融合 NAS 和 SAN 的存储网络设计与实现

韩德志[1,3], 余顺争[1], 谢长生[2]
(1. 中山大学电子与通信工程系, 广东广州 510275; 2. 华中科技大学计算机学院, 湖北武汉 430074;
3. 广东外语外贸大学信息学院, 广东广州 510420)

摘 要: 针对目前两种主流网络存储系统存在的缺陷, 本文提出和实现了一种在 IP 协议下融合 NAS 和 SAN 的统一存储网络系统. 通过全局多协议文件系统, 统一存储网络能同时支持文件协议和块协议, 实现了 NAS 设备和 SAN 设备在 IP 上的无缝融合, 满足了应用开放性、高扩展和海量存储的需求; 通过 iSCSI 软件实现模块, 统一存储网络能同时为客户提供文件 I/O 和块 I/O 服务, 具有 NAS 和 SAN 二者的优点; 通过自主存储代理文件系统, 统一存储网络能同时通过服务器通道或高速附网通道向客户机提供数据, 提高了系统的 I/O 响应速度, 减少了服务器瓶颈. 实验结果显示, 统一存储网络系统具有超高速的文件 I/O 和块 I/O 响应速度, 能为网络提供性能、扩展性、兼容性、性价比都更好的海量存储系统.

关键词: 统一存储网; 多协议文件系统; iSCSI; 自主存储代理
中图分类号: TP303 文献标识码: A 文章编号: 0372-2112 (2006) 11-2012-06

A New Storage Network Integrated with NAS and SAN

HAN De-zhi[1,3], YU Shun-zheng[1], XIE Chang-sheng[2]
(1. *Department of Electronics and Communication Engineering, Zhongshan University, Guangzhou, Guangdong 510275, China;*
2. *School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China;*
3. *School of Informatics, Guangdong University of Foreign Studies, Guangzhou, Guangdong 510420, China*)

Abstract: Nowadays, NAS and SAN are served as two main network storage systems, both of which have their advantages and disadvantages. With the dramatic increase of the network application, however, a new network storage architecture, Unified Storage Network (USN), in which NAS and SAN are integrated on the base of an IP, has been made. Firstly, by way of a Global Multi-Protocol File System (GMPFS), the USN combines NAS and SAN, achieving a high scalability and a large capacity. Secondly, with an iSCSI module, the USN serves the block I/O and file I/O simultaneously, combining both the advantages of NAS and SAN. Thirdly, through an Autonomic Storage Agency (ASA), the USN provides two types of data channels: a server channel and a high-speed network-attached channel, creating a direct access to the storage device for its client. And in the concerned experiments, the USN has turned out to be an improved network storage system of a great capacity with an ultra-high throughput for both the file I/O and the block I/O. Compared with the existed NAS and SAN, this new system has proved to be better in terms of its performance, scalability, compatibity, high-availability and price, etc.

Key words: unified storage network (USN); global multi-protocol file system (GMPFS); iSCSI; autonomic storage agency (ASA)

1 引言

IT 技术的发展经历三次浪潮: 第一次浪潮以处理技术为中心, 以处理器的发展为核心动力, 产生了计算机工业, 促进了计算机的迅速普及和应用; 第二次浪潮以传输技术为中心, 以网络的发展为核心动力. 这两次浪潮极大地加速了信息数字化进程, 使得越来越多的人类信息活动转变为数字形式, 从而导致数字化信息爆炸性地增长, 进而引发 IT 技术的第三次发展浪潮: 存储技术浪潮. 存储技术浪潮的核心是基于网络的存储技术. 目前, 流行的网络存储系统主要有两种: 附网存储 (NAS) 和存储区域网 (SAN)[1,2]. 按照存储网络工业协会 (SNIA) 的定义: NAS 是可以直接联到网络上向用户提供文件级服务的存储设备, 而 SAN 是一种利用 Fibre Channel 等互联协议连接起来的可以在服务器和存储系统之间直接传送数据的网络. NAS 是一种存储设备, 有其自己简化的实时操作系统, 它将硬件和软件有效地集成在一起, 用以提供文件服务, 具有良

