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] 02 号　总第 8 期

项目办公室　编　　2007 年 9 月 25 日

项目研究进展

围绕调整后的课题任务方案，针对中期检查中存在的不足，改造现有的实验平台，以进一步增强课题间的合作。各课题进行深入研究，已取得了一定的进展。

大规模网络存储实验平台的改造。

从去年 12 月到今年 5 月间，对位于武汉光电国家实验室的实验平台中高性能存储区域进行了扩充，增设了 8 台高性能存储节点，形成了 480Gb 背板交换及不少于 10 端口的 Infiniband 连接能力，已经配置好 8 个 Infiniband 端口，2*20Gb、6*10Gb，增加了 64*300GB 共 19.2TB 的存储容量。所扩充的高性能存储节点将主要用于华中科技大学承担的课题一（对象存储系统）和中科院计算所承担的课题二（高速通道）合作开展的 Infiniband 实验环境改进和高性能对象存储系统的实验。同时在规模上也对实验平台进行了扩展，增加了 24 个节点，该部分的扩展主要用于课题三针对大规模网络存储系统的研究。针对中期检查期间显现的因外部可访问 IP 地址不足而影响子项目间的远程合作的问题，光电国家实验室正计划重新调整优化实验区域网络结构，增加 C 类地址以满足科研需求。

课题一

1、对象存储系统的研究。在存储节点的部署、对象存储设备的设计、元数据的管理等方面取得了进展。针对地理信息系统 GIS 的应用需求，改进对象存储了系统的设计结构，实现了通过对父目录的对象 ID 和文件名进行哈希来组织和管理元数据的方法。在对象存储设备 OSD 的研究方面，完成了智能 OSD 的结构设计，新结构将能够构建智能的、自组织的存储集群。将 OSD 上对象的元数据以类似数据库的方式组织起来并置于 NVRAM 上从而进行高效的“增删改查”操作，以实现海量存储系统中高效率的查询操作。完成元数据和对象管理对外统一接口的概要设计，为元数据管理提供两种可用接口形式，即 GET_ATTRIBUTE/ SET_ATTRIBUTE 接口和供检索的接口。针对 EXT2 文件系统索引效率低的状况，提出并实现用 B+树来进行节点的索引。

2、高效 OBS 控制器研究。国内外同类研究工作基本上都是在现有硬件设备上，在文件系统层次实现对对象存储的支持。围绕缩短存储路径的思路，开展了 OBS 控制器的芯片级设计，以在硬件层次上更好的支持对象存储体系结构；并将可重构技术用于存储协议适配功能，对于某些在体积/能耗方面有苛刻要求的关键性应用，可重构技术能使占用有限空间的 FPGA 芯片实现多种功能，将其应用于存储协议适配，有效地解决多种存储协议相互转换的问题。

3、面向下一代互联网的广域网存储研究。主要包括广域网存储资源管理和存储服务。在存储资源管理方面，提出了一种新型的索引算法 APIX，提高了广域网存储资源信息查询系统的效率，初步测试性能高于当前相关工作中具代表性的 APEX；结合 Chord 实现目录代理（DA）间的信息交换，以满足高度动态的广域网环境中 DA 因不确定因素导致失效、连接不稳定、传输性能难以保证的情况下，广域网存储资源管理系统的有效运行，提高广域网存储资源管理系统的可靠性。在广域网存储服务方面，对支持 IPv4/IPv6 的 U-Stor 进行了版本升级。针对当前 Windows 网络文件系统 CIFS 及 Linux/UNIX 网络文件系统 NFS 不能很好的适应广域网下复杂的网络环

境，提出了一种新型广域网文件系统 UstorFS。经测试 UstorFS 实现了对普通应用的正常支持。

课题二

针对本年度的 3 项任务展开研究工作，进展较为顺利。

在学术论文方面，截止 8 月份，本课题组共发表 7 篇学术论文，其中 5 篇论文被 ICPP、IPCCC 和 ISPA 等领域知名国际会议收录，另外 2 篇期刊文章中有一篇在《计算机学报》上发表。此外还有 2 篇文章已被相关期刊录用，拟于 08 年发表，同时有 3 篇文章正在审稿中。在研究生培养方面，上半年课题组有 1 名博士生与 1 名硕士生顺利通过学位答辩。在与课题其他协作单位交流方面，上半年课题组派人到华中科技大学进行了一次工作交流，统一了集成调试环境的平台环境，进一步明确了工作思路。

