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] 02 号　总第 11 期

项目办公室　编　　**2014 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 2014 年 1 月份以来，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 2014 年 12 月，项目研究成果包括在 INFOCOM、FAST、USENIX ATC、HPDC、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 85 篇，申请国内国际发明专利共 36 项，获授权专利 20 项，申请软件著作权 1 项，培养博士后 2 人、博士 14 人、硕士 61 人。项目研究成果获国家技术发明二等奖 1 项，获省部级科技进步一等奖 1 项。

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 35 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 11 篇（IEEE TC 3 篇、TPDS 3 篇、INFOCOM’14 2 篇、INFOCOM’15 2 篇、FAST ’15 1 篇），B 类论文 9 篇。

新增牵头制定国家标准《分布式异构存储管理规范》。申请发明专利 15 项，授权专利 9 项，其中一项美国专利；培养博士后 1 人，博士 4 人，硕士生 21 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: 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，实现设备端的关键软件；提出基于虚拟块并行的混合映射算法（VBP-FAST），通过三级并行，极大降低完全合并操作代价。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。设计了一种基于溯源的视频共享系统 Provis，利用视频溯源来解决近似重复视频给系统和用户带来的存储和传输开销的问题。针对大数据应用中存在数据重复存储效率低的问题，开展数据去重技术研究与优化，提出 Asymmetric Extremum（AE）分块算法，寻找非对称区域内的局部极值来解决边界偏移问题。

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。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。融合存储硬件平台拥有改进丰富的计算和存储资源，相比 Virtex-6 FPGA 采用的 MicroBlaze 嵌入式软核处理器，Zynq ARM 双核处理器具有更强的运算能力（1GHz 主频），适合完成复杂的计算任务，并且具有丰富的外部设备接口，便于与 PS 层的外设和 PL 层的用户自定义逻辑进行交互。

(2)提出了一种新的 FTL 算法，称为基于虚拟块的并行全相连映射算法（VBP-FAST），该 FTL 算法中将闪存空间分为虚拟块（VBlock）和物理块（PBlocks)。VBlocks 可以充分利用闪存中的通道级、die 级和 plane 级并行。通过这三级并行，VBP-FAST 算法可以极大的降低完全合并操作的代价。同时，VBP-FAST 算法利用 PBlock 保留了部分合并和交换操作的优点。通过真实的 trace 模拟结果显示，VBP-FAST 算法在随机负载的情况下比 FAST 算法性能提高了 5.3--8.4 倍，对于顺序负载，它也比 FAST 算法性能提高了 1.7 倍左右。

(3)为了缩小存储系统和应用感知的视频服务之间的差距，提出了一种高效的实时视频检索机制，称为 FastVR，它能支持快速的近似视频检索。FastVR 在大规模存储系统中有着显著的空间效率和时间效率。FastVR 的具体实现是借助空间高效的索引结构和简洁特征表示(有助于基于关键帧的比较)。提出一种基于溯源信息的文件元数据查询系统的技术方案，通过溯源信息提供的关于文件元数据之间的动态的相关性来加快查询。所提出的方法充分利用了元数据的多维静态属性以及统计的部分动态属性的相关特性，而且利用溯源信息所反映的文件在时间维度上的相关特性加快元数据查询的速度。设计了一种基于溯源的视频共享系统，称为 Provis。 Provis 利用视频溯源来解决近似重复视频给系统和用户带来的存储和传输开销的问题。

(3) 数据去重分块算法中，定长分块因为有边界偏移问题，使用定长分块仅能获得较低的去重率，基于内容的分块算法根据局部内容产生数据块边界（即切点），只要局部内容不变，边界就不会改变。这些分块算法都存在吞吐量过小和不能检测低熵字符串等问题。提出的 Asymmetric Extremum（AE）分块算法是 MAXP 算法的改进。AE 寻找非对称区域内的局部极值（最大值或最小值）来解决边界偏移问题，并且将这个点放在数据块的中间而不是作为切点，这样会有少量的去重率的损失，但是极大地减少了分块所需的操作，吞吐量是传统算法的大约 3 倍。

