国家重点基础研究发展规划（973）项目

下一代互联网信息存储的组织模式和核心技术研究

项目编号：2004CB318200

简　　报

[2007] 03 号　总第 9 期

项目办公室　编　　2007 年 12 月 5 日

项目 2007 年度总结与学术研讨会议召开

2007 年 11 月 26-27 日国家重点基础研究发展计划(973 计划)项目“下一代互联网信息存储的组织模式和核心技术研究”(编号为 2004CB318200)在华中科技大学武汉光电国家实验室举行 2007 项目年度总结与学术研讨会。

项目组咨询专家强文义教授、毕光国教授和马建峰教授，湖北省科技厅王红斌处长，项目专家组专家沈绪榜院士、张以谟教授、叶朝辉院士，以及项目下属六个课题组主要学术骨干、项目依托单位华中科技大学科研主管领导、武汉光电国家实验室和华中科技大学计算机学院相关领导等 30 余人出席了会议。

会议由“项目实施工作研讨会议”和“项目工作总结和汇报”两部分组成。

2007 年 11 月 26 日召开了“项目实施工作研讨会议”，与会人员包括各课题组负责人和技术骨干。会议围绕项目总体目标之一——构建 PB 级网络存储系统及其上的 GIS 应用示范，针对项目调整方案建议，重点讨论课题之间的融合，以及如何在项目统一的平台上进行研究成果的集成。经过研讨，达成以下共识:

1、课题围绕项目总体目标开展研究，所取得的成果应通过应用示范系统得到体现。

2、各课题在项目试验平台上将研究成果进行部署与融合，对系统进行验证和完善。

2007 年 11 月 27 日“项目工作总结和汇报”在光电国家实验室会议厅举行，由项目首席科学家、华中科技大学冯丹教授主持。华中科技大学杨勇副校长、湖北省科技厅王红斌处长和武汉光电国家实验室林林书记分别致辞。他们对项目过去的研究成果给予了充分的肯定，并表示将一如既往的关心和支持项目的发展。

会议由冯丹教授进行项目总结报告，围绕项目拟解决的两个关键问题：网络存储的组织模式和 PB 级存储系统的组建理论与方法，从项目总体思路与研究目标、计划执行情况、主要进展及创新性和人才培养与学术交流情况等方面对项目年度的工作状况进行了介绍。然后各课题组汇报了课题研究工作进展及下一年度的工作计划。与会专家对项目研究相关的问题进行了深入讨论，并提出了指导性的意见和建议。

01 课题：研究分析了大量互联网环境下的存储系统的 trace，结果表明网络存储的 I/O 行为具有突发性和自相似性，且读自相似性占主导地位，指导在进行网络存储系统设计时，读优化要多于写优化。对对象存储系统的 I/O 负载进行自相似分析，发现随着存储系统节点数目的增加，自相似系数出现波动。提出基于对象的主动存储概念，扩充 T10 OSD 标准，引入方法对象及策略对象，对 OSD 命令集进行扩充，形成自组织对象存储系统这种新存储模式，从理论、方法、关键技术和系统实现多个层面研究，使系统具有自组织(Self-Organize)、自调节(Self-Adjust)、自安全(Self-Secure)、自愈(Self-Heal)特性。有关标准化建议书提交给信产部标准化组织。设计实现包含三种安全方法 CMDRSP，ALLDATA，CAPKEY 的对象存储系统的安全访问模型，CAPKEY 和 CMDRSP 下性能减少少于 5%, 元数据操作开销增长小于 5%。提出基于热度的多线程重建调度优化的算法，形成提高存储节点及系统的可靠性和可用性的自愈方法，该算法充分考虑用户访问模式，对用户热点访问区域优先重建，减少了重建所需的时间和用户访问的延迟；可根据用户访问模式提前重建好用户即将要访问（读请求）的数据，进一步缩短了用户访问延迟，在保证可靠性的前提下提高系统性能。

