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

[2012] 01 号 总第 6 期

项目办公室　编　　2012 年 7 月 2 日

项目研究进展

第一课题

根据 973 项目的任务和目标，课题一截止 2012 上半年开展了融合存储与泛在服务部分关键技术的研究。取得一些研究成果，在 IEEE Transactions on Computers、IEEE Transactions on Parallel and Distributed Systems、the 7th IEEE International Conference on Networking, Architecture, and Storage 、The 21st ACM International Conference on Information and Knowledge Management、The 26th IEEE International Parallel & Distributed Processing Symposium（IPDPS）等重要存储领域会议或期刊上发表，共发表论文 12 篇，新申请发明专利 5 项。

（1）构建仿真模型和物理验证系统，分析固态盘内部各层次并行性的使用方法，研究各层次并行性的最佳使用方法，以及多个并行性之间的使用优先级，给出多层次并行性利用的设计原则。研究了 SSD 的四种并行性，以及提高并行性的调度策略。四层并行操作的最佳优先级是：通道间并行优于晶圆间并行，晶圆间并行优于分组间并行，分组间并行优于芯片间并行。对于同一通道中的多个任务，为了充分发挥晶圆之间和分组之间的并行需要将任务重新调度，常用的调度策略有以下两种：多分组重调度（Multi-plane rescheduling）和流水线调度（Multi-plane rescheduling）。多分组重调度（Multi-plane rescheduling）将同一 die 中的多个 Plane 上同类型的操作尽量放在一起。这样可以使用 Multi-plane 操作。流水线调度（Multi-plane rescheduling）将同一分组上的同一类型操作尽量放在一起。这样就可充分利用流水操作。在通道任务调度时可以采

用读优先策略、最短队列优先策略、最短估计时间优先策略提高设备性能。以上研究成果已发表 1 篇 ACM ICS，1 篇 IEEE TC 论文。

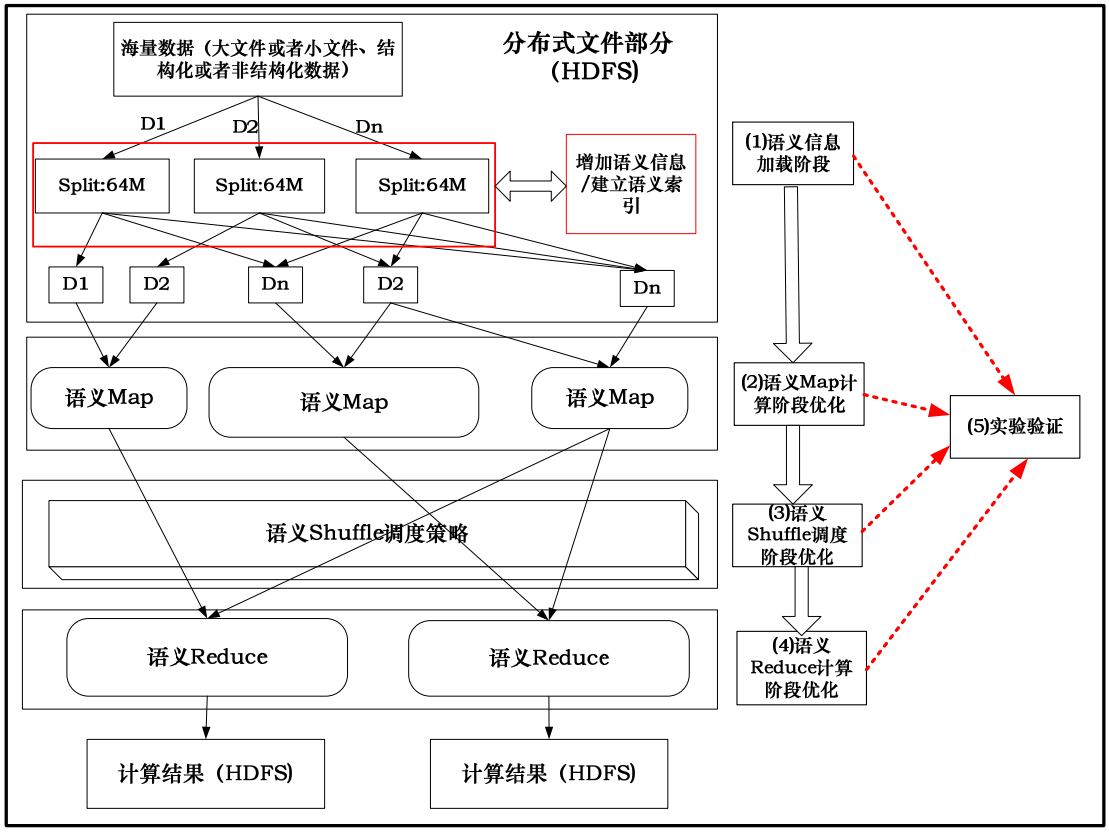
（2）课题提出了一种新的分布式语义感知的元数据组织系统 SmartStore，优化了存储系统的数据组织结构，通过分析工具获取文件数据之间的相关性度量，从而将相关性较强的文件聚集在一个或邻近的组内，聚集分组方法能够提高组内文件之间的语义相关性，支持大规模存储系统中面向文件名的点查询和复杂查询服务，有效实现文件的定位、查询、添加和删除等功能，通过扩展文件系统和用户之间的交互接口，提高系统的可用性、灵活性和可扩展性。同时，为了有效地支持近似查询服务，提出局部性特征灵敏 LSBF(Locality Sensitive Bloom Filter)结构设计，通过用局部性特征灵敏的哈希函数来实现，使得 Bloom filter 不仅能够支持传统点查询，也能支持近似查询服务。在 LSBF 结构设计中，提出了位矢量和主动溢出的方法来验证文件多个属性值的一致性，降低误报率。实验结果表明，与 LSB 树和 SmartStore 进行比较，LSBF 在查询精确度方面平均提高了 11.36%和 17.92%，查询延迟降低了 36.28%和 47.51%，节省了大约 90%和 77%的空间负载。以上研究成果已发表 2 篇 IEEE TPDS,1 篇 IEEE TC 论文。

