量数据的插件式系统。在大数据时代，收集海量复杂数字信息的需求为数字图书馆软件的设计带来了新的挑战，这对建立超大数字图书馆提出了需求。为了解决在开发超大数字图书馆的过程中数据管理的问题，需要设计新的技术和方法。我们设计了一个新的插件式系统 PuntStore，作为用于超大数字图书馆的通用解决方案。PuntStore 支持多种存储引擎和索引引擎，从而能够高效的存储和检索信息。我们还为 PuntStore 设计了一种新的索引结构 pLSM，用以满足数字图书馆应用的特别需求。我们已经成功地将 PuntStore 应用在中国科技史数字图书馆项目中，这说明 PuntStore 可以有效地支持超大数字图书馆系统。论文发表于国际数字图书馆领域重要国际会议 ICADL2013 上。

（3）提出了一种面向大数据的语义安全模型及其语义处理方法。现有安全模型基于 PKI 的体系架构，但是基于 PKI 的安全架构在大数据的云环境下很难实现对文件本身的真实性检查及其在时间和空间发生转移时的文件数量安全性检查。针对该现象，我们对云环境下大数据的安全性进行了研究，并提出了一种面向大数据的语义安全模型，同时针对大数据安全管理过程中提出了相应的语义处理方法。其中论文发表于国际会议 WISA2013 并申请发明专利 1 项。

（4）在大数据的分析上分别针对新闻领域与电子商务领域提出了海量非结构化数据的分析方法。以世界主要国家权威网站上的实时新闻为研究对象，将语言学、信息科学、图书情报学的研究进行交叉融合，通过对每日互联网上产生的不同信息源，乃至不同语言源的最新海量新闻信息的研

究分析，设计一套对新闻信息的智能收集、快速获取、去重去伪，并快速与该信息的背景资料进行有效整合的基础理论模型及其算法，从而找出实时新闻报道与用户兴趣点之间的联系，实现实时、按需推送的目标，并为解决大数据环境下海量信息的实时处理的某些关键难点科学问题提供理论基础和验证。提出了针对电子商务等数据高可靠要求的场景，提出了一种基于 Paxos 的 Key-Value 存储系统。该系统使用基于 Multi-Paxos 的复制方法进行同步复制，保证了高容错性、低延迟，在一般故障场景下可以为用户提供不间断的服务，论文发表于 WISA2013 会议及国内存储会议上，并申请发明专利 1 项。

（5）进行了华鼎平台的进一步完善研发，并开发了基于大数据的移动健康服务平台。现代社会人们生活节奏加快、缺乏锻炼、工作学习压力加大，使人们不自觉地进入“亚健康”状态，特别是随着我国步入老年社会，人们越来越关注身体的健康，推行健康服务可以满足大多数人群的需求。移动健康服务平台通过结合可穿戴设备、智能手机应用和后台大数据处理平台，可以给人们展示长期的基本健康数据和环境数据，并用于比较和分析。可以实现在线的健康数据浏览和监控，使得关注家人和朋友的健康不再是一个梦想。在华鼎平台研究的基础上研究了大数据的索引机制及其云资源的预分配方法等，并申请发明专利 2 项。

课题三：高效能存储系统组建方法

课题三“高效能存储系统组建方法”按照研究计划开展研究。在本年度的研究工作整体分为三个部分，第一部分研究基于新型存储介质的高能效计算机系统结构，设计并开发低能耗存储节点；第二部分针对大规模存储系统，利用纠删码和重删技术提高整体能效；第三部分，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研究，在此基础之上设计并开发大规模存储系统（如数据中心）能耗仿真系统。

利用新型存储器件大容量、高性能的特点，研究基于新型存储器件的高效存储系统结构，设计混合瓦记录磁盘系统 HWSR；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 接口全固态存储卡，针对性地优化传统存储栈；设计适应固态存储器件特性的新型文件系统。另一方面，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，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。