以下是本年度上半年研究工作的具体内容和进展。

首先，研究支持基于对象存储的高速通道的原理与实现方法，以 InfiniBand 与 Ethernet 混合存储网络为基础，提出利用 RDMA 传输机制提高数据对象交互性能的方法，进而提高系统整体 I/O 性能，并探讨基于 RDMA 传统机制优化系统的负载均衡与数据分布的机制；其次，研究海量数据存储网络的多样化、动态的数据通道接口，提出将应用的部分计算任务下放到存储节点来完成，提高存储系统的有效带宽；最后，研究动态调节与重构多虚拟通道，支持存储对象的高效网络传输，提出了利用 FPGA 的部分动态可重构技术，实现数据通道自主适应运行环境的策略，并支持通过网络远端进行配置。

同时，针对后两年的研究目标，一些工作已经开展并取得了一些成果。具体包括开发了基于 RDMA 的 Java 集群消息通信库，下一步将基于该通信库开发与平台无关的自组织数据通道中间件；另外，面向空间应用数据存储的特征，研究了矢量数据的存储管理方法，建立了 VDMS 的原型系统，可支持未来进一步面向应用特征提供自组织数据通道的研究工作。

在下半年，将继续完善以上研究工作，并加强专利申请与软件著作权登记。基于当前工作的进展，预期将很好地完成本年度的工作任务。

课题三

上半年度，我们主要从以下三方面展开研究:

1、用属性管理的方法提升存储系统服务质量（Qoss）。首先使用形式化的方法描述存储系统在不同层面的各种属性，重点是采用 zipf 模型形式化地描述了存储系统中访问热点这个属性；然后基于 OSDT10 和 ISCSI 的相关标准实现对象属性在系统中传递，使处于存储系统最上层的应用与处于系统最底层的设备之间能高效地共享重要的属性信息；最终通过搭建属性管理原型系统 Amss，使 Amss 原型系统能够通过存储对象传递“数据访问倾斜度”属性，并利用热度值这一属性实现相应的副本分布以及其他存储策略。

2、从系统结构和访问控制角度出发，研究安全性灵活的可扩展安全存储系统。针对扩展性，从存储系统结构的层面提出一种服务独立的安全存储服务框架，作为可扩展存储安全系统的参考模型；针对安全灵活性，深入研究权能标识机制，获得灵活的安全策略，并设计出相应的可扩展安全访问协议。

3、研究集群存储系统的高性能可靠性容错机制。该机制采用小型低密度校验码对文件进行编码，然后连同原始数据均匀分布在存储节点中，在恢复数据时采用解码来进行数据重构。在此基础上实现了一种改进的 CDP 机制，即通过引入少量的数据快照信息，既保证了空间和轻量级系统开销的特点，还具有快速数据恢复、防止恢复链条失效及多路径恢复等优点。

课题四

本课题要研究的两个主要问题是提高对等网络中存储数据的可用性和提高对等网络数据检索的效率。针对这两个问题，我们拟定了用工程发现研究、深入研究和验证研究的思路。在一个以对等模式组建的高可用存储系统为基础的桌面备份应用系统两年运行经验的基础上，同时采纳专家的意见，本年度的主要工作集中在对课题涉及的可用性问题进行深入的理论研究，并取得了突破性成果，文章都发表在重要的国际会议和国内一级期刊。

1、 P2P 存储研究的难点。首先，P2P 存储系统运行环境的动态性特征没有被完全认识清楚，节点的动态性本身还没有得到很好的认识。其次，P2P 存储系统中的节点错误完全不同于经典存储系统,使得其需要新型的屏蔽错误的方案。再有，有些系统运行环境的动态性本身极高，在其上实现存储功能有较大难度。最后，由于没有足够的测量数据，很多方案的有效性难于验证。目前对数据高可用性的分析模型和解决方案都还处于起步阶段，研究也表明已有的方案还不能很好的适用于 P2P 存储的复杂运行环境。