收稿日期: 2005-11-30; 修回日期: 2006-07-20
基金项目: 国家自然科学基金 (No. 60173043, No. 60303031); 国家 973 重大基础项目 (No. 2004CB318203); 广东省科技攻关项目 (No. 2006B11201004); 中国博士后科学基金项目 (No. 20060390749)

好的共享性、开放性、可扩展性. SAN 技术的存储设备是用专用网络相连的, 这个网络是一个基于光纤通道协议的网络. 由于光纤通道的存储网和 LAN 分开, 性能就很高. 在 SAN 中, 容量扩展、数据迁移、数据本地备份和远程容灾数据备份都比较方便, 整个 SAN 成为一个统一管理的存储池(storage pool). 由于具有这些优异的性能, SAN 已成为企业存储的重要技术. 但在实际应用中 NAS 和 SAN 也存在很多缺陷, 越来越不能满足 IT 技术的快速发展和数字化信息爆炸性地增长的需求[3,4]. 如 NAS 设备存在如下缺陷: (1) 数据的传输速度慢, 因为 NAS 只能提供文件级而不能提供块级的数据传输; (2) 数据备份时性能较低, NAS 在数据备份时要占用其大部分网络带宽, 其它 I/O 性能受到影响; (3) 只能管理单个 NAS, 很难将位于同一局域网中的多个 NAS 集中管理. SAN 也存在以下缺陷: (1) 设备的互操作性较差, 不同厂家的设备很难互操作; (2) 构建 SAN 成本高, 目前只有实力较大的企业构建自己的 SAN; (3) 管理和维护成本高, 企业需要花钱培训专门的管理和维护人员; (4) SAN 只能提供存储空间共享而不能提供异构环境下的文件共享.

针对 NAS 和 SAN 的优缺点, 目前出现了多种新的网络存储技术, 如: NAS Gateway(NAS head)[5]、基于 IP 的 SAN 技术[6]、对象存储技术[7]. NAS 网关能将 SAN 连结到 IP 网络, 使 IP 网络用户能通过 NAS 网关直接访问 SAN 中的存储设备, 所以 NAS 网关具有以下优点: 能使 NAS 和 SAN 互连在同一 LAN 中, 突破了 FC 拓扑的限制, 允许 FC 设备在 IP 网络使用; 减少了光纤设备的访问成本, 允许访问未有充分利用的 SAN 存储空间. 基于 IP 的 SAN 互连技术主要包括: FCIP(IP tunneling)、iFCP、iSCSI、Infiniband、mFCP, 其代表技术是 iSCSI 技术. iSCSI 技术原理是将 SCSI 协议映射到 TCP/IP 之上, 即将主机的 SCSI 命令封装成 TCP/IP 数据包, 在 IP 网络上传输, 到达目的节点后, 再恢复成封装前的 SCSI 命令, 从而实现 SCSI 命令在 IP 网络上的直接、透明传输, 使访问远程的 SCSI 盘可以像本地的硬盘一样方便. 存储对象具有文件和块二者的优点: 象数据块一样在存储设备上被直接访问; 通过一个对象接口, 能象文件一样, 在不同操作系统平台上实现数据共享. NAS Gateway 虽实现了 NAS 和 SAN 在 IP 的融合, 但不是真正的融合, 因为它不能将 NAS 设备和 SAN 设备融合起来向用户提供统一的存储池, 用户也只能以文件 I/O 的方式访问存储设备. 对象存储虽具有 NAS 和 SAN 的优点, 但需要设计专门的对象存储接口, 需要对现有的文件系统进行修改, 这阻碍了它的进一步普及推广.

