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

[2011] 04 号 总第 4 期

项目办公室　编　　2011 年 12 月 8 日

年度项目进展

课题一：融合存储体系结构与服务架构研究

根据 973 项目的任务和研究目标，本年度课题一“融合存储体系结构与服务架构研究”主要围绕融合存储系统的组织、面向复杂应用环境的按需服务以及真实系统构建示范等三个方面开展相关研究工作，取得了较好的研究成果。

第一，在融合存储系统方面，主要从器件级、存储节点级以及存储系统级开展面向融合系统的研究工作，注重研究基于异构器件、混合介质和数据等系统环境提供具有高性能、高可靠性、高可用性、可扩展、智能性的融合存储系统的基本理论和相关的关键技术。由传统的磁盘存储器件和新型非易失性半导体存储器件进行融合，形成自治的融合存储节点，节点本身具有很强的智能性，实现节点内的存储资源调度与自动适配，达到最优的存储配置。在一个存储区域内部，多个自治的融合存储节点间通过高速互联通道相互连接，实现存储节点间的协作，为提供更安全、可靠、高性能的存储服务奠定基础。在多个存储区域间，通过互联网络，可以实现融合协作，为云存储服务提供技术支持。

第二，在复杂应用环境存储服务研究方面，注重研究面向复杂应用环境中基于异构平台虚拟化技术的基础设施的实现机制，以及提供具有服务质量保证的按需服务的方法，在时间(虚拟机)和空间上(重复数据删除)的性能优化方法和实现技术。存储系统需要面对复杂的应用环境，不同的应用环境在物理设备、运行软件、操作流程、数据布局和服务需求等方面都有较大差别，因此，需要提供具有服务质量保证的按需服务来满足服务应用在高性能、高可靠性、可扩展等方面的要求。通过存储虚拟化将这些异构的物理存储设备以存储池的方式聚合在一起。上层应用将自身的存储需求发送给

存储虚拟化软件，这些需求包括性能、容量、可靠性等方面。同时，按需服务能够灵活地分配物理空间，便于存储虚拟化技术来提供超过存储池的物理空间的逻辑空间给上层应用，并且根据上层应用的使用需求动态分配物理空间，提高存储设备利用率。

第三，在系统构建示范方面，为验证面向复杂应用环境的融合存储系统理论和关键技术的可用性和可靠性，构建了数据共享服务平台和海量数据备份服务系统作为应用示范来评测相关的理论和方法。

相关研究成果发表在国际期刊 IEEE Transactions on Computers (TC)、IEEE Transactions on Parallel and Distributed Systems (TPDS)、ACM Transactions on Storage (TOS)、以及国际学术会议 Proceedings of The USENIX Annual Technical Conference (USENIX ATC-2011)、Proceedings of the 25th International Conference on Supercomputing (ICS-2011)、Proceedings of the 25th IEEE International Parallel and Distributed Processing Symposium (IPDPS-2011)等，同时也申请了 1 项国际专利、10 项国内专利以及 3 项软件著作版权。“基于主动对象的海量存储系统与技术”获湖北省技术发明一等奖。

课题二：海量数据组织与资源共享的方法研究

2011 年研究主要集中在海量非结构化数据的组织与共享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。经过这一年的研究,取得了一些成果，在 BIBE2011,WISA2011(最佳论文)，COMPSACW2011,计算机研究与发展及其通信学报等国际会议与期刊上发表了 10 篇论文；申请专利 5 项；研制一个海量结构化数据处理平台华鼎-C，研发并搭建了一个海量非结构化数据存储管理平台。课题组成功组织并举办了 ICADL2011 国际会议；同时邢春晓作为程序委员会主席成功组织了 WISA2011 会议。

(1)提出了一套海量非结构化数据的组织与共享模型：针对海量非结构化文件（含大文件与小文件）的管理和处理特点，建立数据组织模型。该模型为了兼顾处理好大文件设计了一个文件比较算法。若为大文件，则将这些非结构化大文件直接存储到 THCFS 分布式文件系统中，若为小文件，则将这些非结构化小文件存储到 THCDB 云数据库中。在该模型中，用户可以通过三种方法对海量非结构化文件进行计算和处理。1）可以直接通过海量数据驱动的应用程序进行处理，2）也可以编写类似于 Hive 的简单或者复杂分析计划 Rabbit，然后由 Rabbit 转换成 MapReduce++程序处理机制对海量非结构化数据进行处理。3）当然也可以直接编写语义性能比 MapReduce 更强的 MapReduce++程序对海量非结构化数据进行各种处理和分析工作。本研究针对现有框架 MapReduce、Hadoop++、Haloop、Twister 及其 Spark 等在 Shuffle 阶段存在的瓶颈，改进了 Shuffle 阶段的处理算法，设计了一个新的互联网海量数据处理框架 MapReduce++,同时实现了