2、 理论研究思路。我们首先通过测量的方式了解 P2P 系统实际运行环境的动态性特征，包括节点的在线率、会话时间的规律以及高动态节点的特征等；在动态性规律的基础上，研究分析模型以分析系统屏蔽暂时错误所需的数据冗余度，同时研究存储节点的选择方案；同样基于动态性规律，研究检测系统中节点永久离开的方法；我们还将研究与冗余编码相结合的适用于 P2P 存储环境的安全加密方案；最终将应用上述研究成果，实现一个开放式 P2P 存储系统。

3、节点动态性研究近展。首先，我们测量研究各种 P2P 存储系统运行平台中节点动态性情况。因为所要研究的高可用数据存储方案需要以系统中节点动态性特征为基础，故我们收集包括 PlanetLab、Maze 和微软企业内部桌面机在内的 3 个 P2P 系统运行环境的运行日志，并主要对节点的动态性进行测量。在测量中，我们既关心节点可用性等经典动态性指标，同时更关注节点会话时间及离线时间的分布情况、节点的永久离开率等新动态性指标，因为这些指标是后续方案可否应用的关键。我们通过对系统长期的动态性测量，分析系统演化过程中极高高动态节点的行为特征并寻找方法避免高动态节点对系统的影响。我们比较多个系统的测量结果，分析不同类型系统中动态性的异同。本部分的研究已经发表在 P2P 方向顶级会议 IPTPS’07 上。

4、数据的可用性分析模型和数据分发策略研究进展。我们在假设没有节点永久离开的情况下研究如何根据数据的目标可用性评估数据所需要的冗余度，以及在哪些节点上放置这些冗余的数据。在不考虑节点永久离开时，存储数据的可用性是和其多个副本所存储的节点的在线规律相关的。

我们首先分析时间因素对存储数据可用性的影响，并提出时间相关可用性概念。对于数据的可用性，采用随机过程的方法建模，分析数据副本数与其可用性之间的关系。我们对随机过程模型和经典的数据可用性计算模型做深入的对比分析，并据此说明时间相关可用性概念与经典可用性概念的一致性与区别。评估出需要的冗余度后，进而研究将冗余数据存放在哪些节点集合上才能满足目标可用性要求，并通过实际系统日志上的实验对比本文方法与其它方法的优劣，我们还提出在实际系统中应用所提方案时可能碰到的问题及相关的解决方案。本部分的研究已经发表在 P2P 方向顶级会议 P2P Computing'07 上。

5、数据的概率丢失模型和数据修复策略研究进展。在屏蔽了暂时错误对数据可用性的影响后，我们进一步研究如何屏蔽永久错误对数据可用性的影响。我们首先提出对永久错误的判别方案，也即如何区分节点永久离开与暂时离开。经典的判别方案是使用时间阈值，也即超过此时间阈值的离线被视为永久离开，否则被视为暂时离开。我们认为人工配置时间阈值非常困难，且时间阈值方案不能在高动态系统中使误判率（误判是指将暂时离开判断为永久离开）和漏判率（漏判是指将永久离开判断为暂时离开）同时很小。我们提出一种基于概率的判别方案，其思想是让漏判和误判互相抵消。在判别方案的基础上，本文提出系统整体的数据修复方案。根据实际系统日志，本文验证上述判别方案和数据修复方案的可行性及有效性，并将本文方案与基于最优时间阈值的修复方案的效果进行比较。本部分的研究成果已经向 INFOCOM'08 投稿。

6、信誉模型。信誉问题是近年来 P2P 系统中受到极大关注的问题，无论是共享系统还是存储系统，都需要一种可信的环境来保证系统的可用性。我们在这个方向上的研究思路是，从 P2P 系统信誉机制的角度，激励用户的良性行为，惩罚恶意行为；从可信计算的角度研究对可信系统的攻击行为的分析与检测；从存储安全性的角度研究数据的可靠存储的方法。其中，攻击行为的检测的研究成果已经发表到国际知名会议 ICDCS'07 和 IPTPS'07 上。