本文提出并实现了一种在 IP 协议下融合 iSCSI、NAS、SAN 的统一存储网络(简称 USN). 在 USN 中, NAS 设备、iSCSI 设备和 SAN 设备并存, 用户可以块 I/O 的方式访问 USN 中的 iSCSI 设备和 SAN 存储设备, 也可以文件 I/O 方式访问 USN 中的 NAS 存储设备和 SAN 存储设备, 整个 USN 是一个统一的存储池. 并且, USN 能同时提供服务器通道和附网高速通道向客户机提供数据, 减少了服务器瓶颈, 提高系统的 I/O 速度. USN 既有 NAS 的优点(低成本、开放性、文件共享), 又有 SAN 的优点(高性能、高扩展性). USN 同 NAS Gateway(NAS head) 技术、基于 IP 的 SAN 技术、对象存储技术相比具有明显的优势.

2　USN 总体结构

USN 系统的硬件结构如图 1 所示. USN 由 NAS 设备、iSCSI 设备和 SAN 设备, 以及元数据服务器和应用服务器组成. 用户可以文件 I/O 的方式访问 USN 中的 NAS 设备和经过 NAS 头访问 SAN 中的存储设备, 也可以块 I/O 的方式访问 USN 中的 iSCSI 设备和 SAN 中的存储设备. USN 同时向用户提供服务器通道和附网高速通道, 对于元数据和小数据请求都经过服务器通道完成, 对于大数据请求则经过附网高速通道完成, 这样大大提高整个系统的 I/O 速度, 减少服务器瓶颈. 整个 USN 是用基于 IP 的技术构建, 可以兼容现有的存储系统, 添加和删除存储设备都很方便. 所以, 整个系统的性能、扩展性都很好. USN 真正实现了 NAS 和 SAN 的统一, 即同一存储网络中既有 NAS 设备, 又有 SAN 结构; 实现文件 I/O 和块 I/O 的统一, 即用户可以文件 I/O 方式(文件为单位) 也可以块 I/O 方式(块为单位) 访问 USN 中的设备; 实现了文件协议和块协议在 TCP/IP 协议上的统一, 用户可以 NFS(Unix 用户) 和 CIFS (Windows 用户) 访问 USN, 也可以 SCSI(iSCSI 用户) 访问 USN.

图 2 是 USN 的软件结构图, 其中 GMPFS 是全局多协议文件系统, 位于 USN 系统中的各个应用服务器上, 它支持使用 CIFS 协议的 Windows 用户对 USN 的访问, 支持使用 NFS 协议的 UNIX 用户对 USN 的访问, 也支持使用 iSCSI 协议的块协议用户对 USN 的访问. GMPFS 通过对目前存储系统所使用的元