(4) 复杂应用环境下的数据服务安全保障技术。提出一种基于二次哈希的收敛加密策略（TCE：A Twice-Hash based Convergent Encryption Strategy），TCE 首先计算数据块的哈希来产生密钥，然后再次哈希得到数据块指纹。TCE 实现了先去重后加密的方式消除了对重复数据的加密操作，减少开销；通过增加第一次哈希值的随机长度的填充，降低了哈希冲突的概率，提高了 TCE 的机密性。

(5) 实现并开源了数据去重存储系统 Destor。它将数据去重系统映射到一个 N 维的参数空间，包括键值存储、指纹缓存和预取、抽样等。每个维度代表了一个设计要素，并有多个值可供选择。参数空间里的每个点都代表了一个现有的或者潜在的设计方案，同时也是各个性能指标的权衡：备

份（写）性能，恢复（读）性能，内存开销，存储成本等。通过对整个参数空间进行了详细的分析，并指出了现有方案存在哪些缺点以及优化方法。Destor 作为平台可以帮助研究人员根据需求选择合适的设计方案。

课题二：海量数据组织与资源共享的方法研究

2014 年课题二的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进行研究，在国际会议发表 EI 检索论文 13 篇，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 3 篇、C 类论文 3 篇；申请专利 1 项，获得专利授权 1 项；培养博士后 1 人，博士生 1 人，硕士生 4 人。主要成果包括：

（1）复杂环境中海量数据的安全存储技术：提出一个新的缓存管理策略，根据页面请求频率将缓存区分为三个部分，据页面请求热度选择替换决策；实现自适应动态调整 AB 树的桶结构，根据在线工作负载动态调整桶结构，优化读操作和写操作的表现，并在此基础上提出基于可信固态硬盘的大数据安全保护模型。

（2）复杂环境中海量数据的组织技术：研究数据的索引与检索技术，提出面向海量空间元数据索引数据结构和分组方法，在分组索引基础之上实现可变化参数 PK 树索引，海量空间元数据索引存储和数据检索。提出基于区间窗口方法支持加权持久性 top-k 时间约束和权重约束。针对基于阈值和 top-k 字符串相似性搜索提出统一框架，设计相似搜索算法以兼顾以上两个方面。

（3）复杂环境中海量数据的处理技术：研究视频数据格式转换和版权保护技术，可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。设计视频服务架构 HuaVideo 以解决版权保护问题，利用 HTTP 报文头格式判断请求合法性，从而保证视频服务器的内容安全。提出了一个基于主题模型的 ADLDA 方法来解决混合长度文本聚类，能同时考虑文本集中的长文本的高维性和短文本的稀疏性。

（4）复杂环境中海量数据的分析技术：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；提出连接图像模型与信息扩散处理思想，精确定义基于非渐进扩散模型的主动学习问题，构建迭代贪婪算法 MinSS 解决该最小原数据集问题；提出社交网络环境下三角关联行为模式的挖掘算法，来预测是否有三人将形成动态网络中的一个封闭的三角关联。

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和磁盘阵列，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 编码以及用于异构纠删码存储集群的负载感知恢复方案 LARS。设计采用流水线式编码来提升纠删码数据归档性能，将链式分散机制引入 Mirrored RAID-5 和三副本冗余组，分别设计[D+P]cd 和[3X]cd 两种数据布局，在此基础上设计了两种流水线式纠删码归档方案。

研究以固体盘、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 HAND 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）；设计开发基于神威处理器的全

