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] 02 号 总第 2 期

项目办公室　编　　2011 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，课题一上半年开展了融合存储与泛在服务的前期调研和部分关键技术的研究，包括 FLASH+DISK 混合存储系统的调研和建模、提高数据中心存储服务能力的优化技术、互联网应用服务原型系统的构建等。取得一些研究成果，在 MSST2011、IPDPS 2011、USENIX ATC 2011、ICS 2011、 IEEE Transactions on Parallel and Distributed Systems (TPDS)、 IEEE Transactions on Computers 等重要存储领域会议或期刊上发表，发表 18 篇论文。

（1）开发多层次固态盘模拟器—SSDsim。利用模拟器分析固态盘内部的多个行为对性能的影响，通过实验得到以下结论：1）在多种不同负载环境下，大页会影响固态盘的性能；2）不同的分配策略有不同的部署环境，对于每种负载可以找到一种最优的分配策略；3）虽然在一些情况下，高级命令能够提高性能，但是如果不恰当地使用这些高级命令将影响固态盘的性能与寿命；4）固态盘的并行包括通道级、芯片级、晶圆级以及平面级，他们的使用优先级与高级命令、分配策略紧密相关，并且对性能寿命有明显的影响。

（2）提出一种基于光纤通道的远程访问零拷贝机制，通过虚拟文件接口访问远程 SCSI 磁盘设备，利用 RDMA 机制减少了传输过程中的数据拷贝，提高了传输速度和硬件利用率；提出了一种新型的虚拟机资源调度策略，动态调整各虚拟机带宽分配，实现

存储资源的充分利用，并减少 I/O 访问的延迟，实现 SLO（Service Level Objective）中资源的分配保证和带宽延迟保证。

（3）提出了一种基于日志的 RAID6 磁盘阵列结构 RAID6L。RAID6L 的阵列结构由 RAID6 磁盘阵列外加一个日志盘组成。RAID6L 优化 RAID6 磁盘阵列写性能的基本思想是，在写负载较重时停止校验块的更新，等到负载变轻后再使校验条带恢复到同步状态。同时，为了使 RAID6 磁盘阵列在校验不同步的状态下仍然具有容双盘出错的能力，RAID6L 在处理写请求时将相关数据块的初始数据或更新数据记录到日志盘中，以保证发生磁盘失效后所有丢失的数据块都能够顺利恢复。基于马尔科夫状态转换模型的系统可靠性分析显示，RAID6L 系统的平均数据丢失时间与传统的 RAID6 磁盘阵列相比有所降低，但对原型系统的实验评估表明，RAID6L 系统的写性能相比于传统 RAID6 磁盘阵列有较大大幅度的提升。同时提出一种推广 Parity Logging 方案的方法，使得 Parity Logging 方案能够运用于任何形式的磁盘阵列。从理论和实验两方面对 Parity Logging 方案和 RAID6L 方案进行了对比，证明了 RAID6L 方案比 Parity Logging 方案对 RAID6 磁盘阵列写性能的优化作用更为明显。

（4）标准形式下的容双错编码通常都具有码长限制，极大的影响了容双错垂直码在实际存储系统中的应用。为此，提出了两种适合于容双错垂直码的码长扩展算法，这两种算法均能将标准的容双错垂直码扩展至任意码长。两种码长扩展算法各有自己的优势和缺陷。变更校验块方式的码长扩展算法生成的扩展垂直码是 MDS 编码，但其更新复杂度和编解码计算复杂度都比标准垂直码要高；移除校验条带方式的码长扩展算法生成的扩展垂直码不再是 MDS 编码，但其更新复杂度保持了最优值 2，并且其编解码计算复杂度比标准垂直码更低。

（5）提出一种近乎精确的重复数据删除系统 SiLo，能够有效利用备份数据流中的相似性和局部性达到很高的重复消除率、极低的内存开销和高吞吐量。现有大多数最先进重复数据删除方法或者是基于局部性的算法，或者是基于相似性的算法。而根据我们的分析，这些算法在许多情况下失效。前者在数据集缺乏局部性的条件下，重复数据删除吞吐量性能很差；后者在相似性缺乏的条件下，重复数据删除往往错过了消除大量的冗余数据。SiLo 的主要思想是挖掘和利用合并相关的小文件和分割大文件从而产生更多的相似性，并充分利用备份数据流的局部性来补充重复数据删除的相似性检测。通过相似性算法与局部性算法的结合，SiLo 能够大大减少 RAM 的索引，并且可以达到很高的重复数据删除吞吐量。SiLo 的实验结果显示，SiLo 系统整体性能优于目前的主流重复数据删除系统。