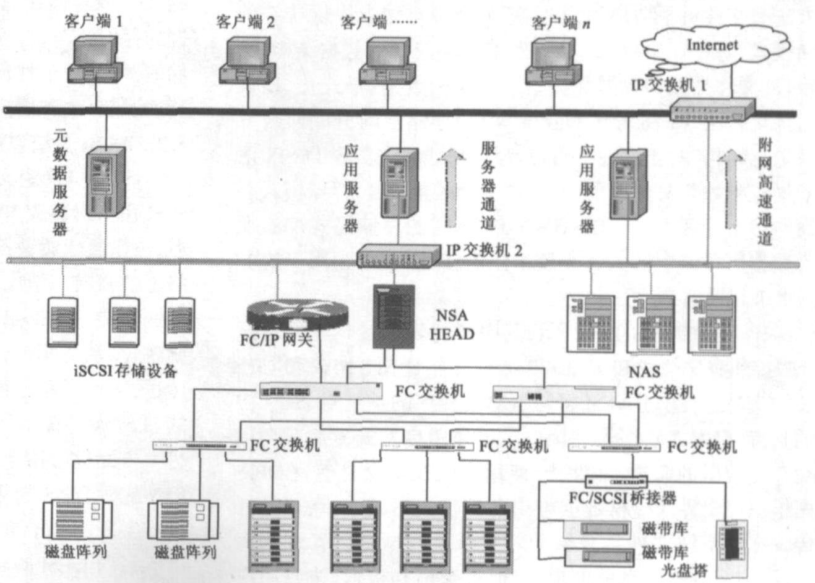

图 1　USN 系统的硬件结构

数据进行扩展, 采用启发式的方法, 收集用户应用信息, 为用户提供统一、方便、快捷的存储访问接口以及合理的数据存储方案. ASA 是自主存储代理模块, 它能够自动地发现海量存储系统中存储设备的种类和可利用的各种资源, 自主地对这些存储设备和资源进行有效的统一管理和优化. ASA 根据应用的不同和应用的具体需求, 安排与应用相适应的存储设备种类、性能以及可靠性和可用性等级等, 并为 I/O 请求选择合适的数据通道, 使应用得到最优的存储资源分配, 从而使整个系统的性能达到最佳.

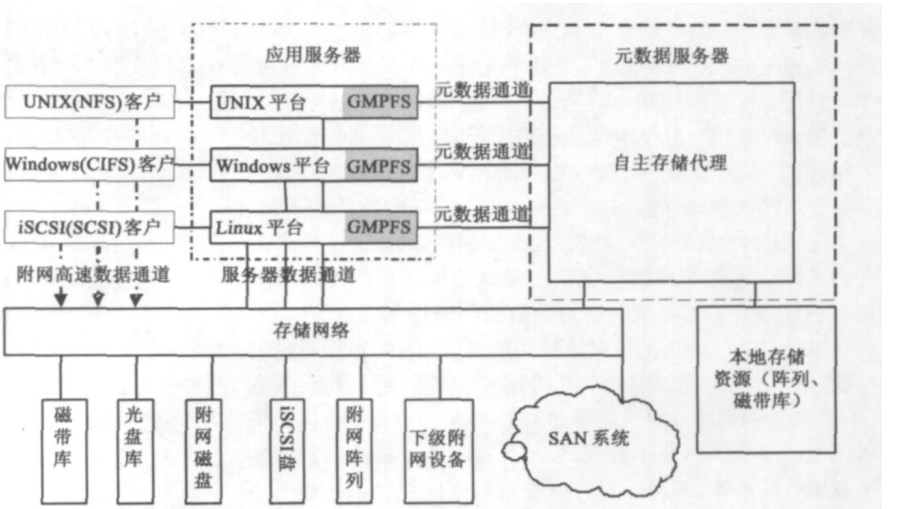

图 2 USN 的软件结构图

3 系统设计

USN 是一个复杂的系统, 涉及到许多复杂的技术, 本文主要论述其核心技术的设计和实现, 即 GMPFS、ASA 和 iSCSI 系统的设计与实现. GMPFS 可以驻留在多种操作系统平台上(UNIX, Windows, Linux), 支持各种协议用户的访问(NFS, CIFS, iSCSI), 为用户或应用程序提供对网络存储系统的数据访问服务. ASA 将多种存储技术(这些存储技术各有所长, 也各有所短) 整合为一个统一的海量存储系统, 充分发挥各种存储技术的优势, 使得该存储系统对特定的应用程序而言服务性能达到最优, 有效地满足多方面的应用需求. iSCSI 真正的实现了块 I/O 和文件 I/O 在 IP 网络上的统一, 文件协议和块协议在 IP 协议上的统一.

3.1 全局多协议文件系统的设计

GMPFS 保留了分布式文件系统的灵活性和高性能的优点, 而克服了其在不同 I/O 协议支持方面的缺陷, 能同时支持 NFS、CIFS 和 iSCSI 协议用户的访问. GMPFS 在提供文件存取的方法和文件目录结构的同时, 还为每种存储卷提供特定的存储模式. 每种存储模式包含某种文件系统的元数据结构, 操作接口(文件类型和数据块类型), 功能函数集(格式化, 检索等), 优化方法(cache 方法和预取等) 和存储空间分配回收方法及数据结构. 对于文件卷而言, 存储模式包含实现 POSIX 语义的操作函数和文件目录结构; 对于分区卷而言, 存储模式必须面向特定分区类型, 如 NTFS, ext3. 所有的存储模式都必须在元数据服务器中的 ASA 系统中注册, 以便 ASA 为用户的 I/O 请求进行通道选择.