国产固态盘阵列。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。设计高性能磁盘阵列方法 ThinRAID。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 14 篇，其中 4 篇被 IEEE Transaction on Parallel and Distributed Systems 录用，有 1 篇被 IEEE Transaction on Computers 录用，有 1 篇被 IEEE Transaction on Dependable and Secure Computing 录用，有 1 篇论文被 International Symposium on Reliable Distributed Systems（SRDC2014）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 6 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 年年度，本课题组在分布式文件系统元数据集群、网络分簇 RAID 的高性能扩展及重构技术、广域文件系统的目录缓存技术等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 延迟的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，降低了元数据或者状态写入延迟与状态规模相关性；提出并实现了一种异步两阶段提交协议，使得分布式文件系统元数据操作的平均响应时间降低

了 30%。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。本年度并进一步开展了 IO 性能、重构性能和扩展性能的优化工作，并开始进行支持纠删码的高可靠存储系统的设计。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效果时，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 6 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。课题开展了以下几方面工作：

围绕“兼顾服务成本与服务质量”这个主题，我们针对低成本的资源分配机制展开研究，提出了分布式求解算法 DREAM（-L），实现高效的资源配置；同时，我们发现多个并发应用经常会使用同一份数据。因此，我们设计出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，大大节省了数据对云端稀有资源（如缓存）的占用。为了保证应用程序的数据在用户端和云端平滑转移，我们设计了一种普适

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。能够极大减少云端和用户端同步数据的开销。

针对云存储设备 SSD,我们设计了基于页面差异性感知的 SSD 纠错码，提高 SSD 性能和使用寿命，提出了弱化 SSD 缓存系统中的纠错码，减少解码开销，提高缓存性能；对于读密集型的应用，性能提升甚至达到了 60%。该系统还减少了固态盘用于垃圾回收的时间(20%左右)。更重要的是，在各种不同闪存错误率(RBER)条件下，该系统的性能仍然比普通的固态盘性能平均高出 17%。

我们还研究了云存储数据的分析处理，包括三维对象特征提取、文本数据的多维匹配，为高效数据组织提供借鉴；安全方面， 研究了自安全的数据共享方案，降低了云端的安全管理开销并消除了性能瓶颈。

今年课题有 2 项重要成果，在理论研究方面，提出的“对等网络数据共享系统设计的理论与方法”荣获 2014 高校科学研究优秀成果奖（自然科学二等奖）. 该成果提出一套完整的模型，准确的计算系统服务质量和所需要的维护开销之间的关系，寻求最优的平衡。依托本项目研发的“混合云存储系统关键技术”荣获湖北省科技进步一等奖，本成果从系统结构与资源调度、数据组织、安全保障三个方面开展了研究，提出一种混合云存储系统结构，对云存储资源调度建立数学模型以保障云存储服务质量。

课题今年发表 10 篇论文，包括著名国际会议 INFOCOM，IMC，ATC，HPDC 等，获得授权专利 4 项，申请 6 项。获得两项云存储设备国家标准获批立项。

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，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。根据专家意见，面向最终项目结题，模拟面向数

字城市应用的时空数据访问负载，形成大规模并发访问负载，逼近真实的访问负载，接近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，课题组提出一种基于动态计算相关度的大数据布局方案，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统计数据的点云聚类压缩算法。通过对空间统计数据的空间数据信息映射成空间点云，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，将各阶段数据处理特点与内存数据库、关系型数据库和分布式数据库不同类别数据库性能优势相结合，通过的分布存储和协同调度，满足 GIS 实时应用的限时性。利用内存数据库建立了“不落盘”的传感器数据流式接入处理机制，支持实时数据流的在线处理；利用关系型数据库建立了关系完整性约束和结构化数据高频更新的主题数据库，支持用户决策；利用 NoSQL 数据库建立持久化存储海量非结构化传感器等归档数据。针对 GIS 实时分析是传感数据和历史数据的综合分析，课题组设计了支持存算联动的实时数据内外存混合索引，在考虑内外存调度方式和数据粒度的差异性基础上，建立分

而治之的管理内存/外存全局索引，以原子级的对象管理技术保证并发访问中的数据一致性，并通过自适应的缓存算法，提高高频访问数据在内存中的命中率，以满足内外存统一管理的需要，保证 GIS 综合分析中新旧数据的无缝调度。