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

[2008] 02 号　总第 11 期

项目办公室　编　　2008 年 12 月 20 日

项目研究进展

课题一

围绕对象存储系统展开研究，包括多层次、可扩展存储对象结构优化及其实现方法；动态环境下数据的在线优化方法；大规模网络存储系统的能耗管理和性能优化方法研究；数据生存阶段最佳放置策略的研究；研究网络存储资源无关、应用无关特性的存储资源管理体系结构。主要取得下几个方面进展。

（1）完成 OSD 行业标准讨论稿的撰写。创新性地提出方法对象，并在标准中对方法对象的定义、执行、作用空间、属性页等各方面作了详细的规定。基于国际发布的 OSD 标准版本 T10-OSD-r04，增加主动存储方面的建议，包括方法对象 CREAT、READ、WRITE 等命令及相关属性、与数据对象的关联触发机制等，提交英文版建议“OSD□3 Active Storage Enhancement”，目前正在网上进行讨论，并应邀参加 SNIA Symposium 2009 会议将于 2009 年 1 月 23 日与 OSD Technical Working Group 成员进行详细讨论。

（2）在对象存储系统的理论研究和原型系统优化过程中，实施了基于热点数据的特征分析和研究提高系统性能的方法，特别是热点数据更新的有效机制和方法。通过对海量数据访问模式的分析来获取有效的“热点”数据，并针对“热点”数据的变化来更新和替换数据，提高“热点”数据访问的效率，同时，改进“热点”数据存储的结构设计，以有效地表示和存储“热点”数据的多维属性信息，节省所需的存储空间，支持快速、准确的查询操作，降低相关操作的复杂性。

相关成果发表在著名国际会议 the 28th International Conference on Distributed Computing Systems (ICDCS '8)和 the 37th International Conference on Parallel Processing (ICPP '08)。“基于对象的智能网络存储系统”通过湖北省组织的专家鉴定，认为达到国际先进水平。

（3）围绕 OBS 高效能控制器，研究了存储节点一级节能方法，通过在镜像磁盘阵列的基础上加入一个日志盘来提高 RAID10 磁盘阵列的性能（RAID10L），能效（GRAID），可靠性（MP-RAID）。基于 I/O 负载重定向的重建优化算法，相对于现有重

建算法，大大减少出错硬盘数据重建时间，提高系统可靠性；同时降低用户 I/O 响应时间，在系统降级模式下快速响应用户请求。相关研究成果发表在 the 16th Annual Meeting of the IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS '08)和 the 13th IEEE Asia-Pacific Computer Systems Architecture Conference (ACSAC '08)。其中，题为："WorkOut: I/O Workload Outsourcing for Boosting RAID Reconstruction Performance"的论文被存储领域国际旗舰会议 The 7th USENIX Conference on File and Storage Technologies (FAST '09)录用。

（4）研究十多组典型磁盘负载中 I/O 到达间隔时间的相关性，考察相应自相似负载的高斯性，提出了 alpha 稳定磁盘 I/O 负载模型。通过与其它 I/O 负载模型的仿真实验结果进行对比，发现 alpha 稳定磁盘 I/O 负载能准确地描述 I/O 到达过程，无论是独立同分布型的，还是高斯自相似型的，或者是非高斯自相似型的，都具有适用性。与传统模型对比实验结果显示，由推荐模型生成的合成 trace 比其它方法更能准确地捕捉 I/O 突发到达行为。另外，alpha 稳定磁盘 I/O 负载模型包括 5 个参数，每个均具有相应的物理意义，这使得我们可以针对不同的应用环境很方便地转换 I/O 负载模型。相关成果发表在著名国际会议 the IEEE International Conference on Cluster Computing (Cluster '08)和 the 16th Annual Meeting of the IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS '08)。

课题二

结合未来网络存储技术的发展，结合具体应用的需求，针对大型网络存储环境的特征，研究了面向互联网数据服务的高速数据通道技术，为未来网络存储系统在高性能、高可用、可扩展以及面向服务的发展提供了创新与前沿的思路。

在前期动态可重构的多虚拟通道的研究基础上，下半年度进一步提出数据通道的动态自组织策略，使得数据通道能够自适应地调节，实现随需应变。在多个物理上分离存储网络子系统间使用 InfiniBand 网络进行互联，构建混合网络系统中的数据缓存。在底层统一的物理存储系统之上，依据不同应用的特点构建专用的存储支持，即在传统文件系统接口之上依据应用对数据的需求特征进行定制，使得不同的应用拥有自己独立的存储视图，以便更好的支持不同应用具体的存储需求，同时物理存储介质上各个不同应用所有的数据可以彼此共享;并针对高性能计算以及 Web 应用这两种典型的访问模式，建立了一个可以提供多种视图的数据通道的原型系统。完成的主要关键技术包括数据访问路径优化，可定制的预取策略，以及分层协作缓存管理策略三个方面。实验结果表明，相比于传统的并行文件系统，采用高速数据通道可以提高上述两种典型应用模式的数据访问性能。

