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

[2006] 02 号　总第 6 期

项目办公室　编　　2006 年 09 月 05 日

项目研究进展

第一课题组

围绕对象存储系统进行深入研究，在柔性元数据服务器、主动存储服务机制、自组织对象存储系统组织模式、高性能存储系统构成方法及核心技术等方面取得阶段性进展。

1、提出并完善自组织对象存储系统概念和方法。研究对象存储系统主动服务机制，以对象存储系统为平台，扩展对象的定义，即将现有对象包含数据和属性，扩展为包含数据、属性和触发器，通过形成方法库和策略池，当用户访问对象时，触发器触发方法执行，使存储系统主动服务于不同应用环境。具体方法包括借助对象属性记录分析和预测网络上用户数据消费行为、数据访问趋势等，提出一种基于统计、预测和反馈的系统主动调整方法；采用反馈信息评估系统调整是否满足用户或系统需求，并通过机器学习，实现策略的自我更新。每一对象存储设备具有主动服务能力，通过网络互连构成系统时，进一步使其具有自组织特性，系统内部根据外界应用变化进行数据布局、负载均衡等调整，使系统总体性能达到最优。

2、面向可靠性、可扩展性进一步完善元数据服务器的设计。作为整个存储系统的关键组件之一，MDS 的可靠性必须得到充分的保证，提出并实现 MDS 链式备份方案极大提高系统的可靠性。采用“柔性对象数据分布”算法，根据 OSD 的实时负载来决定对象分布，实现 OSD 的负载平衡。

3、在对象存储设备方面，研究实现了存储对象的逻辑实现与物理实现的不同映射方法，提出并实现持久对象文件系统（LOBFS），实现高效的磁盘内部数据组织；研究并提出两种对象放置策略：基于阻塞概率模型和基于响应时间模型的对象放置策略，分别适用于流媒体及事务处理两种应用对象放置策略，实现磁盘间负载均衡，从而使系统性能最优；通过扩展对象的“地址”属性以及隐目录技术，改进了 OSD 中查找对象和访问数据的速度。研究的对象存储智能型，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，研究并实现了方法对象的高效运行机制；研究并实现了 OSD 的层次化 QoS 支持，通过上层的对象请求调度、中间层的对象预处理及底层的块请求调度实现精确的存储 QoS 控制。

4、构建了可扩展、高效的对象存储文件系统 CFS：提供通用文件访问接口 (POSIX 语义接口)，将上层发送的基于文件的命令转化为基于对象的命令，与 OSD 进行数据传送，具有高的集合带宽和良好的可扩展性；提供用户安全访问管理，防止非法用户访问和避免用户非授权操作；具有高效的元数据缓存机制和自适应的更新策略；采用“基于对象的分布式数据锁”保证数据对象的一致性，实现对共享的数据和元数据对象资源的高效并发控制。与 Lustre 进行对比测试显示写性能超过 Lustre，读性能比 Lustre 略好。

5、对象存储控制器硬件设计的电路原理设计、PCB 设计及仿真、PCB 印制、元器件购买及贴片等已完成，进入 PCB 的调试阶段，目前已完成电源网络、时钟网络、复位单元电路以及 JTAG 单元电路的调试工作。实现了光纤通道协议 FC-1、FC-2 层协议功能的 IP Core，实现了千兆以太网 MAC 层协议 IP Core 的收发器部分，在验证平台上验证通过测试；设计完成符合光纤通道 FC-2、FC-3、FC-4 层协议的固件程序，正在进行与硬件的联调。

申请发明专利 3 项，文献专著一本，两个项目“基于虚拟接口的高速网络存储系统”和“一种新型并行存储系统——磁盘树”于 2006 年 6 月通过了湖北省科技厅组织的鉴定。

第二课题组

从未来网络存储系统的网络体系结构入手，提出了面向服务的层次网络存储模型，并作为指导在未来高速互连网络环境下的广域网络存储资源整合以及面向服务的存储系统架构等相关研究工作。基于面向服务的层次网络存储模型的三要素：广域网络存储资源、服务引擎和网络高速通道，目前在统一的高效存储网络通道、服务引擎两个方面展开以下研究：

1、深入研究了存储网络高速通道的新型体系结构，提出了构建广域与局域统一高效的存储网络高速通道体系的设想，为未来广域网络中各类存储资源共享服务提供一个可扩展高性能的传输通路。具体研究工作包括：(1) 可大幅降低网络存储系统通信延迟的新型高效单边操作通信机制；(2) 可支持多存储层次的动态远程存储资源共享系统。

2、研究面向服务的存储加速引擎，使之能高效汇集与分发广域范围内的异构海量数据。具体研究工作包括：(1) 高速存储通道与 IP 网络的接口模式研究以及可配置高速网关阵列的设计实现。目前在普通 PC 服务器结点上，软件实现的高速网关阵列可支持 2Gbps 带宽的数据聚合与分发；而通过专有固件优化，系统的数据聚合与分发能力将可扩展到 10Gbps 带宽；进一步采用阵列组相连方式，可望将聚合带宽提高到 120Gbps；(2) 兼容多种异构存储资源的存储加速引擎设计实现。该系统可为上层应用系统提供统一的存储访问接口。此外，基于高速通道提出消除存储资源 I/O 访问瓶颈的机制，使之能够更好地支持面向对象的存储系统；(3) 面向空间海量数据访问的服务加速引擎设计实现。基于高速通道的单边通信机制，结合海量数据空间计算的特征，设计了组件级别的细粒度负载均衡系统，以及多层次对象管理缓存机制，在服务系统和存储资源之间建立高效的资源汇聚与分发通道。