02 课题：在高速通道的体系结构中，赋予通道三种特征：a）智能特征。通道不仅可以根据对象的属性组织数据分布、预取与缓存数据，而且具有计算特性，实现“计算到数据的迁移”，从而使得通道可以为应用提供更好的性能支持；b）虚拟化特征。应用对通道的访问可以是透明的，不仅保证与传统数据通道相兼容，而且可以在两种通道之间实现动态的负载均衡；c）动态特征。通过建立 Jdib 通信库和对 FPGA 的网络配置机制，可以支持系统动态地调整与配置通道。为了充分利用各个节点之间空闲的内存资源，在以集群方式组建的高速通道系统中，实现并部署了一种多接口、多功能

的远程存储资源共享池（Cooperative Memory Pool）。它可以为高速通道系统提供如下扩展功能：（1）构建虚拟的网络数据通道，（2）屏蔽后端复杂的存储结构，（3）根据应用需求提供相应的存储资源分配。由于传统的关系数据库管理屏蔽了许多应用的访问特征，在高速通道的存储适配层中，采用基于文件系统的矢量数据管理系统，来支持应用访问特征在分布式文件系统上的映射。同时，通过分布式文件系统来直接访问与管理空间数据，也能够有效避免数据库系统在管理大规模空间数据时所遇到的性能瓶颈。在高速通道的应用适配层中，提出建立 Jdib 通信库，使之支持 Java 平台在 InfiniBand 网络环境中进行高效通信。

03 课题：扩展原型存储系统。扩展现有存储原型系统的规模，使其更能够体现海量存储系统的复杂性和多样性，并使其支持更大规模测试的需求。在此基础之上构建多个功能区，面向不同方向研究的工作。海量存储系统运行规律研究，以及性能分析及测试。定量理解海量网络存储系统这种复杂的大系统的运行规律和典型特征，并能够通过模型来定量或定性的描述相应的运行规律，在此基础之上建立使其长效运行机制，揭示复杂大系统能长期可靠运行的方法和规律。海量数据的组织和管理原理及机制研究。增加对于数据集的自我描述属性定义，并在系统设计过程中充分利用这些属性，解决海量数据快速查找和存取，支持海量存储系统可伸缩性扩展；分析海量存储系统多样性负载特征，确定合理的负载分析和性能预测方案。海量存储系统中关键技术研究。针对海量存储系统数据的高可用性需求，提出并实现了分散式体系结构的高可靠文件存储系统；为提高海量数据安全性，建立一种适应存储系统结构的安全存储服务架构；为了提供存储系统中海量数据的连续数据保护功能并优化其性能，研究并实现一种连续数据保护机制。

04 课题：研究典型 P2P 系统的节点行为规律。对等模式下对象的组织需要以构成对等系统的节点的行为特征为基础。研究精确的数据对象可用性分析模型。数据可用性是存储系统的基本属性，由于对等模式下节点行为差异大，因此现在广泛使用的平均可用性模型会带来较大误差：既无法保证数据的目标可用性，也浪费系统资源。研究高效的数据对象修复策略。P2P 存储系统中的节点错误完全不同于经典存储系统，不符合经典的“fail-stop”模型。其中最难的一个问题是系统含有大量的临时错误，而系统本身又无法清楚的辨别一个错误是临时还是永久的。如何高效的修复数据，是当前国际上研究的一个热点。研究 P2P 系统中的信誉机制。此项研究的目的是保证 P2P 系统的良性状态，尤其是存储应用对 P2P 系统的稳定性要求非常高，因此从用户行为入手，发现不良行为，建立稳定用户群。

05 课题：所研制的高速系统数据备份技术基于多种策略的管理模型进

行任务管理，可以根据应用的特点灵活调整。在融合虚拟存储技术的基础上将备份数据流与任务控制流有效分离，为系统客户提供高吞吐率和高扩展能力的数据备份，并能够大大提高系统可靠性。提出一种新型的数据恢复策略——后端融合技术，实现只做增量备份（只有第一次备份为全量）、按需恢复、减少数据存储量、提高可靠性和系统资源利用率。