大表之间的 Join 连接算法，初步实验表明新的框架将较大提高处理效率及其查询精度。论文发表在计算机研究与发展等期刊及其投稿在 ApWeb2012 上。

（2）针对海量文件的管理与索引：提出了一种处理海量 GIS 小文件的分布式文件管理方法及一种基于 HPK 算法的元数据索引机制。针对互联网应用的主流分布式文件系统仅适用于大文件管理的局限性，提出一种用于管理海量地理空间数据中小文件的方法，并对其存储空间利用率和访问性能进行了优化。实验结果表明，改进后的算法与原有的处理方法无论在读海量小文件还是写海量小文件上，速度均得到了明显提高。论文发表在通信学报等期刊上，申请了 1 个发明专利，同时投稿论文在 IPDPS2012 等上。此外，基于 Fedora 与 HDFS 实现了数字对象管理服务 DaaS（Digital Object Repository as a Service），为现有基于数字对象仓储管理的数字图书馆应用移植到云环境提供关键一步，论文投稿在 JCDL2012 上。

（3）针对海量数据的分析与挖掘：针对多个领域的应用分别取得了一定的研究成果。针对海量数据的数据挖掘方面，联合美国及其日本方面的研究者，设计了一套有效的海量关联规则描述模型及其挖掘算法和机制，该算法运用于生物医药的智能处理上，取得了较好的效果，论文发表在 BIBE2011 及其 WISA2011（最佳论文）上，同时申请了 2 个发明专利。为了较好分析海量数据，我们将海量非结构化的 Twitter 的 1000 多万条 Tweets 数据及其新浪微博 1000 多万条数据应用于基于复杂社会网络挖掘中的信息推荐中，分析比较了各种算法的效果。

（4）平台方面：开发了针对海量结构化数据的华鼎—C 平台及其针对海量非结构化数据的海量分布式文件存储平台华鼎—S 平台。华鼎—C 平台对易程苏州研究院 80,000 个站点共包含 29,200,000 条记录进行了有效测试，站点模型计算结果表明，计算 20,000 个站点数据的平均值只需 5 秒，与以前的解决方案比较，极大地提高了计算速度。项目组已经成功搭建了华鼎—S 平台，作为一种大规模的文件共享存储系统，它面向海量文件数据、高并发访问的环境。在性能上能达到近线性增长。如： 1）聚合带宽随着存储集群规模扩大而线性增长；2）充分发挥硬件性能，如网卡性能利用率达到了 80%。华鼎系列平台申请专利 2 项。

（5）国际学术交流与组织方面：2011 年 10 月 24 日至 27 日，主办了主题为“数字图书馆——文化传承，知识传播，未来创新”的第 13 界亚太数字图书馆国际会议（ICADL2011）。来自国内外二十四个国家和地区的一百多位学者出席了本次会议。本课题的负责人邢春晓为本次大会的程序委员会主席（Program Committee Chair）、本课题的骨干成员周立柱教授为本次大会主席（General Chair）、张勇为本次大会的组委会本地

主席（Local Chair）、李超为本次大会的研讨会主席（Workshop Chair）。第八届 Web 信息系统与应用会议 WISA2011 于 2011 年 11 月 4 日—6 日召开。邢春晓教授作为本次大会的程序委员会主席，成功组织了 WISA2011 主会及其两个研讨会 SWON2011、EGTA2011 的各项工作。两个会议由 Springer 及 IEEE CPS 分别出版了论文集。

课题三：高效能存储系统组建方法研究

本年度的工作整体分为三个部分，第一部分偏重基础测量及其实验方法研究，期望获取第一手的真实数据，为后面的研究打下坚实的基础；第二部分是构建基本系统，其中包括低能耗存储节点，高效能大规模存储系统等，为后续研究构建基本的原型平台；第三部分针对具有的问题提出新的解决方案，以提高系统的整体效能。