第三课题组

针对海量存储系统的构建原理、机制和优化方法的相关理论和技术进行研究。具体表现在以下几个方面:

1、海量存储系统原型系统设计和测试。原型系统具有150多个TB的存储容量，包括存储集群系统，光纤和SCSI320接口的磁盘阵列，NAS，iSCSI等多种异构存储结构和设备。所有设备通过接入的千兆网络和核心的万兆网络互联。在构建原型系统之后，进行了广泛网状的测试，测试参数包括请求大小，分区大小，结点数，存取模式等等。从中发现影响系统性能的因素，并着手解决，例如stripe大小对于性能的影响。

2、网络存储安全方向的研究现状。针对目前网络存储中的被动安全状况，借鉴生物免疫系统的内在机制：免疫系统能识别和排除“异己”、学习、记忆和模式识别的功能，将人工免疫技术引入到存储安全系统的研究中，建立一种主动的存储安全系统，并借助人工免疫建模语言对存储安全系统进行定量化评估。

3、研究集群存储系统的数据分布机制。目前集群存储系统的数据分布机制基本上采用固定分条策略。而分条映射机制基本采用HASH算法。我们目前准备使用索引机制替代HASH映射机制，设计更加灵活的海量数据分布方案。

4、进化存储系统中的数据迁移研究。研究并实现了一种满足 ESS 动态数据迁移需求的数据迁移解决方案，该方案借鉴生物学中的进化理论和控制理论中的反馈控制技术，依据 ESS 不同应用场景执行相应的数据迁移策略，并且可以根据系统状态的变化动态调整迁移策略。经过对照试验的数据分析证实，该方案能高效地实现ESS数据迁移。

5、结合多媒体应用对存储的需求。设计并实现了一个可靠的分布式高性能存储集群系统。该系统采用层次化的元数据管理保证了元数据的安全性和一致性，同时采用动态的成员管理模式管理成员节点的加入和退出。并结合第一部分工作，引入了动态迁移的机制，保证了系统的负载均衡和高I/O性能。

6、研究基于属性控制机制的分布式存储系统 Qoss（存储服务质量）问题。研究 Qoss 的两个大的方向一个是服务区分，另一个是性能保证。目前具体的研究大都集中于存储系统得自适应管理，磁盘和磁盘子系统的性能、调度机制，以及存储系统中存储资源的自动提供和自动分配等问题。目前的研究思路是首先研究存储系统的需求目标，然后通过属性来描述需求目标，并通过一定的机制根据属性将数据映射给设备。其中牵涉到的具体的技术是面向对象的存储方法，通过存储对象这个有别于块的基本存储单元，来对底层设备表达更丰富的信息，以解决服务区分和性能保证这两个问题。

基于 OSD 和 iSCSI 的相关标准，搭建了面向对象的磁盘阵列控制器原型系统，通过属性传递机制，先实现了用户指定负载特征的对象放置策略，下一步的研究目标是利用扩展其它属性的可能性，实现存储系统（目前以磁盘控制器为研究目标）动态预测及动态调整能力，为实现 Qoss 提供依据，并将研究目标扩展到海量存储系统。

第四课题组

提出了“P2P 网络环境下开放的存储体系结构”的构架，目标是将分布式存储系统中一些公共功能剥离出来，形成开放的存储支撑模块和抽象的接口，并将其和 OpenDHT 服务整合起来，为广域网上基于桌面机的存储系统的搭建提供给平台。我们最终目标是希望实现一个实用的提供海量信息存储、检索及下载的信息服务系统。具体工作进展情况如下:

1、已实现了 OpenDHT 底层 P2P 路由协议的设计和部署，其特点是可以兼顾共享、存储和协作等不同应用的需求，以标准接口的形式为上层提供传输、路由和存取服务。目前该重叠层网络已经在多台服务器上运行。

2、在广域网上的多台服务器上部署已有的 P2P 存储系统，观察和测量系统在实际运行中的网络特性和用户行为。

3、完成新的、开放的 P2P 存储体系结构的系统设计和技术方案，并对系统中的关键技术进行了深入研究，取得一些理论成果。其中包括:

(1)对于 P2P 存储系统而言，广义上讲，数据存放在第三方节点上。本课题提出了一种新的安全的编码方法，能够做到即使恶意用户得到数据碎片也无法得到完整的有意义的信息。

