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

[2014] 01 号　总第 10 期

项目办公室　编　　**2014 年 07 月 01 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”2014 年上半年，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 2014 年 6 月，项目研究成果在包括 INFOCOM、USENIX ATC、AAAI、HPDC 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 59 篇，申请发明专利共 45 项，获授权专利 17 项，申请软件著作权 1 项，参与标准制定 1 项，培养博士 24 人、硕士 92 人、博士后出站 1 人。

课题一：融合存储体系结构与服务架构研究

课题围绕从融合存储系统的组织、面向复杂应用环境的按需服务开展了研究，发表了 18 篇论文，其中在中国计算机学会推荐的 A 类国际期刊和会议上发表学术论文 7 篇（包括 IEEE TC 3 篇、TPDS 2 篇、INFOCOM 2 篇），B 类论文 7 篇（包括 ICDCS 1 篇、USENIX ATC 1 篇、Performance 1 篇、DCC 1 篇、IWQoS 2 篇）。申请专利 11 项，专利授权 7 项；参与标准制定 1 项；培养博士后 1 人，博士 2 人，硕士生 41 人。

通过研究，在融合存储系统和按需服务关键技术等方面取得一定成果：进一步构建和完善了能够支持Flash/PCM等多种异构存储介质的融合存储平台，通过采用多通道技术提升固态存储设备的容量和速度，提出海量多媒体数据基于关联特征的组织管理与查询，实现了海量存储系统中分布式锁及其一致性的解决方案；提出基于信誉激励的多维存储服务质量保障机制和固态存储性能隔离技术，针对复杂应用环境下，多样和突发的企业应用需求合理地分配系统资源以保证应用服务质量；深入开展了面向大数据应用中重复数据删除技术的研究与优化方法；针对云存储的私密性问题和可靠性问题，开发了基于高效编码的混合云存储系统等。

主要研究成果包括：

(1) 通过设计完成新型的 NAND Flash 和 PCM 硬件子卡，进一步构建和完善了能够支持 Flash/PCM 等多种异构存储介质的融合存储平台；通过多通道技术提升容量和速度；进一步优化了主机和设备端的关键软件，包括 FTL、中断、DMA 等底层驱动软件，初步实现软件定义固态存储思想。

(2) 基于关联特征的数据组织与面向异构非结构化数据的近似查询技术。提出 Necklace 新结构，进行数据特征和行为模式的挖掘、获取和快速分类，使用增强 cuckoo hashing 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。

(3) 提出一种面向多维存储服务质量保障的 I/O 调度算法 Courier。云存储系统环境下现有策略不能同时支持应用对带宽和延迟的需求，导致其不能广泛整合到云存储服务中。Courier 调度算法动态切换于基于反馈的延迟控制器和基于信誉分配的带宽分配算法之间，并采用了更精细的优先级粒度，通过实时监控请求最后期限，并采用请求级别的优先级策略以向紧急请求提供优先服务。通过采用合成和真实负载进行大量对比测试，调度算法不仅可以满足不同应用的存储性能需求，同时存储系统整体性能也提升了约 20%。

(4) 提出一种面向固态盘设备的调度算法 FBCQ，为并发应用提供性能隔离并优化存储系统整体性能。FBCQ 采用二层调度框架，上层调度器负责提供性能隔离，底层调度器专注于提高固态盘访问性能。为了实现性能隔离，上层调度器通过管理应用服务时间进行 I/O 调度，并且根据固态盘性能特征区别对待请求的读开销和写开销。调度器通过在底层维护基于通道的队列以挖掘固态盘的内部并行性。FBCQ 调度器可以有效地实现应用之间的性能隔离，并提高固态盘存储设备的整体性能，在性能隔离方面和性能优化方面都优于其它调度算法。

(5) 提出了基于历史信息的重写算法 HAR。在数据去重备份系统中，每一次备份的数据块经过数据去重后被离散地分布在容器中，这导致了数据碎片问题，降低系统的恢复性能。提出了基于历史信息的重写算法 HAR，通过挖掘历史信息更准确地识别和重写碎片数据块。由于 HAR 将有效地数据块聚集在容器中，减少合并操作次数；同时提出了容器标记算法，识别有效的容器而不是数据块减少了垃圾回收的元数据开销。因此 HAR 可以非常有效地清理稀疏容器，避免重写乱序容器，从而在保证去重率的前提下，提高恢复性能。