研究基于纠删码的大规模存储系统。设计新型纠删码编码，包括提出了新阵列编码V2-Code码提高重构性能;提出了一种采用并行化思想的RAID数据分布方案 S2-RAID；针对 RAID-5 扩容，提出一种具有最小迁移数据和最短在线迁移时间的基于校验块迁移方法 PBM;提出基于多级 Cache 算法中数据隐示算法 Hint-K；提出优化后台重构和前台 I/O 性能的 VDF 算法；针对异构纠删码存储集群提出感知服务质量的读优化策略 ROS;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，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。设计并实现面向大规模虚拟机镜像的重复数据删除。

针对大规模数据中心环境，开发一个低恢复带宽，低维护成本的云存储系统 BMCloud。提出了一种基于编码技术的面向大数据备份的优化算法 ICRS。设计并开发数据中心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提高存储整体的能效。

在 2013 年，在国外学术期刊和会议上发表论文 15 篇，其中在国际学术会议 IEEE Cluster2013、IEEE MASCOTS 2013、IEEE BigData 2013、IEEE NAS2013 和 IEEE HPCC2013 上发表多篇论文，在期刊 IEEE Transaction on Parallel and Distributed Systems 上发表或录用 4 篇,在 ACM Transaction on Storage 发表 1 篇。计划明年向顶级学术会议和投稿投稿 6 篇以上。已

申请和授权专利 8 项；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 7 人，硕士研究生 36 人。

课题四：存储服务关键支撑技术

中科院计算所存储中心在 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”中承担课题 4 的研究任务。2013 年度，本课题组在文件系统、网络 RAID 集群方面取得了一定成果。

在文件系统方面，课题组继续深入研究机群文件系统基于卷的元数据集群技术及负载均衡技术、客户端低延迟访问技术，实现了相应的原型系统，并将系统移植到最新的 pNFS 平台上，通过了 ltp 测试。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，典型小文件读取性能提升 10 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，课题组已经基于资源分配、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。目前已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%，通过理论分析，证明在复杂应用环境下，版本确定比例可以达到 95%以上。高维冗余的事务化技术验证了事务与 RAID 冗余过程解耦的效果，经过优化，目前事务 RAID 读性能与原始 RAID 相同，顺序写入性能达到原始 RAID 的 95%以上，延迟增加控制在 4%以下，CPU 开销控制在 1%以下。并且，通过结合日志结构的精简配置技术，将随机写性能提高到原始 RAID 的 106%到 206%。网络 RAID 存储系统平台保障正常 I/O 路径和节点故障恢复路径的数据访问，目前系统通过了 72 小时的正确性和稳定性测试。

此外，课题组在知识产权方面共获得 1 项专利的授权，完成 4 项发明专利的申请，完成博士论文 1 篇，录用会议论文 1 篇，已投稿和正在撰写的论文 4 篇。

课题五：云存储服务和保障机制研究

根据项目中期的调整，2013 年本课题的任务提出服务目标量化指标，研究基于协作的云存储服务质量优化机制。课题今年重点围绕“经济”这个服务指标，设计相应的机制帮助用户节约使用云服务的成本。课题开展了三方面工作：

（1）基于多层协作的混合云存储系统。为了达到经济这个指标，我们提出了多层次的协作模式，包括用户—用户间的协作、用户—云之间的协作和云—云之间的协作。基于这种多层协作模式，我们构建了一种由组员节点、超级节点和中心存储组成的混合云存储系统 M-Cloud，目前已经完成了原型搭建，提供了优化机制部署的平台。

（2）在不同的协作层次，设计了节省云服务成本的保障机制。首先，提出一种新的用户—云之间的数据同步机制。我们测量了以 Dropbox 为代表的众多云存储系统，发现云存储的核心操作—“数据同步”消耗巨大网络流量，即用于维护同步过程的“控制流量”远超过实际要同步内容的“数据流量”，从而给云端、客户端带来沉重却非必要的流量负担。我们对云存储内容分发过程进行了一系列精心设计的测量，从网络协议、操作系统两个层面分析出该问题的发生原理；然后提出并实现了“高效批同步算法”解决该问题。其次，我们还提出了一种基于多云的数据存储机制。我们发现云的价格策略的不同，有些云对带宽的收费较高而存储的价格较低，有些云的收费策略则恰恰相反。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，有效的帮助用户减少云服务的费用。