进一步深入研究了 uStream 通信协议，为包括 Java 应用在内的上层应用使用高速网络进行数据传输提供高效的通信机制。uStream 是一个基于 InfiniBand 网络架构的零拷贝用户态流通信协议。由于 uStream 是在用户态实现的，所以能够消除用户/内核空间上下文切换的开销，它利用 InfiniBand 网络的 RDMA 特性，可以达到了与 InfiniBand

Verbs/RDMA 网络层访问接口相当的性能，并远远超越 InfniBand 上传统的传输层通信协议 SDP（Sockets Direct Protocol）。uStream 使用的关键技术主要包括：独立双通路（数据/控制）策略；缓冲区回绕算法与预注册，非实时确认策略，以及异步发送机制与零拷贝。

结合开放、动态的广域网络环境，提出在广域网络环境下存储对象的高效数据传输机制。面向互联网规模的应用，研究其海量数据的组织方式，以及计算任务的调度问题，使得后端 PB 级的存储系统更好地服务于上层应用。其核心思想是，在数据通道上建立一种有效的映射方法，将计算任务与数据组织更加紧密地结合，从而尽量减少在数据传输上的性能损失。设计并实现了小文件重组归并策略，分组预取策略，以及延迟垃圾回收等关键技术。我们通过一个具体应用——面向 web 的地图浏览来验证优化机制的有效性，测试结果表明，该机制不仅减少了对磁盘的访问次数，而可以大幅度降低分布式文件系统中的节点内存使用率，同时加快了文件的访问速度。

课题三

着重对海量存储系统的组织结构、运行机制、数据组织方式及相关理论的研究，并在此基础上构建相应的原型系统。根据这个目标，研究工作主要从构建海量存储原型系统和存储仿真系统、分析存储系统的性能特征和一些关键技术等几个方面展开。

在整个研究中，我们需要构建一个海量存储原型系统作为整个研究的平台，使得我们研究的关键技术能够在这个平台上加以体现。现阶段，我们已完成了客户端到存储服务节点的数据传输模块；实现了基于 OSD 的文件系统，能够对存储服务节点中的对象存储空间实施有效的管理，并针对一些特定应用（如 VOD）实现了性能优化；完成了元数据服务器的元数据管理和数据分布管理。整个分布式文件系统已经能够正常提供存储服务，为下一步分布式锁管理和客户端 cache 机制的实现打下了很好的基础。

针对海量存储系统运行规律研究方面，我们对集群存储系统性能分析，建立了集群存储系统的读写 fork-join 模型。我们通过建立理论模型描述 Incast 现象。由于实际产生 Incast 的因素设计到 TCP 传输协议中重传机制的细节，涉及的参数太多，其精确模型很难通过理论方式得到。

而对于集群存储系统而言缺乏有效的仿真平台也是制约对存储系统研究深入的一个重要原因。我们通过在 NS2 的基础上构建集群存储仿真系统。对于在 OST 端实际产生的 I/O 延迟通过两种方式模拟，一是通过建立 I/O 响应模型；另一种方法是把到达 OST 的请求记录下来，发到 Disksim 仿真平台，获取其响应时间。在 NS2 中根据上述得到的 OST 响应时间加上通过模拟网络传输过程得到的网络传输时间可以计算出全局的响应时间。

存储系统中不同的数据组织方式有性能、安全性和可靠性差异。因此评估不同的数据组织的可靠性就是非常重要的工作。传统上模型仅仅考虑磁盘的失效，没有进一步考虑磁盘中 sector error 所造成的数据丢失。在假设 sector 失效和修复时间也符合负指数分

布，采用 Markov 方法进行了分析。理论上通过对传统模型的扩展可以把 sector 失效，甚至能够其他数据丢失问题进行理论分析。

同时，我们把集群文件系统中分布式虚拟磁盘的概念借鉴到现代的大规模 RAID 中，提出了一种新的大规模阵列系统结构—聚合子阵列（Collection of SubRAID，CSRAID）。CSRAID 在阵列存储系统的块设备级实现，它把所有阵列磁盘构建成若干个“子阵列”，而这些子阵列共同组成一个全局的存储空间；子阵列由部分磁盘组成，其结构与传统阵列完全相同，但各子阵列可以根据其组成磁盘的数目、容量、性能、以及所服务负载的特点不同而相应地采用不同的数据分布方式和性能优化策略。因此，通过构造不同的子阵列可服务不同的应用请求，使整个 CSRAID 能够对所有用户表现出更佳的性能；子阵列的数目可以由系统任意添加，从而 CSRAID 能更好地合并新添加的磁盘，实现存储系统的可扩展性。