通过上述各项工作，本课题组已经取得了部分研究成果，其中包括已受

权专利一项，授权号 ： ZL2005 1 0002915.7，专利名称: 基于数据分块冗余和虚拟化存储的在线备份方法;2007年共发表由本项目支持的论文10篇。

课题五

针对课题调整方案，围绕存储服务和质量进行了以下几方面的研究。

1、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。针对目前备份策略的不足，结合数据的演化特点，本数据恢复策略提出了相应的解决方法: 只做增量备份，备份数据最少; 通过增大版本粒度，减少重复数据; 通过在后端对数据的按需融合，完成信息生命周期的管理，提供灵活的恢复方式。对用户而言，备份是简单的，恢复是按需的，而只做增量备份的方式显著地减少了备份数据量，提高了资源利用率。

2、提出了设备级备份系统的按需恢复方法。在现有的设备级备份系统上，提出一种支持按需恢复的机制——虚拟恢复技术。通过将虚拟设备的实现集成到备份系统中可以完成设备级、块级和文件级的恢复。并针对虚拟存储的特点，主要针对系统的可靠性和性能的评价方法进行了研究。

3、研究了加密文件系统中的基于组密钥服务器（GKS）的密钥管理机制，并在 Lustre 上实现了基于 GKS 的加密文件系统。主要工作包括: 1）提出了基于 GKS 的集中密钥管理模型，优点是密钥由 GKS 专门保存并从不离开 GKS，增强了系统被攻击时的安全性，并减少了用户密钥管理的负担。2）通过使用锁盒子和访问控制块减少 GKS 的计算和存储需求，从而有可能以硬件方法确保 GKS 的安全性，进而保证整个系统的安全性，同时便于在 GKS 上实施灵活的访问控制。3）提出一种高效的用户权限撤销机制，采用了以文件块为粒度的密钥管理策略，并结合密钥版本技术，有效降低了权限撤销的开销。4）在 Lustre 上实现了 GKS-CFS 的系统原型并进行了测试。测试结果表明由于避免使用了公钥密码算法，加密引入的额外开销可以接受。5)提出了一种易于实现的高容错高存储利用率的编码方法，以适应大规模磁盘存储系统发展的需要。

课题六

根据项目中期验收后课题六调整方案的计划研究内容，课题组迅速调整各项工作的重点和人员配置，使各项研究内容得以顺利展开，在基于对象存储的空间数据模型、空间数据的收集、示范应用系统的进一步研究和网络 GIS 系统性能测试等研究工作上取得了一定进展，并加强了专利成果的申请工作。具体情况如下:

在基于对象存储的空间数据存储模型方面，我们一方面对支持空间数据库引擎技术的数据库（包括 Oracle Spatial、PostgreSQL 等）进行了研究，以期使用这一数据库技术能够将空间数据和属性数据有机地集成起来，并在此基础上，建立起有效的空间索引，实现对空间数据的有效存储和管理，能够进行查询和各种分析操作。另一方面，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，对 Client 读流程进行了初步分析。

加强海量地形数据、中高分辨率影像等重要空间数据的收集、加工和整理。首先对 2006 年 1 月至 12 月的 MODIS 数据进行了加工和处理，对全球范围地形数据的 1～7 层数据进行了校验和整理，并开始收集 8～12 层的地形数据。同时，采集处理了一定数量的多媒体视频和监控数据。

进行示范应用系统的进一步研究与设计。对 GlobeSIGht 中的三维空间信息的可视化和插件机制也进行了一些研究，空间信息的可视化方面主要从三维模型的可视化和 LOD 模型两方面展开了研究。针对系统的应用层、服务层、传输层和存储层中可能存在的系统瓶颈，对 GIS 服务器群的负载均衡和调度算法进行了初步研究。目前阶段，在存储层方面，搭建了基于对象存储的集群实验系统，对系统的性能进行测试和分析，并与传统的网络存储系统进行了比较，掌握了一些系统测试和分析方法。

2007 年上半年新发表论文 10 余篇，新增 EI 收录论文 5 篇，申请专利 3 项。