（3）现有重复数据删除主要有基于局部性的重复数据删除算法和基于相似性的重复数据删除算法，挖掘备份数据流的局部性增加了 RAM 的利用率，减少访问磁盘上的索引，从而减轻重复数据删除的磁盘瓶颈，或者利用数据备份流的相似性，通过提取备份流相似特征减少 RAM 的使用。但这些算法存在很大限制，局部性算法经常在数据流局部性欠缺的时候性能很差，相似性算法经常在相似性很弱的情况错失掉大量的重复数据。课题提出的一种相似性方法，将局部性和相似性进行互补利用，从而提高重复数据删除的整体性能。该方法下，避免 segment 的所有指纹都放入内存中，而仅仅只需要保存代表指纹到内存即可。比如系统设定一个 segment 平均 2MB 大小，设置重复数据删除平均分块大小 8KB，每个指纹的索引开销 60B，那么重复数据删除 1TB 的数据，就只需要 30MB 的内存开销，仅仅使用了以前的全局内存索引表方案的 1/250 内存开销。而重复数据删除的局部性算法，通过对存储系统中的备份数据流局部性的挖掘并缓存数据的局部性到内存中，可以对相似性方法进行补充和加强，从而能够找到更多重复数据；而且因为数据流的局部性缓存到内存中，可以避免频繁访问磁盘索引，提高缓存命中率。以上研究成果发表在 USENIX ATC 上。

第二课题

2012 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 Journal of Computational Information Systems（EI 源）、WAIM2012、WISA2012、 SSS2012 等国际期刊与会议上录用或发表了 7 篇论文（EI）；申请专利 1 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)提出了一套基于语义的海量信息处理编程模型 MapReduce++：



针对现有海量信息处理模型缺乏语义的特征，提出了一种基于语义的海量信息处理模型，内容包括：（a）数据块语义索引信息、语义索引结构的描述及其语义索引信息自动加载方法。通过自动加载算法，对存储在 HDFS 上的海量数据块增加相应的语义索引信息，为 MapReduce 在计算处理过程中增加智能；（b）对 Map 与 Reduce 阶段的各种语义操作(语义排序、语义分组、语义合并及其语义查询等)，给出大表划分方法及其基于语义的大表 Join 连接方法；（c）基于语义的 Shuffle 调度策略。Shuffle 调度一直是 MapReduce 的主要瓶颈，通过研究基于语义的 Shuffle 调度策略及其算法，提高 Shuffle 的智能调度能力。本部分的研究成果，能较好地地优化 MapReduce，增强其处理效率和

