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

[2009] 01 号 总第 12 期

项目办公室　编　　2009 年 5 月 28 日

项目结题工作部署会议召开

2009 年 5 月 13 日在北京湖北大厦召开 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”项目结题工作部署会议，参加会议人员有科技部基础司钱小勇博士、宋海刚博士，华中科技大学科发院夏松处长，项目咨询专家强文义教授、毕光国教授、钱华林研究员和马建峰教授，项目组专家叶朝辉院士、沈绪榜院士和何炎祥教授以及各位课题负责人等。大会由项目首席科学家冯丹教授主持。

冯丹教授介绍了项目的主要研究进展情况，并表示，项目研究已经开展了 5 年，即将结题，为了更好的完成结题工作，举行此次会议，希望大家能够在最后的三个月中，认真总结课题的研究内容，完善各项理论成果，争取在项目结题时交出一份漂亮的答卷。

会议中，来自华中科技大学、中科院计算所、北京大学、清华大学、武汉大学和解放军通信指挥学院的的各课题骨干分别介绍了课题的研究情况，对照立项时任务书，检查了课题五年来取得的各项成果，为即将到来的课题验收做好准备。

针对课题的研究情况，各位专家提出了后期工作的主要方向，并给出了宝贵的意见。专家组认为，项目经过五年的研究，已经取得了很多的成果和创新，如何将创新点突出，围绕立项时的关键问题，体现解决重大科学需求，是后期工作的重要内容。

会议最后讨论并部署了课题结题的主要工作和演示方案等。

项目研究进展

课题一

围绕对象存储系统展开研究，在大规模存储系统元数据管理和数据检索机制；存储系统容错纠错技术；广域网存储资源管理和发现等方面研究取得阶段性成果。主要取得下几个方面进展。

（1）提出一种提出基于元数据多维语义特征的文件组织模式 SmartStore 和一种支持多种复杂查询方式的数据组织模式 BR-tree。SmartStore 通过分析元数据多维属性信息的语义特征，将相关的文件组织在相同或相近的组内，多个组构成 semantic R-tree 结构，这样，使得相关的查询、添加/删除和更新操作可以在有限的小区域内完成，降低操作的执行代价。BR-tree 结构主要是通过在传统的 R-tree 节点上添加 Bloom filter 结构实现结构的扩展，这个结构充分考虑了 R-tree 在支持范围查询和 Bloom filter 支持点查询方面的优势，融合的结构不仅能够有效的支持上述两个操作，更能有效的支持 Cover query 和 Bound query，新的查询服务能有效的支持海量和不确定数据查询。相关研究成果被超级计算领域顶级会议 ACM/IEEE Supercomputing Conference (SC'09) 和 IEEE Transactions on Computers (TC)所录用。

（2）提出一种新型具有最优属性的 RAID-6 编码技术 P-Code，该技术表述简单直观，通俗易懂，实现容易。而且 P-Code 扩充了这两种最优 RAID-6 技术，它在阵列磁盘数目为素数和素数减一时都具有最优性，同时 P-Code 具有其他潜在特性：它的校验块均匀分布在阵列的各个磁盘上，使得其小写性能比常规的将校验块集中在专用的校验盘上的 RAID-6 技术要好；P-Code 是一种灵活而富有弹性的 RAID-6 技术，在它的编码结构中，数据块和它们的标签之间具有十分灵活的对应关系，使得 P-Code 有许多种变形形式，而且这种灵活性还可能在分布式存储系统中得到很好的应用。相关研究成果发表在超级计算的旗舰会议 23rd ACM International Conference on Supercomputing (ICS’ 09) 上。