（3）设计实现了基于混合云存储的示范应用。首先，我们设计并实现了（2 万 9 千余行 C++ 代码）一个基于 P2P 存储的低开销内容预约系统 Beehive。Beehive 有效组织用户机器上闲置的存储和带宽资源，用来缓存

和分发用户的预约内容，避免了对商业云存储的使用。其次，我们还设计并实现了大规模备份系统 YuruBackup 的原型系统（约为 12,000 行代码）。系统能够以较小的流量、在较短的时间内完成数据的备份与恢复。降低备份和恢复过程中的带宽使用，减少存储开支。

此外，课题组在知识产权方面共发表论文 14 篇，其中包括 3 篇 A 类、3 篇 B 类，5 篇 SCI，7 篇 EI。获得授权专利 6 项，申请专利 11 项，软件著作权 4 项。

课题六：面向数字城市的实时跨媒体信息存储与公众服务

本阶段，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）基于三维地理空间框架的多源动态时空信息内容描述，研究三维地理空间框架的实时多维动态数据表示模型、数据组织方法和索引方法。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难题，课题组以时空变化为核心，建立多维动态地理信息数据在时间、空间、语义方面的集成表达，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。针对传统面向分区存档管理的视频数据组织模式缺乏视频数据之间、视频数据与地理环境之间时空关联的问题，课题组设计地理视频内容自适应的数据组织方法：利用生成的“对象-过程-事件”三域描述模型，基于视频数据的地理语义、内容语义和元数据语义相似性，对地理视频数据多层次分组聚集。研究支持广域范围地理视频大数据可认知计算与推理的时空关联分析，为广域范围的大数据搜索任务提供必要的时空约束，有助于危机事件的感知理解。地理视频数据的复杂逻辑结构及其视像内容的时空关联、多维动态特性使得直接对视频流数据进行索引和检索十分困难，课题组提出一种语义感知的地理视频数据时空索引方法：建立视频数据的多层次语义描述模型，解析视频内容语义，以事件为核心，基于事件和时空相关性，针对语义内容丰富

的视频元数据设计动态自适应的分布式索引机制，为具有良好逻辑结构的视频单元索引提供依据，实现动态三维空间数据的快速查找、插入、更新和删除。

（2）研究数字城市数据存储与访问优化。针对用户访问不均匀的特性，考虑异构分布式存储系统各服务器处理能力差异，课题组提出一种基于用户访问密度的负载均衡方法，根据各服务器缓存大小自适应调度用户访问，排队论最小化数据请求代价，均衡各服务器负载，获得最佳响应时间，防止集中式访问热点数据，实现存储系统对用户访问的高效实时动态适配。深度挖掘用户访问特性，从用户访问任务角度出发，改进 PLSA 模型分析访问文件、系统应用和访问任务之间的相关性，提出基于用户访问任务的合并与预取模型，尽可能依据访问任务合并小文件，根据访问任务间转移概率选择预取文件集；针对海量时空数据小文件，结合用户的访问行为特征和访问数据自身属性，序列化和用户请求，依据时间、空间和类型三个属性参数化请求访问的文件，提取文件特征序列，通过模板匹配发现访问规律合并相关文件。可以有效降低 MDS 负载和请求响应时延，提高数字城市小文件的存储与访问性能，可以作为解决各类小文件问题的重要基础。

（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，继续各种原始数据和产品数据的搜集、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实现空间环境感知数据的时空关系及变化联合展示；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；研制无人机低空遥感与视频实时采集系统，实时采集航拍静态影像、航拍视频影像和无人机 GPS 飞行轨迹信息，地面设备实时获取、动态展示采集数据；研制湖北省农村饮水工程信息管理示范应用系统，实现对农村水厂基本信息，包括地理信息、饮水工程专题信息和部分主要工程实时监测信息的普查登记和统计查询；利用这些示范应用系统考察数据存储系统对实时动态数据获取、存储、分析和管理能力，对时空分布的海量小文件存储与访问处理能力以及应对海量多源异构数据处理和应用的能力。