处理能力，为海量数据的智能处理提供理论和技术积累。有 2 篇论文发表在国际期刊 Journal of Computational Information Systems 上。

（2）针对地理信息系统中海量小文件的管理与处理面临的问题：提出了一种基于 HDFS 的、结合地理信息系统中海量小文件空间相关特点的一种文件打包方法。实验表明，该方法在对地理信息系统中海量小文件的处理上具有较为明显的优势。有 1 篇论文录用在 WAIM2012 国际会议上。

（3）针对海量数据的分析与挖掘：我们研究了云环境下的程序设计方法，并且定义了一种面向海量信息处理的数据挖掘语言，能够将各种完成任务所需的 Map 任务或者 Reduce 任务进行连接或者组合，完成用户所需的各种复杂的服务。本部分内容申请了 1 个发明专利，录用了 1 篇论文在 WISA2012 上面。

（4）在海量数据处理的应用方面：我们研究了一种针对海量数字仓储的资源推荐方法 GARSS。该方法针对海量数字仓储资源，通过增加标注信息，以较好地实现推荐功能，论文发表在 SSS2012 国际会议上。针对海量心理学图书文献信息，我们构建了一种合著关系网络，对其中的合著关系提出了高效的分布式并行分析方法，论文录用在 WISA2012 上。针对海量文本信息，我们提出了双哈希表模式下的 Top-K 查询机制，通过实验表明，我们的算法比单哈希表的查询机制具有较好的性能，论文录用在 WISA2012 上。

第三课题

研究了大规模存储系统运行过程，分析其能量消耗的根源，提出针对性的解决方案。对于存储系统其功率消耗主要分为两个方面，其一是维护系统运行所消耗的能量，其二是针对数据存取过程所消耗的能量。如果系统处于空闲状态时，第一部分能量在理论上是白白浪费了，甚至在低负载情况下，这一部分能耗效率也是很低的。针对上述两个方面，目前重点从以下四个方面对于大规模存储系统的高效能进行了研究。

针对减小低负载下的维持功耗，目前提出两种解决方法，第一种是设计高能效存储节点，基于 Marvell DB-88F6282-A1 芯片设计具有 4 个磁盘接口和两个千兆接口的低能耗存储节点，在使用 SSD 驱动器情况下，其工作能耗小于 8 瓦；第二种称之为 MES（多引擎存储系统）也就是保障请求响应时间的前提下在低负载时使用较低配置（能耗也很低）的计算部件，在强负载下使用高配置计算部件，系统能够根据负载变化自动进行平滑的切换，能够节约能耗 40%。

其次在设计低能耗节点的同时，考虑其系统的软硬件设计能够利用可再生能源（目前主要是太阳能）的进行供电。在硬件结构上，设计太阳能供电系统，保证系统能够在太阳能环境下尽量长的工作；在软件结构上，保证能够在系统电源波动情况下的持续可靠的工作；最后，把标准存储协议（iSCSI 和 SAMBA）移植到低能耗存储节点上。

再次研究设计新型存储结构，在保证可靠性和性能的前提下，减低运行时能耗。这包括弹性磁盘阵列 ERAID，利用单个磁盘上的空闲空间进行磁盘间冗余；LDF 利用存储节点的计算能力，通过优化校验过程，提高纠删码性能；DROP 划分磁盘阵列中的特定区作为 Cache 区，根据每个服务器对阵列的负载进行动态的调，从而将热点数据从原来的分散分布改为集中分布，减少了磁头的移动距离；设计 PERAID 从 RAID 磁盘中划出一部分作为一个虚拟的写缓冲区，把许多随机小写合并成大写的方式再下发到磁盘中，避免了经典的小写问题并且能够延长空闲磁盘的休眠时间；针对纠删码集群的更新写，根据大小写特性分别提出 DUM、PUM，及其适用于任意更新粒度的混合式更新方案（DUM+PUM-P）。

最后针对大数据应用环境，研究重复数据删除和云系统配置优化技术。重删重点研究海量存储系统中的重删原理及其优化技术；提出基于公有云存储平台的具备鲁棒性的决策支持和成本估算框架；提出了 pCloud 算法能够根据访问模式制定合理的资源分配，实现自适应分布式 I/O 资源调度，提高系统的 I/O 资源利用率和收益。