(6) 数据去重感知的差量压缩方法。传统重删技术中对相似数据判断有较大的计算和索引开销，提出一种差量压缩方法 DARE，通过采用重复数据删除感知的相似查找模块，简化了差量压缩的数据查找过程，充分利用现有重复数据删除信息快速查找相似数据。同时 DARE 采用基于相似性的相似查找，补充基于局部性的相似查找，通过对剩下的非相似非重复数据的数据块计算超级指纹，尽可能查找更多相似数据，提高压缩效率。DARE 提出的结合重复数据删除感知和基于超级指纹的差量压缩进行相似数据查找压缩，具有相似数据查找开销少、数据压缩效率高并且吞吐率高等优点，可应用在数据备份系统、归档系统、数据远程复制与迁移等场合，节省数据存储空间，提高数据传输效率。

课题二：海量数据组织与资源共享的方法研究

2014 年研究主要集中在对复杂环境中结构化、半结构化和非结构化数据的存储、组织、处理、分析等方面的关键技术研究。经过今年上半年的研究，取得了一些成果。在 AAAI2014、WSDM2014 和 WISA2014 等国际会议上录用或发表了 EI 国际检索论文 11 篇，其中 CCF A 类论文 2 篇、C 类论文 3 篇；申请发明专利 1 项，获得专利授权 1 项；培养博士后 2 人，博士生 5 人，硕士生 8 人。主要成果包括：

（1）在复杂环境中海量数据的安全存储技术方面，提出一个新的 SSD 缓存管理策略 Tri-List，采用“桶”的数据结构，把多次小的随机写转化成为一次顺序写的操作，并执行一次批量插入，对于已有索引项的更改将会被采用一种延迟写出的方式从而避免小的随机写操作。此外实现了一种自适应机制来动态调整 AB 树的桶结构，可以根据在线的工作负载动态调整桶结构，从而优化了读操作和写操作的表现。

（2）在复杂环境中海量数据的组织技术方面，研究了数据的索引与检索技术。提出了一种面向海量空间元数据索引的数据结构和分组方法，该方法在分组索引的基础之上实现一种可变化参数的 PK 树索引，并在该索引结构的基础上实现了海量空间元数据的索引存储和数据的检索算法。提出了一种基于区间窗口的方法对经典的 top-k 算法进行扩展，支持加权持久性 top-k 的时间约束和权重约束，将持久性 top-k 问题被转化为能够用经典 top-k 算法解决的子问题集合，然后汇总而得到加权持久性 top-k 结果。

（3）在复杂环境中海量数据的处理技术方面，研究了视频数据的格式转换和版权保护技术，以及针对健康领域提出的社交媒体数据的主题检测、文本聚类和自动问答技术。提出了一种将 GIF 格式转换成视频格式的方法，可以在不被察觉视频质量受损的前提下显著地压缩视频传输的大小，实验证明 GIF 的加载速度提高了 8 倍。提出了一种视频版权保护方法 HuaVideo，充分利用了 HTTP 报文头的格式来判断请求的合法性，用户每次访问相同的视频时都需要通过不同的 URL 且每个 URL 只能使用一次，从而保证视频服务器中的内容不会被非法盗链。提出了一个高级的深度学习方法来检

测互联网医疗健康相关的主题，和基于 LDA 的方法从文本集中的长（短）文本为短（长）文本获得辅助信息，从而减缓短文本的稀疏性和减低长文本的维度，实现文本集中稀疏和维度的平衡。健康问答技术充分利用和挖掘已有的海量问答数据的价值，即时、直接、准确、全面地回答用户提出的新问题，可以便捷快速地服务用户，大大减少医生的重复劳动。

(4) 在复杂环境中海量数据的分析技术方面：通过将社交网络用户发布的图片及其好友对该图片的评论相结合进行建模，提出了一种通过社交网络分析作者情感的方法，可将与揭示一张图片内在情感紧密相关的那些评论和不相关的那些评论区分开来，通过在开放的网络相簿数据集上的实验，提出的模型可以显著提升对图片发布者情感的分析准确率，研究成果发表在国际顶级学术会议 AAAI2014 上。同时，提出了一种全新的连接图像模型与信息扩散处理的思想，精确地定义了基于非进行扩散模型的主动学习问题，并给出解决方案 MaxCo，该方案通过构建一个简化的问题集合并证明其为最小原数据集，以及一个迭代贪婪算法 MinSS 解决该最小原数据集问题，同时从最优解的下界开始寻求自动学习的解空间。研究成果发表于重要国际会议 WSDM2014 上。