针对虚拟存储的特点，主要面向系统的可靠性和性能的评价方法进行了初步的研究和探索。（1）可靠性评价：将磁盘阵列的分析方法应用到网络存储系统上，整个系统被划分为双层冗余模型，下层是设备内部的磁盘冗余，上层是设备之间的设备冗余，从全局角度研究系统的可靠性。（2）性能评价：采用在线分析系统状态和负载类型特征，建立一种自适应调节机制的性能评价体系。整个评价体系由一系列方法构成，各种方法联合运行，完成系统的自适应评价和调节。

针对现有的条带卷扩容方法中元数据写操作和数据读写都很频繁的缺点，研究了条带卷快速扩容方法，从而通过增加条带卷中的磁盘数既能扩大存储容量又可提高 I/O 性能。针对大规模网络存储系统中，跨越系统崩溃或断电事故后仍旧可用、庞大遗留系统零合并时间等问题，研究了面向大规模网络存储系统的带外虚拟管理系统。

针对基于磁盘控制器级的远程镜像技术须专用磁盘控制器、费用昂贵且通用性灵活性差，而基于主机驱动程序级的远程镜像技术需修改所有前端主机的驱动程序，系统可扩展性差等不足，提出了一种基于 I/O 节点的驱动程序级远程镜像技术，并设计实现了一个远程镜像系统，同时基于此，正在进一步研究一种基于存储区域网络 SAN 的卷级别目标器端远程镜像系统。针对现有共享加密文件系统的密钥管理方面难以同时满足安全性、灵活性和高效性的需求，研究了加密文件系统的密钥管理。

06 课题：研究了一种可扩展的基于双重集群的多媒体网络 GIS 体系结构，采用 LVS 服务器集群解决空间数据服务器提供网络服务带宽瓶颈，采用基于对象存储集群解决 PB 级海量空间数据存储 I/O 带宽瓶颈，并实现了一种大型多媒体网络 GIS 系统原型 GlobeSIGht。进一步完善了 GlobeSIGht 的功能，优化了系统的数据管理方式。采用多层次可扩展的地球空间信息海量数据组织模式。进一步研究了基于对象存储的空间数据存储模型，初步建立了一个适合空间数据的空间数据对象存储模型，将空间数据组织成空间存储对象存储在基于对象存储设备中，并设计了该模型中空间存储对象访问接口。研究了一种基于矢量基对的地形存储对象分布策略，地形数据及其属性数据组织成地形存储对象存储在基于对象存储集群中，地形存储对象在多个基于对象存储设备中的分布采用基于矢量基对的地形存储对象分布策略。

听取了各课题汇报后，项目咨询专家强文义教授、毕光国教授和马建峰教授，项目专家组专家沈绪榜院士、张以谟教授、叶朝辉院士对项目研究工作进行了指导，主要内容包括:

1、肯定了项目的工作，认为各课题的体系结构很清晰，研究内容很丰富，取得了许多的成果，系统中闪现出了一些亮点，项目年度研究任务完成的较好。

2、针对了项目的现状，对未来工作提出了指导性意见和中肯的建议，主要包括:

- **作为国家重点基础研究发展计划，通过研究，项目能够提出什么样的重要理论。**指出项目可适当加大在理论研究上的力度，突出在基础研究方面的创新；争取能够提出标志性的理论成果，在国际上能得到认同，并产生重大的影响。
- **项目研究的技术核心要突出并具系统性。**指出项目研究涉及的内容很多，围绕项目拟解决的两个关键问题，技术核心要凸显，以增强整个研究工作的系统性。

此纪要呈报:

科技部基础司 973 计划办公室

教育部科技司基础处

湖北省科技厅基础处

各课题组所在单位科技处

项目组咨询专家：钱华林研究员、强文义教授、毕光国教授、马建峰教授

项目专家组专家：沈绪榜院士、张以谟教授、叶朝辉院士

973 项目(2004CB318200)办公室

2007 年 12 月 5 日