在测试方面，首先建立动态能耗测试和监控系统，能够以 1 毫秒采样率并发测试 64 个采样点的电流、功率和温度等参数，同时能够触发 Traces，Benchmark 和应用程序发生实际的 workload 到目标系统；其次分别测试磁盘、处理器和服务器的实际工作能耗、性能和温度等；最后分别从嵌入式低能耗存储系统构建与测试，不同配置情况下服务器和文件系统的性能与能耗研究与测试，磁盘阵列系统的能耗优化技术三个方面展开研究，初步构建低能耗存储设备，分析和研究硬件配置、系统软件配置、应用负载和服务器能耗之间的关系，为下一步构建相关分析模式和优化方法打下基础，其中磁盘阵列能够节能 40%。

在原型系统构建方面，针对面向复杂应用环境的大规模数据存储系统特点，分别研究和构建面向文件的分布式文件系统、面向卷的分级存储系统和面向应用的键值存储系统；分级存储系统实现基于逻辑卷的多节点间动态数据迁移，系统根据前台负载的变化自动完成数据在系统中的重分布，在保证系统性能与可用性的前提下，降低系统整体的能耗。分布式文件系统从能在文件级实现性能与能耗随着负载的变化而自动调节之外，还能实现数据的高效能组织和管理工作；针对高效存储设备，在分析现有硬件能效基础之上，开发采用嵌入式加固态盘的高效能存储系统，其能耗效率超过 1.0MB/焦耳。

在关键优化技术方面，研究多用户情况下的数据精简技术，包括自动精简配置和重复数据删除技术，最大化的提高存储设备的利用率；研究存储系统的动态扩容研究，在保证前台响应性能的情况下，其性能提高了一倍；针对磁盘实效情况，提出基于失效盘优先（VDF）的新高速缓存替换算法，能够同时提高重构和前台服务性能；提出了基于 K 步数据隐示信息的多级 Cache 算法：Hint-K。它利用数据块多次升级或降级的历史信息，简单快速地得出该数据块的活跃度（热度）；针对 RAID6 校验块写开销大问题，提

出了新的 RAID-6 编码：H-Code，由于每行末端的数据块和下一行初端的数据块共享相同的斜向校验块，因此能有效解决跨行写的问题并提高存储系统的性能；为了综合横向编码和斜向编码的优点，提出了一种新的横向编码：HDP 编码，HDP 编码将横向校验块平均分布在每个磁盘上，从而实现较好的 I/O 负载平衡，减少超过 30%的双盘恢复时间；提出基于最优路径的 RAID6 重构 PDRS 方法。

本年度研究成果在国际学术会议 ATC2011，DSN2011，IPDPS2011，NAS2011 等上发表；并且 1 篇在 IEEE Transaction on Computers 上发表；另外已投出 IEEE Transaction on Computers 和 IEEE Transaction on Parallel and Distribution 各两篇，向 ISCA，EuroSys，SIGMetrics 各投稿一篇，向 FAST 投稿两篇，前期申请专利共 11 项，获得二项授权。

课题四：存储服务关键支撑技术研究

中国科学院计算技术研究所存储中心在 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“存储服务关键支撑技术”的研究任务。课题组在课题负责人许鲁研究员的带领下，成立了分别以卜庆忠博士、刘振军副研究员、沈玉良博士、张军伟博士为负责人的研究小组。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，课题组正在开发和构建蓝鲸 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，用以展示和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，提高系统的可用性。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。本系统采用缓存节点之间相互镜像的方式维护缓存集群的可用性。针对其中存在的共享缓存管理问题，使用缓存分区进行应用区分，并提出了 FAME 方法来解决各分区间缓存容量的动态调节。典型负载测试表明，相对于同步更新（缓存可靠，性能基线），负载的响应时间平均降低 68.35%、请求吸收率平均提高 32.53%、读响应时间平均降低 19.31%；平均性能可以达到异步更新（缓存不可靠，性能上限）的 91.6%。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存储访问接口，如：NFS、pNFS、CIFS 等多种标准客户端访问接口；对下可兼容使用块/卷存储设备、本地文件系统存储以及标准网络文件系统等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。

在 2011 年，在知识产权方面共完成 6 项发明专利的申请，发表 5 篇学术论文。

课题五：云存储服务和保障机制研究

云存储作为未来信息存储的重要服务模式，吸引了大批研究人员来关注云存储的发展。很多研究集中在和数据中心相关的问题上，却忽略了用户到数据中心之间的互联网空间上存在的问题对云存储服务质量产生的影响。例如：由于互联网在复杂环境下性能的不可预测性，导致海量数据传输时间代价巨大，没有可靠性保证；数据密集型访问会造成可用性和效率的显著下降；此外，随着信息的不断增多，对存储容量的要求越来越大，不仅云存储服务基础设置的硬件成本越来越高，用户的存储费用也随之增加。已有的对相关问题的研究往往针对一个个孤立的问题，而本申请则把这类问题和云存储服务质量结合起来，在整体的构架下探讨云存储外延服务质量问题。