在存储系统中，合理的 cache 机制能够较大的提高系统的性能。目前大多数栈式 cache 都采用了三类工作栈：recency 工作栈、frequency 工作栈以及 history 栈。如何设计这些栈以及如何在这些栈之间取得平衡是获得高命中率的关键。我们研究了几种具有代表性的高性能栈式 cache 替换算法，并提出了一种新的 cache 替换算法－Adaptive Dual LRU(AD-LRU)，此算法其目标是尽可能高效的使用 buffer cache 页面。并通过动态调整 history 栈、recency 工作栈与 frequency 工作栈的大小来提高整个 cache 的效率。在与其它的类似的栈式 cache 替换算法如 ARC、LIRS、LRU 的比较中，其显示出较高的命中率。

课题四

2008 年下半年 04 课题组的工作重点是实际系统的开发和部署。经过几年来的深入的研究,我们形成了一些理论成果,为此，我们在系统研发中，不断的尝试将这些理论结合到系统中来，以验证理论成果的效果。具体来讲，我们的验证目标是复本的管理策略以及和系统可用性的关系。

我们在理论上提出了三个主要的方法：（1）用半马尔可夫过程刻画副本的再生过程，在此基础上基于节点的历史信息，对节点是否永久失效进行判别，（2）基于对节点离线行为规律的认识和大量的统计，设计了更加准确的存活复本估算器，（3）引入最小平均绝对误差估计来决定修复的时机，将修复代价降为最优值。

作为这些理论应用到实践的载体是我们新一代 P2P 文件共享存储系统 AmazingStore。系统研发所面临的最大挑战是动态性和稳定性之间的矛盾。由于上述三个理论成果的目的就是针对动态性的，因此我们据此形成了复本管理控制策略，将其应用到系统中以期屏蔽动态性带来的干扰。为了保证系统的可靠性，我们将提供存储空间的节点分为两类，一部分是实验室提供的可靠的服务器，一部分是用户的桌面机，每个可靠的节点作为 super node 负责一部分节点的信息搜集和数据维护。我们设计复本的分发策略，一方面将重要性高的数据分发到的可靠机器，另一方面也利用普通用户数多，

空间大的特点，用更多的冗余晃去可用和可靠。

在理论到实践过程中的我们主要解决的问题是信息搜集的问题。无论是节点失效判断还是副本数估计，都需要根据节点的一段时间的历史行为，为此要搜集大量的信息并对其进行处理。Super node 结构就是针对搜集信息而考虑的，把大量的节点分成簇来管理，信息从集中处理变成分布处理，为系统更大规模的运行提供了结构上的保障。

总之，半年来，系统的开发部署在经历了无数的测试->修改、再测试->再修的过程，使系统逐步稳定，各项功能也逐步完善，课题的总体按计划进展。

课题五

针对各种典型集群应用，建立有效的系统评价策略和方法，研究相应的存储资源调度和计算资源调度算法和策略，重点研究了大规模存储系统的性能优化技术。针对磁盘容错的纠错码、存储 Cache 策略和存储安全技术展开深入研究。取得了以下进展。

研究了高效的数据块关系挖掘方法和基于数据块关系访问特征的主要优化技术，使得挖掘开销相对传统方法下降一到两个数量级，系统的平均响应时间下降 50%-79%。主要进展如下:

1. 提出了基于噪音过滤的数据块关系挖掘方法。首先分析影响数据块关系挖掘效率的主要原因——噪音，并提出了支持噪音过滤的数据块关系挖掘方法 Z-Miner。Z-Miner 在经典的深度优先挖掘算法的基础上，根据不同噪音的类型使用全局分支裁剪和分支聚类方法对模式树(FP-Tree)进行裁剪，从而减少噪音产生的冗余分支。实验证明，Z-Miner 的挖掘性能相对经典挖掘方法提高 10-67 倍。而 Z-Miner 挖掘结果指导的数据预取使平均响应时间下降 26%-66%，优化效果是经典挖掘方法的两倍。

2. 提出了数据块关系指导的缓存替换方法 BDP(Block-correlations Directed cache replacement Policy)。利用数据块关系指导缓存替换需要解决数据块关系识别、预测和减少预测失败惩罚几个关键问题。BDP 使用时间戳方法识别访问流中按数据块关系进行访问的请求；利用历史访问信息和实时访问信息预测数据块未来的空间局部性特征；根据预测结果决定数据块的替换策略，减少预测失败带来的惩罚。本文同时给出了 BDP 缓存模型，优化 BDP 的参数配置。BDP 的失效率相对 LRU 算法下降 11%--38%，平均响应时间下降 13%-31%，优于现有的考虑空间局部性的替换算法。