半年内，已投稿论文 3 篇，发表论文 6 篇，申请发明专利 3 项，申请软件著作权 2 项。

第四课题

中国科学院计算技术研究所数据存储研究中心在 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“存储服务关键支撑技术”的研究任务。在课题负责人许鲁研究员的带领下，成立了四个研究小组分别在存储服务的有效区隔和 QoS 保障机制、资源使用和数据访问模型、可叠加的网络文件系统以及存储和数据的动态模型四个方面展开研究工作。

目前，课题组正在开发和测试蓝鲸 BW-RAID 系统、存储缓存集群系统 MFC 和可叠加的网络文件系统三个原型系统，用以验证本课题组所研究的各项关键技术。其中，蓝鲸 BW-RAID 系统是针对通用存储设备之间的存储和数据资源的动态组织和调度，设计的一种可线性扩展的高可靠性的网络 RAID 技术，用以能够容忍存储设备节点的故障，同时和三副本冗余技术相比，存储空间有效利用率可提高 1 倍以上。存储缓存集群系统

MFC 针对集中存储环境中缓存集群管理问题，在保证数据的一致性前提下，增强集中存储系统的可用性和可扩展性，其平均性能开销仅为 8.4%。可叠加的网络文件系统是通过一系列核心技术，建立可叠加的文件系统架构，实现了对上可提供多种标准客户端访问接口；对下可兼容使用多种已有存储设备资源，保护用户已有投资，发挥已有存储设备性能；对内能够以模块形式建立面向不同类型应用的可优化模块和策略组合，用以满足多样化应用需求。针对文件系统元数据访问性能的瓶颈问题，设计完成了可扩展、高可用、均衡化的元数据服务器集群。其采用动态 Referral 技术和创新的动态元数据自动均衡分布机制；同时，支持元数据服务器的横向在线扩展。经测试，元数据访问性能达到近线性扩展，在 5 个元数据服务器环境下，元数据访问吞吐率达到每秒 23,000 个文件创建。此外，本课题实现了高效的元数据高可用技术，达到数秒级的失效恢复，支持元数据服务透明接替。已通过包含各种故障注入的 72 小时压力测试，平均故障切换恢复时间为 8 秒。

在 2011-2012 年，在知识产权方面共完成 8 项发明专利的申请，发表 6 篇学术论文。

第五课题

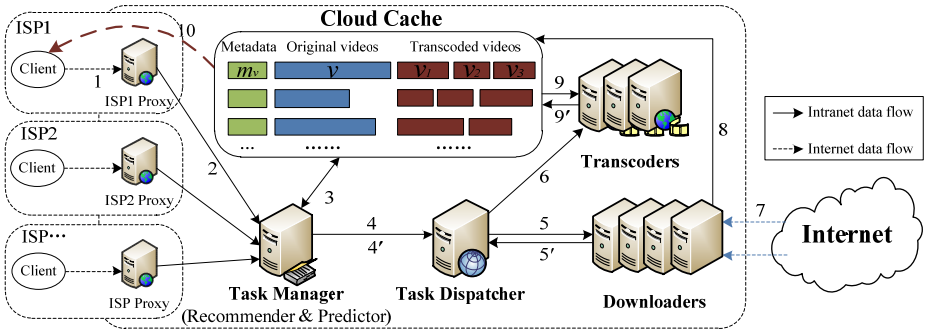
2011 年本课题的工作集中在给出云存储系统原型；在此基础上，2012 年本课题的主要工作集中在原型系统的测试和利用原型优化现有的云存储相关应用。通过真实的、大规模的数据负载驱动，和在实际应用中部署，改进用户协作的相关的理论和机制，为基于协作的云存储服务的质量优化提供理论基础和实际展示。

课题组本年度在外延云存储架构和保障机制方面开展了如下工作：

（1）利用社会化网络的大规模数据驱动测试外延云存储架构，通过使系统自动感知用户数据的局部性，提出了自适应的数据划分和存放机制，保证系统的并行读能力。

（2）研究混合式 CloudP2P 内容分发的“带宽放大效应”，通过分析“QQ 旋风”系统的大规模真实测量数据，找出了的关键影响因素，从而设计了一个细粒度的性能模型 FIFA 算法来合理分配有限的带宽，使得系统总的“带宽放大效应”最大。