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”根据研究计划开展。在本年度的工作整体分为三个部分，第一部分研究面向数据中心、基于纠删码的高能效大规模存储系统；第二是研究基于固态盘和光盘的新型高效能存储介质、设备和系统；第三实现低能耗存储节点，并构建基于测量的高效能存储系统原型。

研究面向数据中心的大规模存储系统，设计平衡存储效率、运行能耗、性能和可靠性的新型纠删码布局及其存取优化方法。设计 PUSH 机制引入流水线式 I/O 调度来提升各存活节点的资源利用率的流水线式重构；设计多等级容错存储系统 MFTS，建立数据特征（包括数据静态属性和数据访问模式）和冗余编码方案之间的最优映射关系；设计一种高效的集群扩容方案 Scale-RS，发挥纠删码的结构特性来优化数据迁移和校验更新，不仅让数据

块迁移总量达到理论下界，而且能最小化校验更新所导致数据访问量。设计一种新的编码平衡 P-Code 编码，设计用于异构纠删码存储集群的负载感知恢复方案 LARS。

研究以闪存、磁内存和光盘为代表的高能效新型存储器件和系统。提出和建立基于性能和能耗的固态盘行为放大模型（Bamp），计算单位用户写数据量的能耗下 NAND Flash 内编程、多余编程、多余擦除和多余读等操作的总能耗；从 NAND Flash 的编程角度，提出高精度 NAND 固态盘写放大研究模型和测量方法（RB-Explorer）。针对海量大数据长期保存问题，研究和设计大容量并行存取光盘库及其关键技术。

设计并实现低能耗存储结点和系统级能耗优化。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，I/O 存取性能超过 100MB/s。继续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能耗模式及其引入新能源供电优化方式。研究并设计并实现数据中心海量虚拟机镜像的分布式重复数据删除技术。

在 2013 年底至今，在国外学术期刊和会议上发表论文 12 篇，其中 2 篇被 IEEE Transaction on Parallel and Distributed Systems 录用，1 篇被 IEEE Transaction on Computers 录用，1 篇被 IEEE Transaction on Dependable and Secure Computing 录用，1 篇论文被 International Symposium on Reliable Distributed Systems，SRDC2014（计算机学会 B 类会议）录用。已申请和授权专利共 22 项；在人才培养方面，培养博士研究生 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。2014 年上半年度，分布式文件系统元数据集群、网络 RAID 存储集群在技术深化和系统成熟度方面取得了一定进展。

在文件系统方面，课题组在基于卷的元数据集群技术及负载均衡技术的基础上，继续深入研究文件访问负载的快速迁移机制，为元数据集群技术的高可用化打下基础。在该方面，提出了一种基于日志读取和迁出端恢复状态的元数据服务迁移方法，将平台系统原有迁移方法 90s Grace Time 延迟问题优化控制在百毫秒级；提出了一种细粒度的状态迁移控制方法，降低了状态迁移开销与状态规模相关性。在小文件低延迟访问技术方面，实现了基于 readdir++的 layout 批量预取技术及相应的原型系统，对比 pNFS 系统，Readdir++技术可将海量小文件读取访问过程中元数据性能提升到 14.27 倍，总体性能提升到 1.78 倍，元数据耗时占比由 47.11%下降到 5.87%。

在高可靠阵列级存储系统方面，课题组基于资源分配、快照、缓存、迁移等一系列阵列技术的堆叠构建的高可用存储集群系统平台 BWRAID，其冗余单磁盘单节点的系统原型已经由系统原型发展至产品级成熟系统。目前正在设计支持纠删码冗余的存储阵列系统。在本方面 2014 年已申请专利 3 项。

针对广域存储系统面临的广域大目录修改后的本地缓存更新问题，研究并实现了目录项分块检测的精确更新算法，基于目录项分块检测的更新算法在大目录下的小量更新时，性能有约 10 倍的提升。

此外，在知识产权方面，课题组发表或录用论文 5 篇，完成 3 项发明专利的申请，已投稿和正在撰写的论文 2 篇。

目前，通过存储设备集群 BWRAID、支持分布式元数据集群的 PNFS、广域存储系统等一系列技术的积累，将逐步配合构成具备高可用、集群化的水平扩展式存储系统平台。

课题五：云存储服务和保障机制研究

2014 年本课题的任务是通过实际部署协作云存储，发现实际问题，提高系统对外服务的性能和降低系统的服务成本。具体来说，我们要回答下面几个重要的实际问题：（1）如何以低成本管理上层应用的数据；（2）