本课题最终目的是提出高效的＂云存储服务和保障机制＂，重点研究在复杂的网络环境下用户协作的效能评价及其相关的理论和方法；研究旨在提高云存储服务的可用性和可靠性的混合型云存储新型构架；研究基于协作的云存储外延服务的质量优化问题，研究云存储服务的安全保障体系，从而达到经济、高效、灵活、方便的使用云存储服务的目的。

2011 年本课题的主要工作集中在给出云存储系统架构设计及可行性分析；提出服务目标量化指标，提出用户协作的效能评价及其相关的理论和方法，为下一步研究基于协作的云存储服务的质量优化提供理论基础。

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) 提出外延云存储系统架构设计，并通过深化上一期 973 的成果 AmazingStore 系统，实现了外延云存储服务模型，用系统运行的实际数据验证可行性和有效性；

(2) 测量和建模了云存储外延的节点的异构性，为用户协作的效能评价及其相关的理论和方法提供建模和设计依据；

(3) 依据异构性模型，提出稳定性最优的外延节点分组存储方案，使系统在稳定性与可扩展性之间保持良好的权衡；

(4) 给出节点预期带宽能力的预测模型，为设计外延云存储数据分发的保障和调度机制提供基础；

(5) 提出和实现了云下载方案 VideoCloud，通过使用云设施来保证数据健康度和提升数据传输率，从而提供高质量的外延空间上的内容分发服务。同时能够聚合多个下载到单个节点上，节约了外延云存储的电力消耗。

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：

(1) 提出一种基于身份的两阶段分布式认证机制，与具有中心决策支持的非集中式访问控制机制相结合，提高了海量存储系统的访问控制效率，为云存储系统应对大规模用户并发访问提供了一种高效的存储安全访问控制机制。

(2) 提出一种流式重复数据检测方法，通过构建计时型布隆过滤器阵列 TBFA，在滑动窗口模型内灵活高效地检测重复数据，提高了云存储系统重复数据查询的空间效率。

(3) 提出了一种在 SSD 内部实现连续数据保护(CDP)的方法 ShiftFlash，其目标是使得 SSD 自动具有连续数据保护功能。

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，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. 面向数字城市的多源时空数据存储与管理，针对异构融合存储体系的静态或动态信息存储与访问要求，研究数字城市中的多源空间数据特点，分析数字城市地形、地物等静态离散对象、地面车辆等动态离散对象以及气象要素、电磁场等动态连续对象特征，提出一种多维时空要素及其相互关系统一表示的概念模型，描述时空环境中各种对象的几何、语义、物理和行为等方面的属性和演化规律，为时空环境信息模型标准化，以及各类应用的数据共享、互操作奠定基础；同时，面向数字城市多源时空数据存储与管理方法，研究了粒度与结构统一的多层次三维城市模型数据组织方法，提出一种高效的磁盘、内存、显存三级数据存储粒度与结构一致的数据组织方法，减少 I/O 次数，提高调度效率。

2. 基于数字城市的跨媒体信息共享与服务，结合虚拟现实、三维地形漫游，研究公众信息服务实现方法。首先通过计量经济学模型模拟演练，研究广域网公众用户的行为特征，提出用户行为预测模型；同时通过多源、多时相、多分辨率的遥感影像，对大雾、冰凌等灾害进行研究，建立数字城市大气环境灾害理论模型；基于用户访问行为规律的表达和计算，提出一种空间数据访问流行度的集群缓存预取模型。该模型基于 Zipf 分布来表征和计算用户对瓦片长期稳定的访问分布规律和基于 Markov 模型来表征用户

对瓦片访问的时空局部变化规律，构建了能适应用户对瓦片访问的长期流行特征和短期流行特征的集群缓存预取模型，提高公众信息服务能力。

3. 面向数字城市的实时跨媒体信息存储与公众服务系统要求，构建反映异构融合和泛在服务理论与技术的原型系统，基本完成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的构建方案，以及三维 GIS 可视化工具软件 3DViewer 原型系统研制。在空间数据收集、加工、整理方面，完成了部分 MODIS 原始遥感数据、FY 卫星原始数据、环境减灾小卫星 CCD 数据，以及水温、叶绿素等各类指标数据，为下一阶段的应用示范实验提供数据基础。