（3）一种新颖的存储资源管理及发现系统，其中整合了 SLP（服务发现协议）和 XML Native 数据库，借鉴服务信息检索技术优化了管理信息的查找，并首次引入了 P2P 技术来改善存储管理信息的传递。针对广域网存储服务的规则管理，首先从应用特点出发，设计实现了一种可以通过应用特点定制规则管理命名空间的方法，优化复合应用对海量文件存储、服务系统的访问。相关的国际标准建议 “Scalability Support for Service Discovery of SMI-S using a Community-based P2P Network”，应邀参加 2009 ISO/IEC JTC 1/SC 25 会议，参加 ISO/IEC 13186 标准 “Information Technology — Storage Management” 讨论。

课题二

课题二近期在高速网络的协议，互联网应用数据访问优化，以及高速通道的智能化处理等方面取得了突破性的进展。

首先，进一步改进与优化了高速网络上的数据传输协议。目前，越来越多的应用希望通过高速网络来提高数据传输性能，但底层复杂的 Verbs 语义难以使用，而传统的上层通信协议 IPoIB 与 SDP 未能充分发挥高速网络的性能，且不能满足应用的多样性需求。因此，我们分别设计了 uStream 和 iSocket 两种用户态传输协议，在带宽与延迟上均取得了出色的性能。其中 uStream 提供了缓冲区管理策略，适用于包括 Java 在内的上层应用程序使用；而 iSocket 将缓冲区的分配与管理交由应用自身，提供了更高的灵活性，并且 iSocket 采用了事件驱动机制，使得 CPU 的占用率大大降低。

其次，针对大规模互联网信息服务中大量小文件访问的特征，基于流行的云存储平台 Hadoop 为 WebGIS 应用设计了一种小文件的优化策略，使得数据的注入时间大大缩减，并有效地提高了数据访问性能。同时，改进了大规模数据的存储计算框架 MapReduce，设计了基于 InfiniBand 的分布式缓存 iCached，并通过 iCached 改进了 MapReduce 的数据交换方法，即将基于磁盘读写的方式改为分布式缓存的交互方式，减少了 I/O 次数，大大提高了大规模数据处理的性能。

最后，在高速通道智能化处理方面，分别提出了集中式与分布式两种基于 FGPA 与多处理机的混合计算架构，可以实现对万兆流量的实时处理；其中，针对多种网络应用需要精确时间信息的需求，在集中式预处理系统中还设计实现了高精度时钟同步电路；另外，随着像分布式系统的广泛使用，负载均衡算法的研究将会变得越来越重要，为此提出了一个自适应的万兆链路负载均衡算法，达到了良好的性能，同时还提出了一个流量重映射算法，最大程度地降低了数据包乱序率。

课题三

近半年来，课题组工作进入整理、提炼和完善阶段，一方面，遵照课题的计划任务书，对所取得的科研成果进行排查；另一方面，对项目验收平台进行了规整和升级，包括对存储设备和网络设备的梳理和替换。

完成了一部题为《海量网络存储系统原理和设计》的专著写作，目前已经签订出版合同，进入出版程序。申报了 2 个发明专利和 2 个软件著作权，并向海量存储技术标准委员会提交了一份题为《绿色存储能耗测量标准规范》的采标，参加了两个采标的制定，分别是《磁盘阵列通用规范》和《对象存储设备指令集标准草案》。自主开发完成的多协议磁盘阵列控制器已经小批量生产，并被多个重要国防单位主动采用，其更高的安全性已受到这些部门的肯定，改变了此前我国关键部门不得不用国外存储设备存储重要数据的尴尬局面，满足了国家数据安全的重大需求。

课题四

以“对等模式下基于对象的数据组织和索引机制”为主线,全面研究了基于对等（后面的部分用 P2P 来代替“对等”）网络建立存储系统所涉及的基础理论，模型，核心技

术以及构建真实系统中必须解决的工程问题。2009 年主要的研究进展在体现在 P2P 系统的优化理论研究和系统构建方面:

数据的可用性分析模型和数据分发策略研究中提出时间相关可用性概念，分析了数据副本数与其可用性之间的关系。研究了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，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。

数据的可靠性分析模型和修复策略研究中分析出可靠性、系统开销和系统动态性三者的关系，提出如何优化配置系统的方案，同时基于误判抵消漏判的思想，合理解决了不能可靠区分永久错误和临时错误的关键问题。