GMPFS 的结构如图 3 所示. 其中协议转换接口主要通过 NFS 的扩展程序模块和 samba 模块的组合对 NFS 协议和 CIFS 协议的支持, 并通过 iSCSI 目标器驱动程序的扩展对 iSCSI 协议的支持. 启发式数据管理接口主要是用启发式方法获得用户对存储数据的需要, 如性能、使用率以及安全性等. GMPFS 数据组织逻辑界面提供数据组织的逻辑视图, 这一点正是针对传统文件系统文件目录结构对于海量数据难以管理的弱点, 在增加元数据信息的前提下, 通过查询和检索, 按照用户需要提供各种类型文件视图, 例如根据文件创建的用户和时间进行分类. 扩展的文件或卷操作接口、数据组织与分配管理、元数据组织结构和 I/O 定向器等主要是保证与传统的文件系统操作语义兼容, 实现程序级的数据访问. 应用程序无需修改就可以使用 USN 系统中的数据. 提供与元数据服务器中的 ASA 及存储资源的接口和通讯, 能充分利用 ASA 系统所掌握的存储资源, 合理组织数据, 满足用户或应用程序对数据存储的多方面、个性化要求. 如通过同时提供服务器通道和附网高速通道, 改善用户的 I/O 性能服务, 减少服务器瓶颈.

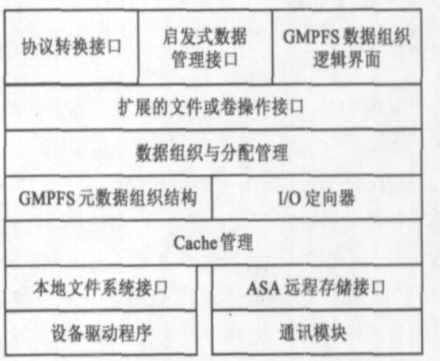

图 3 GMPFS 的结构

3.2 iSCSI 系统设计

iSCSI 协议定义的是 SCSI 到 TCP/IP 的映射, 即将主机的 SCSI 命令封装成 IP 数据包, 在 IP 网络上传输, 到达目的节点后, 再恢复成封装前的 SCSI 命令, 从而实现 SCSI 命令在 IP 网络上的直接、透明传输. 它整合了现有的存储协议 SCSI 和主流网络协议 TCP/IP 等两种主流协议, 实现了存储和网络的无缝融合. 从应用的角度看, iSCSI 一方面通过 SCSI 命令的远程传送, 实现了和远程存储设备的命令级交互, 使用户访问远程的 SCSI 设备像本地的 SCSI 设备一样方便, 而且具有高速度; 另一方面也可用于改造传统的 NAS、SAN 技术, 实现 NAS 和 SAN 的融合. iSCSI 系统是 USN 系统的核心部分之一, iSCSI 的设计实现了基于 IP 的数据块访问机制.

目前 iSCSI 的实现方式可以考虑采用以下三种方式: 纯软件方式、智能 iSCSI 网卡实现方式、iSCSI HBA 卡实现方式[1].

由于我们是设计 USN 的原形系统, 所以只采用纯软件方式, iSCSI HBA 卡方式是下一步产品化我们将实现的目标. iSCSI 系统整体设计模型如图 4 所示(不包括管理模块). 服务器端(Target) 采用 linux 操作系统, 客户端(Initiator) 采用 Windows2000. SCSI 微端口驱动在系统中生成一个虚拟的 SCSI 磁盘, 过滤驱动截获系统发给 SCSI 磁盘的 SCSI 命令, 通过核心态的网络接口发给服务器处理.