（6）针对 WAN 带宽较低，云备份的备份和恢复时间太长，而现有重删技术只考虑减少备份时间而忽略了恢复时间的问题，提出了基于重复数据删除中的因果关系、可改善云备份和云恢复两方面性能的 CABdedupe 方案。利用三个关键部件：文件监控器（File Monitors），文件列表（File List）和文件谱（File Recipe），首先获取和保存数据集各时间版本之间的因果关系，研究因果关系的信息来快速识别修改的文件和数据块，通过删除每个备份或恢复操作传输过程中未修改过的数据来提高备份/恢复的性能。CABdedupe 可集成到任何现存的备份系统中的中间件。把 CABdedupe 集成到两个现有备份系统中并用真实的数据集测试，结果表明：备份时间和恢复时间减少的比例高达 103:1。

（7）构建两个互联网应用服务的原型系统：U-Stor 互联网文件服务、“晒点”学术交流社区网平台。研制并构建了聚合带宽达到 60GB/s 以上的分布式后端存储系统。

第二课题

围绕海量数据的组织与资源共享展开研究，在复杂应用环境下海量数据资源管理框架与方法，基于分布式文件系统的海量地理空间数据存储及索引技术，数据多副本管理等方面取得阶段性成果。主要取得以下进展。

（1）提出了一种云环境下海量数据资源管理框架，包括物理存储层、海量存储网、数据转换层、数据管理层、安全管理层、资源组合层和应用层，该框架可以实现对云环境下的异构海量资源统一管理和组织。基于该框架提出了一种云环境下海量数据资源管理方法，该方法包含云环境下的数据资源管理与共享的各个方面，主要有副本动态管理与迁移、负载均衡、数据溯源、用户资源使用方式及资源组合等。并针对海量信息处理提出了一个技术架构，包括：结构化、半结构化及非结构化数据与资源的对应管理方法，能兼顾海量大/小数据文件处理的分布式文件系统，分布式列存储的云数据库系统，查询语义能力介于 MapReduce 和 SQL 之间的一种分布式查询计算方法。为云环境下的海量、异构数据组织和共享设计了一套完整的解决方案。

相关的研究成果主要包括：已受理发明专利一项（专利申请号：201110147807.4），第二届云计算与 SaaS 会议（C4S2）已录用论文一篇(EI 检索)，WISA2011 已录用论文一篇(EI 检索、优秀论文、拟推荐到计算机学报发表)，已投稿 ICADL2011 国际会议和第 17 届全国信息存储技术学术会议各一篇。

（2）提出了一种基于分布式文件系统的海量地理空间数据管理优化方法。该方法针对互联网应用的主流分布式文件系统（如 GFS、HDFS）仅适用于大文件管理的局限

性，提出在空间元数据的索引建立时，采取地理空间数据编码加块内索引的方式，将小文件的块内索引分布到数据节点上，以缓解元数据服务器节点的负载，并基于这种方式对小文件的存储空间利用率和访问性能进行了优化。初步的实验结果显示出该方法的有效性。

相关的研究成果主要包括：已提交发明专利一项，第二届云计算与 SaaS 会议（C4S2）已录用论文一篇(EI 检索)，已投稿 CCIS2011 国际会议一篇。

（3）对海量数据的组织和共享中所涉及到的多副本管理问题进行了较为全面的调研。在分布式海量存储系统中，为了保证高可靠性、高性能，多副本管理是其中的一项关键技术。多副本管理作为传统分布式存储系统中的关键技术，其创建方法、放置技术、选择、删除策略及一致性管理等方面的研究已日趋成熟。而目前在云环境下的多副本管理方法缺乏一个全面系统的综述。针对云环境中海量数据的组织和共享有关多副本管理问题进行了较为全面的调研，详细地分析了各种副本管理技术方案的优势与缺点，并提出了下一步的工作主要围绕基于数据访问频度的动态副本调度策略，针对同一块内小文件冷热差距的多副本调度策略。