在传统的 P2P 共享系统中，资源保存在用户的机器上，文件越流行，资源的副本越多，下载就越高效。但是对于一些热度下降的文件，用户会选择删除。因此，一些非热门文件在共享系统不容易获得。我们在设计新系统时，把共享和存储结合起来，定期把热度下降但并非毫无价值的文件自动存储到 P2P 存储系统中作为经典资源保存起来。用户不仅可以通过共享快速的获得热门文件，也可以通过检索找到被存储起来的冷门文件。基于这个思路，首次提出并实现了 P2P 共享存储系统—AmazingStore。一方面利用原有共享系统中的用户群，构成存储系统的存储空间，另一方面，利用获得的存储空间，保存副本不足的资源，做到共享和存储的优势互补。系统运行证明，这个思路是正确的。

课题五

遵照既定的课题计划，09 年上半年主要进行合作单位之间科研成果的集成。

依据所提出的虚拟管理架构模型，主要在存储资源虚拟化、计算资源虚拟化方面进行了技术集成。

在存储资源虚拟化方面，主要集成了条带卷快速扩容方法。该方法通过削减元数据写操作和合并数据读写技术，比传统的扩容方法最多快 40％。该条带卷快速扩容方法克服了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，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。

另外，为了进一步提高存储服务质量，增强磁盘数据的可用性，集成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。GRID Codes 技术具有如下优点：（1）完全基于异或运算且具有规则的编码结构，实现容易；（2）它的容错能力能够高达 15，甚至更高；（3）在编码方式固定的前提下，随着一个条带中条块数目的增加，它能够提供高达 80%，甚至更高的存储利用率。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。

在计算资源虚拟化方面，集成了服务动态部署技术。该技术基于存储与计算分离的思想对服务可重构计算机进行管理，通过调度和绑定计算资源与存储资源，动态构成计算系统对外提供服务，从而灵活满足多种应用需求。并且可以快速部署大量的计算系统满足大规模应用需求。另外，当所提供的某个服务撤销时，提供该服务的计算资源和存

储资源可以被用来提供新的服务，从而提高资源利用率。

课题六

1. 空间数据 P2P 存储模型

充分研究地形系统瓦片特征、用户漫游特征以及节点属性特征，研究一种基于兴趣域的节点分组对等网瓦片共享机制，根据地形数据金字塔模型建立基于金字塔的节点分组模型，对节点进行按兴趣分域，按服务质量分组。在相同区域漫游的节点具有类似的地形数据要求，提高节点间瓦片的共享效率。相关成果发表在《测绘学报》、《武汉大学学报(信息科学版)》等期刊，并申请软件专利一项。

2. 空间数据的缓存技术

研究了空间数据缓存技术。研究并设计瓦片缓存索引，以高效地对空间瓦片缓存进行操作。基于瓦片索引，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 TAIL。TAIL 置换算法综合了先进先出 FIFO, 最近最少使用 LRU, 最不经常使用 LFU 三种算法，巧妙地利用瓦片访问的时间局部性和空间局部性，在瓦片访问的长期流行度和短期流行度间取得一个平衡，即有利于瓦片访问的整体优化，又可以适应于不同用户地形漫游时对瓦片访问的模式变化和突发性访问。

更进一步，提出了一种基于瓦片寿命和访问热度的空间数据缓存置换算法 TCLEPR。考虑缓存中瓦片存活寿命和已经被置换的瓦片的平均缓存寿命，TCLEPR 置换存活寿命超出平均缓存寿命最长、访问热度最低，既老化程度最高的瓦片。相关成果发表在《测绘学报》、《武汉大学学报(信息科学版)》等期刊上。

3. 课题融合测试

与其它课题一起搭建融合测试环境，并对各课题进行了融合测试，得到了测试结果。结果表明，基于对象的存储系统，在各个方面性能表现优异。