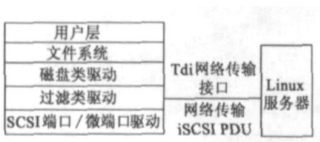

图 4　iSCSI系统整体设计模型

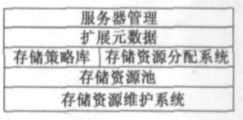

图 5　ASA 结构

3.3　自主存储代理系统的设计

自主存储代理 ASA 的一端面对海量存储系统. 目前的存储系统有 DAS(直连存储) 、NAS、SAN、iSCSI 等, ASA 能够自动地发现海量存储系统中存储设备的种类和可利用的各种资源, 自主地对这些存储设备和资源进行有效的统一管理和优化; 根据应用的不同和应用程序的具体需求, 安排与应用程序相适应的存储设备种类、性能以及可靠性和可用性等级等, 使应用程序得到最优的存储资源分配.

ASA 的另一端面对应用程序(GMPFS) . ASA 通过对目前存储系统所使用的元数据进行扩展, 采用启发式的方法, 收集用户应用信息, 为用户提供统一、方便、快捷的存储访问接口以及合理的数据存储方案; 根据用户 I/O 请求所涉及数据的属性, 选择客户端与存储设备交互数据的通道, 即元数据(目录、卷信息等) 和小数据 I/O 请求, 选择服务器通道, 对大数据 I/O 请求选择高速附网通道. 大、小数据 I/O 请求由 ASA 自主地根据整个系统的 I/O 信息量进行调整. ASA 系统结构如图 5 所示.

4　客户端与 USN 交互流程

USN 系统中包括三类用户: Windows 文件 I/O 用户(使用 CIFS 协议) , Unix 文件 I/O 用户(使用 NFS 协议) , iSCSI 块 I/O 用户(使用 iSCSI 协议) . 用户在客户端与 USN 系统交互流程与图 6 所示.

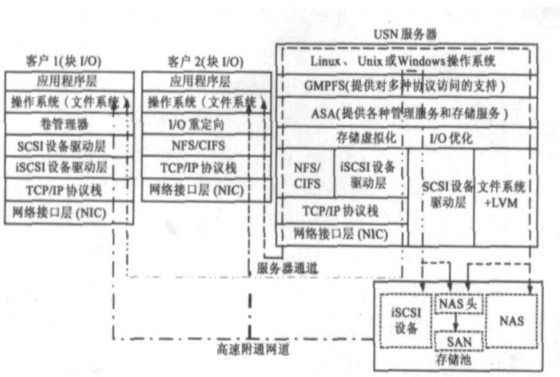

图 6　块 I/O 客户和文件 I/O 客户对 USN 的访问流程

块 I/O 客户的具体的数据读写流程为(如图 6) : (1) 客户 1 上的应用程序发出的块 I/O 命令(SCSI 命令) 经 iSCSI 设备驱动层和 TCP/IP 协议栈之后, 封装成 IP 数据包, 在 IP 网络上传输; (2) 封装后的 SCSI 命令达到 USN 服务器之后, 经解封装, 恢复成封装前的 SCSI 命令, USN 服务器利用这些 SCSI 命令对 iSCSI 存储设备发出块 I/O 读写请求; (3) 被请求的数据块经 iSCSI 设备中的 iSCSI 层和 TCP/IP 协议栈封装成 PDU, iSCSI 设备传送的 PDU 到客户端可经两个途径: 一种是经过服务器转发, 一种是经过高速附网通道直接传到客户端; (4) PDU 经 IP 网络上传输返回到客户 1 后, PDU 经客户 1 解封装并由其文件系统组合成文件.