3. 提出了数据块关系指导的多副本数据分布技术。同一数据块可能存在多种数据块关系，导致用来顺序分布的数据块集合规模过大，使预取效率下降和磁盘负载增加。本研究提出了多副本数据分布技术，分析模式树获得适合数据分布和预取的副本集合，使数据块在磁盘上存在多个副本。在应用访问时选择合适的副本进行预取。多副本数据分布技术在保证预取效率的同时控制预取的数据量和预取的 I/O 次数。使用多副本数据分布技术可以使平均响应时间下降 4.6%-13%。

在磁盘容错的纠错码、存储 Cache 策略和存储安全技术三个方面进行了深入和进一步的研究。其主要研究进展如下:

1、随着存储系统规模的不断扩大，磁盘容错已成为大规模网络存储系统设计中一

个十分重要的方面。目前已有的多副本镜像技术成本高，而现在 RAID 技术（如 RAID6）最多只能容忍两个磁盘出错。为此，我们研究了磁盘阵列的高容错纠删码技术，提出了一种易于实现的高容错高存储利用率的纠删码 GRID Codes 技术，具有如下优点：实现容易，容错能力能够高达 15 以上，存储利用率高达 80%以上，它的这些特征使得它非常适合于大规模磁盘阵列存储系统。

2、随着存储体系结构和应用的发展，不同的 Cache 策略在不同负载下表现不同，还没有一种 cache 策略能适应所有的负载。基于此，我们提出了一种 cache 策略的自适应选择方法 SOPA。在一个实际系统测试中，在采用了 SOPA 以后，系统的平均响应时间比采用 LRU 时降低最多 20.3%，比采用 ARC 时降低最多 11.9%，比采用 MQ 时降低最多 10.5％。

3、继续在存储安全方面进行研究，在研究了网络归档存储中数据持有性检查方法和文件系统的可信赖审计日志的基础上，对基于可验证的只添加文件系统的数据保护技术进行了研究，提出了一种利用可验证的只添加文件系统保护数据完整性的方法，来满足存储内容不可变数据时的法规遵从的需求。在 ext3 上实现了只添加文件系统的原型 ext3ao。测试结果表明：审计过程花费的时间比创建过程花费时间增加 1.72%，表明审计程序是高效的。

课题六

在前期研究的基础上，进一步完善了大型多媒体网络 GIS 系统原型 GlobeSIGht 的功能，对服务器端的空间数据存储管理进行了优化。研究了 GIS 系统性能测试指标与方法，提出了整体系统测试和部署方案的建议。以 MODIS 和风云卫星数据为主要数据源，在环境与灾害监测领域开展了多方面深入的研究和应用。进行了快速三维 GIS 可视化技术研究。并对网络 GIS 系统的动态负载均衡进行了深入的研究。2008 年下半年在国内外期刊和国际会议上发表相关论文 16 篇；毕业博士 3 名。

针对海量空间数据服务中的高可靠、高性能空间应用服务器集群，提出了一种基于最小总代价的网络地理信息服务负载均衡算法。针对大规模空间数据服务中所涉及到的数据量海量特征和空间数据服务质量要求，在分析了空间数据访问的时间和空间相关性的基础上，提出了一种基于瓦片访问平均时间间隔的空间数据置换算法。该算法实现简单，又具有明确的物理意义。

在快速三维 GIS 可视化技术研究中，研究了部件可视锥，通过预处理建立可视锥有效减少了实时可视化过程中可见性识别时间并显著提高了效率；研究了提高部件可视锥的预计算算法效率和使用其他遮挡查询方法来进行辅助判别的算法。研究了通过组法线快速判断对象部件是否可见的遮挡剔除算法和限时自适应绘制算法中的模型 LOD 级别切换策略；研究应用重要性因子进行模型 LOD 层次选择的评价指标，对选择结果进行量化分析。

提出基于“云参数法”遥感干旱监测方法，有效克服了传统方法的不足；建立基于“封冻指数（FI）”的 MODIS 数据冰凌监测模型，实现了黄河冰凌的日常动态监测业

务；提出了基于图像过渡区边缘高程和区域统计特征技术的雾检测方法，使遥感雾检测精度有较大的提高；提出了一种 MODIS 数据 Bowtie 效应快速消除算法，有效提高了海量数据的处理效率；基于亮度温度的动态阈值检测技术较好地解决了大尺度的热点检测；改进了一种基于 MODIS 数据的近海二类混浊水体遥感大气校正算法，反演的光谱反射率精度接近海上准同步测量结果。