(2)从新的角度提出了一种存储系统可用性的定义及形式化描述的方法，引入了用户体验可用性的概念，将用户访问规律和存储结点在线规律的结合起来研究系统的可用性。

(3)根据 P2P 存储系统用户差异大的特点，建立不同的用户分组，提出了差异存储服务的概念，由此确定了数据分发策略的主要因素：(a)从可靠性考虑，需要知道节点机器的物理性能和用户的信誉，把数据存放到可信的节点上；(b)从可用性考虑，需要了解用户的访问规律和其它节点的上下线规律，一个用户的数据放置到和他的上下线规律相同的节点上，可以找回自己的数据的概率就高，可用性得到保证。

(4)研究了 P2P 系统数据检索的问题，研究的关键问题是根据语义相关性进行查询和协作，提出了一种无结构 P2P 环境下基于层次模型兴趣树的语义搜索算法，将 P2P 网络中的用户进行兴趣提取，并根据相同的兴趣（语义相关）进行分组，在分组的基础上建立分布视索引和查找。

第五课题组

关于存储服务质量方面，取得如下阶段性成果:

1、建立多维度、多层次的网络存储服务质量模型，设计了针对不同应用需求的高性能服务质量体系结构，在此体系结构基础上，提出了存储资源预定及分配算法，给出了命令队列的动态调度方法;

2、实现了存储服务质量的一个原型控制系统，在该系统上对提出的预定和分配算法以及动态调入方法进行了实际测试;

3、研究了用高性能内存虚拟磁盘 SSD 代替普通磁盘大幅度提高存储访问性能的技术，支持高达 40000 的读写 IOPS，比同等环境下的磁盘提高了 200 余倍。

4、研究了海量存储环境中自适应的虚拟存储空间的组织方法和动态调整方法、面向海量存储环境的数据副本分布技术;使用 trace 重放和 iometer 等测试工具对存储空间组织方法以及数据副本分布算法进行了测试，定量研究了数据分布对数据访问性能的改善;

5、设计并实现了一个异构环境下的存储管理系统，并对该系统进行了详尽的测试;

6、研究了对象存储网络系统特殊的数据访问模式，确定了对象存储网络中存在的安全威胁及对象存储网中数据安全的基本特征；提出了多层次的对象存储网全局安全架构；研究了对象存储网中针对对象数据和针对映射信息的加密机制，并进行了初步测试;

7、论文专利方面，发表（包括已录用）论文 19 篇，其中 SCI 检索 14 篇，EI 检索 2 篇；申请发明专利 4 项，另有 2 项获授权。

在网络存储对象化及其动态部署方面，秉承计算资源与存储资源分离的思想，研发出了服务部署系统。该系统研究、设计并实现了本课题所提出的主要思想、概念、模型、机制、方法和技术。所涵盖的部分包括虚拟存储、数据克隆、资源管理和数据备份等多个方面。本课题在多个方面已经取得了阶段性研究成果，特别是数据快速克隆技术已经成熟，可以在数秒内（超过预期指标）动态生成上百个服务并完成部署。在此基础之上，通过将各方面的研究成果集成，形成了一个相对完整的原型系统。原型系统还集成了用户认证策略，以加强系统的安全性。目前，计算资源可以通过普通网卡或者自行研制的虚拟磁盘卡动态定位网上的存储资源，可以支持 Linux、Windows 等多种主流操作系统。该系统不仅在原理上验证了原有设计的可行性，而且也明确了进一步的研究方向。

通过以上工作，我们研究并实现了计算资源通过和承载某种服务的存储资源动态绑定，构成计算环境对外提供服务，并且可以根据应用的需要在不同的服务之间灵活切换。根据研究成果和实际数据分析，我们已经开始布局和研究虚拟机技术和层次化存储在服务部署系统中的研究与设计。

第六课题组

针对目前数字地球、数字海洋、数字城市等大型GIS应用系统存在的PB级海量数据组织管理和存储存在的问题，研究多比例尺、多分辨率、多数据源空间数据管理关键技术，进行大型多媒体网络 GIS 集成系统的原型实验，在集成环境下的客户端和服务器原型开发。进一步研究空间语义存储对象、多层次空间元数据服务和 GIS 数据分布式管理的中间件，研究海量卫星遥感数据存储关键技术，研究网络环境下的多源空间数据获取机制。研究了虚拟地球可视化技术、三维地球建模与显示技术等。

在应用示范系统研究开发方面，集成各子系统功能，设计了一种支持国家对地观测需求的地球空间信息系统原型，包括客户端和服务器的设计，实现从全球-中国-湖北-武汉的由全貌到细节、由整体到局部、由低分辨率到高分辨率的快速、无缝的空间数据漫游和浏览；支持多层图象引擎和多数据集，实现全球500m分辨率的真彩色地球遥感影像观测，具备地形，海洋渲染的效果，可以动态切换图象引擎，通过调节地图月份来观察地球的季节等变化。实现全球 30m 分辨率的主要卫星遥感数据观测，研究实现了中高分辨率全球空间数据的观测和无缝漫游机制，实现了北京、武汉等城市的高分辨率影像无缝漫游。初步集成了多媒体应用功能，可实现基于GIS的多媒体监控和数字旅游服务等应用。完善海量 MODIS 卫星数据管理与在线分发服务系统。与课题一和课题三模型（系统）进行了联调测试。