（3）提出了基于协作云存储的云转码系统，利用协作局部性优化系统的缓存和存储性能，并和腾讯联合实现和部署，使得系统在实际中得到了大规模的应用。目前系统每天大约收到一万的用户请求。



“云转码”系统架构

（4）提出和实现了增加系统并行写能力的快速写和索引更新机制，将数据快速并行写入磁盘，然后再将索引更新到系统，同时，数据的读和写采用流水线方式，保证系统的高效性。

上述工作发表在 NOSSDAV'12，IWQoS'12，ACM Transactions on Internet Technology，ACM Transactions on Multimedia Computing Communications and Applications 。

同时，课题组还研究了云存储系统应对大规模用户并发访问的存储安全访问控制机制：

（5）提出一种基于属性的云存储系统的安全访问控制方法，解决现有的安全访问方法存在的密钥管理和分发复杂问题，以防止外部攻击者对存储系统的侵入。该方法采用 CP-ABE 密码算法，通过明确的属性定义来描述用户、文件和访问权限。其中，用户信息、文件信息和文件操作类型均以属性值集合的形式来表示；文件的访问控制以访问控制字符串的形式表示，访问控制字符串由文件属主制定；属性密钥由客户端负责生成，由认证端负责管理和分发。认证端使用用户属性、文件属性和文件操作类型，通过文件的访问控制字符串判断用户是否具有文件操作类型的权限；认证端为具有权限的用户分发属性密钥。该研究内容已申请两项国家发明专利（公示中）。

（6）提出一种针对 SSD 的块级多版本数据保护方法 BVSSD，通过跟踪 SSD 动态状态的历史变化并利用闪存不能覆盖写的特征来实现连续数据保护功能。与现有的连续数据保护相比，BVSSD 具有轻量级、性能影响小、容易实现等特点，更重要的是它基本不需要对上层文件系统和应用做改变。使用企业级负载 trace 的仿真实验表明 BVSSD 只有 3%到 8%的性能下降。同时，考虑到当前 SSD 的用途以及其容量不断增长的趋势，BVSSD 将具有实际的应用价值，即能保证有实际意义的连续数据保护窗口。该研究内容发表在 the 5th annual international system and storage conference(systor 2012)上。

第六课题

本阶段，针对课题定位和任务计划安排，围绕三维数字城市数据自动生成、空间环境和灾害监测应用和典型数字城市小文件的云存储理论和方法进行探索，所取得的研究进展主要体现在以下几个方面：

（1）研究了基于混合三维 R 树的大规模点云数据组织方法，利用基于八叉树和三维 R 树的混合空间索引结构 3DOR 树层次性地组织点云数据，基于节点的视距，实现大规模点云数据的近实时索引构建并保证良好的存储利用率；同时提出一种复杂三维建筑物内部构件几何及其拓扑关系和语义的多尺度一致性综合方法，同时考虑模型的语义约束，保证多尺度综合下 LoD 模型的外观相似性和语义一致性；以及构件模型的连接特性，使得综合后的构件模型能够保持结构上的正确性，为大规模三维城市模型数据的多级 LOD 数据自动生成提供了技术支持。

（2）基于湖北省水环境遥感监测系统，研究了环境小卫星 CCD 影像的水体提取方法和利用环境小卫星 CCD 数据的水质参量反演方法；同时，利用时序遥感影像和面向对象的方法，实现了大范围的雾灾检测；利用非线性 NDSI 模型进行积雪覆盖率的反演。针对数字城市空间小数据特征，提出一种结合 RDBMS 和 Hadoop 云存储的两级存储方法，实现空间结构化小文件数据的高效存储和管理。

（3）针对示范系统，面对专业应用，进行了多源多分辨率遥感影像的搜集和处理，包括 QuickBird、HJ-1A/B 和 MODIS、FY 等高中低分辨率的遥感影像；初步建立湖北省水环境遥感监测系统，支持辐射定标、几何校正等功能，实现了数据的自动化处理；同时初步完成了三维数字城市演示模型；面对公众应用，研究了数据的可视化技术和人机交互新技术，开发出基于视觉的增强现实技术 C-nerve，增强现实系统的可操作性和用户体验；并对前期以广域分布的典型环境监测数据如空气质量、环境噪声等及大数据量的连续视频数据为对象进行数据的实时采集、存储与处理示范应用数字城市原型系统进行了完善和应用拓展。