相关的研究成果主要包括：在 NDBC2011 已录用论文一篇(拟在计算机研究与发展发表)。

（4）提出了数据对象框架中对象标识符的一种编码设计与实现 CDOI（Common Digital Object Identifier）。数据对象框架中的对象标识符用于唯一标识数据对象，以支持分布式环境下的互操作，最终达到资源共享的目的。设计和实现的 CDOI 满足以下要求：a.代表和确认数字对象，且与其物理位置、副本数量、应用协议、存储和处理要求无关；b.确认数字对象的版本编号及版本之间的联系；c.提供逻辑的数字对象与数字对象的具体物理位置的连接；d.提供数字对象与其元数据的连接。在借鉴了 DOI，cIDF 和 OpenURL 等标识系统的基础上，提出了 CDOI 设计，并实现了 CDOI 原型系统。

（5）提出了一种无共享架构的分布式列数据库系统及其实现方法。在现有的列数据库华鼎系统的基础上，为了更好的支持大规模数据组织与共享，实现了分布式处理功能。该方法将对数据请求的处理工作从代理服务器转移至数据服务节点上，使代理服务器功能简化，缓解了其对整个数据库系统的瓶颈限制；对于查询请求，任一数据服务节点均可作为委托管理节点负责管理对所述查询请求的处理工作，提高了查询效率；并通过采用基于列存储的数据库，节省了存储空间，提高了处理速度。

相关的研究成果主要包括：已提交发明专利一项、完成了原型实验开发。

第三课题

在上半年的工作中，整体分为两个部分，第一部分偏重基础测量和基本系统的构建等实验方法研究，期望获取第一手的真实数据并构建基本的原型架构，为下一步研究打下坚实的基础；第二部分针对具有的问题提出新的解决方案，以提高系统的整体性能。

在节能存储方面，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%。

在原型系统方面，分别研究和构建分级存储系统和分布式文件系统。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。

在关键技术优化方面，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；针对磁盘失效情况，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。

前期工作成果较为明显，其研究成果在国际学术会议 USENIX ATC2011，DSN2011，IPDPS2011，NAS2011，ICPP2010，Cluster2010 上发表；并且 1 篇被 IEEE Transaction on Computers 录用，待发表；另外投 IEEE Transaction on Computers 和 IEEE Transaction on Parallel and Distribution 各一篇，前期申请专利共 6 项。

第四课题

该课题组分为四个研究小组，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 系统和可叠加的网络文件系统两个原型系统，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，能够容忍存储设备节点的故障，提高系统的可用性。本课题基于带外管理方式，融合了虚拟存储、数据缓存、动态扩展、瘦预留（Thin Provisioning）、高并行等高端存储技术，有效地将多个通用网络存储设备聚合成大容量、动态可扩展的虚拟空间，同时保证存储设备间的数据可靠性。可叠加的网络文件系统是本课题组通过一系列核心技术，通过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。本课题拟在上述系统架构基础之上，建立高性能、高可靠的网络文件系统原型。

2011 年上半年，课题组在知识产权方面共完成 6 项发明专利的申请，发表 2 篇学术论文。

第五课题

该课题重点是研究在异构网络环境下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。同时，课题组还研究了云存储系统对大规模用户并发访问的存储安全访问控制机制。分别研究了利用外延节点的存储方案，以及建模数据的传输能力，同时研究了保证外延云存储的安全机制。

（1）稳定性最优的外延节点分组存储方案。由于外延节点一般极不稳定，如何利用外延提供存储服务是一个挑战。主要思想是将不稳定外延节点按照一定规则组合成“稳定节点组”，一个节点组在存储服务能力上相当于一个稳定的服务器节点，将系统中近似同构（即稳定性相近）的外延节点分到相同的组，并将分组问题形式化为“同构的最大稳定性分组问题”（H-MSG Problem），在统计模型下给出了最优理论解，得出稳

定性最优的外延节点分组存储方案，如图 1 所示。实验表明，该分组方案使系统稳定性得到大幅提升，同时容量损失是可接受的。具体工作发表在国际期刊 TPDS 上。

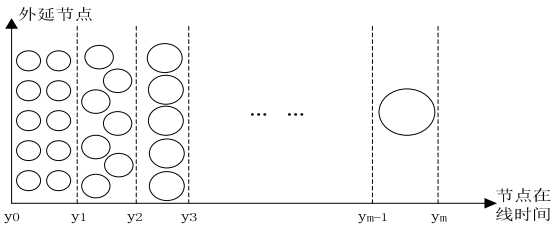

图 1 稳定性最优的 Internet 外延节点分组存储方案示意图

（2）外延存储的数据传输能力模型。为更好的利用节点的闲置带宽，辅助云端应用，需要预测节点在给定的一段时间内的数据传输能力是一项基础性的研究工作。节点的带宽受多种因素影响，波动很大，其中最主要的因素是节点的上下线行为。可以使用如图 2 的马尔科夫链描述节点的动态行为。