如何以低成本持久存储冷数据；（3）如何以地开销提高缓存性能；（4）如何调度资源以低成本保证应用服务质量。课题开展了以下几方面工作：

(1)基于共享的数据管理机制，我们发现多个并发应用经常会使用同一份数据。因此，设计了一个新的数据管理系统 Seraph。该系统支持多个并发任务在内存中共享使用一份数据，大大节省了数据对稀有资源（如缓存）的占用。

(2)基于用户态文件系统实时云同步机制。目前应用程序在用户端和云端平滑转移，对本地文件系统中的文件实时同步到云端的需求越来越大，为此设计了一种普适性的高效实时同步策略，同时给予这种同步策略设计了文件版本控制和协同操作的机制。

(3)基于弱化固态盘纠错码(ECC)的缓存机制。为了解决基于闪存的固态盘可靠性变得越来越低的问题，在固态盘内部采用了可编程芯片控制器(PMC)，缓存管理器通过专门的访问接口将数据块的存储需求传递给固态盘。该系统能够获得约 30%的性能提升。该系统还减少了固态盘用于垃圾回收的时间(20%左右)。

(4)低成本的资源分配机制。目前云平台上的资源分配方法并不能很好的满足视频点播应用对服务质量的需求，且系统开销较高。对云平台的带宽和存储等资源的调度建立综合的数学模型，提出可行的分布式求解算法 DREAM（-L），能够很好的保证云平台对视频点播应用做出的服务等级承诺，而且极大的降低了云平台开销。

知识产权方面，课题 5 今年发表 6 篇高质量论文，包括 HPDC，INFOCOM, ATC, TKDD, 获得授权发明专利 3 项，申请 6 项，软件著作权 1 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，围绕应用原型系统搭建和理论探索进行研究，所取得的研究进展主要体现在以下几个方面：

(1)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，继续完善和更新武汉市数字城市示范系统功能。本阶段完成了以事件为对象并基于 GIS 构建城市公共事件时空分析与应急辅助决策系统，提升当前数字城市及未来智慧城市的服务能力。同时面向最终项目结题，构建面向数字城市应用的时空数据访问负载模型，形成大规模并发访问负载，模拟真实的访问负载，以反应真实的用户访问情况，为最终系统性能测试提供标准访问输入源，同时生成访问日志，为预取、缓存和数据布局提供数据依据，为后续测试分布式存储系统应对高并发、低延时、高聚合带宽访问时的数据处理和分析能力提供基础。

(2)针对复杂计算和数据密集型计算在大规模分布式存储系统产生的跨数据中心数据调度问题，课题组根据 “数据共用”现象和计算执行频次定义数据集之间的动态计算相关度，提出一种基于动态计算相关度的数据布局方案，将动态计算相关度高的数据集尽可能部署在同一个数据中心，最小化跨数据中心数据调度次数。该方法实现复杂度低，对于细粒度划分的海量数据集具有良好的性能，数据中心的增加或减少对方法的实现复杂度几乎无影响，非常适合应用在实际的分布式系统管理中。针对空间信息的大规模用户访问，提高海量空间数据访问形成的空间访问统计数据传输服务质量，课题组提出一种云计算环境下空间访问统计数据的点云聚类压缩算法。通过对空间访问统计数据映射成空间点云，空间数据的访问次数信息映射成点云向量，将空间访问统计数据的压缩转换为空间点云的压缩，借助空间点云聚类梯度剔除偶发性访问形成的离散点，并通过空间聚类提取对空间统计数据进行压缩，可以大大节省空间统计数据量。

(3)针对传统 GIS 数据库引擎是以离线式存储管理地理实体的空间和时态信息为核心，难以支持物联网和传感网观测数据的动态更新与实时处理的问题，课题组研究设计了一种内外存协同的 GIS 数据库引擎，主要包括实时接入、自主加载和综合存储三大功能模块。实时接入模块利用内存数据库的“不落盘”特性，建立面向传感器数据流式接入处理的内存数据库，支持实时数据流的在线处理特别是变化探测；自主加载模块发挥关系型数

据库关系完整性约束和结构化数据高频更新的性能优势，建立支持用户决策的主题数据库，保证高效的信息服务；综合存储模块发挥 NoSQL 数据库可扩展、高性能的架构和性能优势，建立持久化存储海量非结构化传感器历史接入数据和主题案例备份数据的综合数据库，并支持实时计算和主题数据自适应聚合的高效检索和调度。