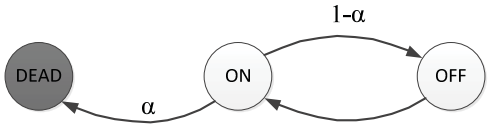

图 2 动态性模型

通过求解该模式，得到节点在未来时间 t 内期望累积在线时长 $R(t)$ 。假设被请求文件的副本存储在 m 个节点上。节点可用带宽是 u，则这些节点未来 t 时间内的数据的传输能力是 $u\sum_{i=1}^{m}R_i(t)$ 具体工作发表在国际会议 ICPP2011 上。

（3）基于身份的两阶段分布式认证机制。为了不影响系统的并发度，系统的认证机制必须能够为上百万的用户提供可扩展的认证方式。提出一种身份认证与文件服务相分离的两阶段分布式认证机制，该机制将用户身份认证与文件服务相分离，将认证过程分为身份认证和 I/O 认证两个阶段，实现了认证的可扩展性；认证服务器支持现有的任意网络认证机制，包括实现跨域标识和基于公钥密码技术的认证方法；存储节点认证用户身份只需要进行两次低开销的散列计算，实现了简单高效的 I/O 认证。

（4） 流式重复数据检测方法。通过构建一个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。本发明在滑动窗口模型下工作，对元素的监测可以精确到一个元素，从而使基于本发明的统计结果具有稳定性，另外 TBFA 中的部分计时器组可以被卸载到磁盘中，从而减少内存开销。理论分析和实验数据表明，TBFA 在加载 10%以内数据内容到内存的情况下，能够保持 95%以上的查询效率，从而使本发明在空间效率和可扩展性上优于现有技术方案。

后续研究计划包括：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；基于外延的云存储服务资源分配和调度策略；云存储系统访问控制与数据安全保障机制，重点研究云存储系统数据自安全策略、云存储系统数据公开审计机制；在云存储系统数据重复删除方面，继续研究提高重复数据检测效率的新方法。

前期研究成果发表在 ICFIT 2010、IMC2010、Grid2010、ICPP 2011、P2P Computing 2011、SCIENCE CHINA(Information Sciences)等共 9 篇文章。申请专利 3 项。

第六课题

课围绕理论探索和应用原型系统搭建开展研究，在以下几个方面开展了研究：

（1）针对数字城市中多源空间数据及实时动态信息的特点，研究设计了多维时空要素及其相互关系统一表示的概念模型。此模型能够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的标准化，以及各类应用的数据共享、互操作奠定基础。同时，针对造成大规模三维城市模型实时应用效率低下的数据调度延迟问题，课题组提出了一种高效的三级数据存储粒度策略与结构一致的数据组织方法。实验表明，该组织与存储方法能够大幅度减少 I/O 次数、提高数据调度效率，为大规模三维城市模型的实时绘制奠定了基础。

（2）为了使存储服务能更好地适配用户访问行为，基于历史访问数据开展了针对用户行为和偏好的理论分析和定量表达研究，提出了一种基于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空间数据长期稳定的访问分布规律和基于 Markov 模型来表征用户对空间数据访问的时空局部变化规律，构建了能准确适应用户对空间数据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，结合空间数据的地域相关性及多维特性动态划分和元数据访问粒度可伸缩性管理，能有效适配用户访问行为和偏好规律，提高访问效率。

（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，进行了基本存储服务系统以及以广域分布的典型环境监测数据和大数据量的连续视频数据为对象的实时数据采集、存储与处理示范应用软硬件系统设计、设备选型和系统环境搭建工作。目前已基本完成示范应用系统的设备选型、招标、采购和系统环境搭建工作，为下一阶段应用示范系统的运行、测试提供了比较完善的基础实验平台。收集、加工、整理多种数据类型的真实实验数据，包括离散小容量、大规模的大气环境监测数据和大容量、连续视频数据以及 MODIS 卫星遥感数据和环境监测、灾害监测遥感数据，融合 GIS 示范数据，有望快速形成初步的基础示范应用数据集，为下一阶段的应用示范实验提供基本的数据基础。