当 USN 系统提供 File I/O 服务时, 其数据读写过程(如图 6 所示) : (1) 客户 2(文件 I/O) 向 USN 服务器发出文件读写请求(其工作方式和传统的 NAS 相同) ; (2) USN 服务器接到客户端的文件读写请求后: 一方面, 将该 I/O 请求发给对应的 NAS 设备或 NAS 头, NAS 设备或 NAS 头将所请求数据传给 USN 服务器, 再经 USN 服务器传到客户端; 另一方面 USN 服务器不把文件 I/O 请求传到 NAS 或 NAS 头, 而是将 NAS 或 NAS 头的 IP 地址传给客户端, 客户端通过该 IP 地址直接与 NAS 或 NAS 头进行数据交互.

这里的 NAS 头主要是支持 FC 协议的 SAN 设备能直接挂到 TCP/IP 网络, 支持 NFS/CIFS 用户的访问, NAS 头也可安装 iSCSI 目标器驱动程序支持 iSCSI 用户的访问. 不论是块 I/O 请求还是文件 I/O 请求, 都可通过附网高速通道实现客户端与存储设备的数据交互.

5　试验评估

从客户端对构建 USN 的各子存储系统以及整个 USN 进行功能和性能评测, 并作进一步的比较. 我们从两个方面对统一存储网进行测试: 功能测试和性能测试. 功能测试包括: (1) 构建 100M 及 1000M 以太网环境, 将 iSCSI 存储设备与服务器连接; 在服务器操作系统中安装 iSCSI 软件包后, 使用户能够通过网络获得 iSCSI 存储设备提供的存储空间, 并能象使用本地硬盘一样对其进行操作. 本测试项测试服务器端 iSCSI 盘安装、设置、管理和使用等各项功能; (2) iSCSI 存储设备作为 NAS 头的存储设备, 与 NAS 头组成一个 NAS 存储系统, 本测试项测试 iSCSI 盘在 NAS 中的安装、设置、管理和使用等各项功能; (3) iSCSI 盘与本地盘、FC-RAID 盘构成各种冗余度的 RAID, 本测试项测试各种存储盘在 RAID 中的安装、配置、管理和使用等各项功能; (4) 多个 NAS、iSCSI 设备、NAS 头连接 FC-RAID 通过多 GMPFS 和 ASA 构建成 USN 海量存储系统, 本项测试测试 GMPFS 和 ASA 系统在融合 NAS、iSCSI 和 SAN 的系统中的安装、配置及使用等各项功能. 性能测试包括: 测试在 100M 和 1000M 网环境中不同工作负载下 NAS 存储设备、iSCSI 存储设备、FC-RAID、本地硬盘以及它们组成的海量 USN 系统的数据传输性能: 包括单位时间内的 IO 次数、一次 IO 的平均响应时间、数据传输率和 CPU 利用率. 该项测试的主要思想是针对不同的网络应用环境, 对各种存储设备和各种传输通道进行频繁的 IO 处理, 在确定时间内统计并计算 IO 率、数传率、响应时间、CPU 利用率等性能参数, 从而得到的各种性能评估.

5.1　测试环境

iSCSI 存储设备: P4 2.0GHz CPU, 256MB DRAM, IBM DPSS-

318350 18G 硬盘, Redhat Linux 9.0 操作系统; LINUX 服务器: Pentium 4 2.66GHz (FC-PGA) CPU, 256MB DRAM, 80GB Ultra ATA/100 7,200rpm 硬盘, Redhat Linux 9.0 操作系统; WINDOWS 服务器端: XEON 3.06GHz CPU, 512M DRAM 内存, Smart Array 6i(板载)存储控制器, Qlogic QLA2300 PCI FC Adapter 光纤适配器, IBM 36.4GB (32P0726) 10Krpm 硬盘, Microsoft Windows2003 操作系统; FC-RAID: NexStor 4000S, CPU 600MHZ, 512M SDRAM, 10×ST314680FC 硬盘; 普通 NAS 存储设备: P4 2.66GHz CPU, 512MB DDR, Maxtor 160G 硬盘, Redhat Linux 9.0 操作系统.

网络连接: iSCSI 设备和普通 NAS 设备都使用 100M 以太网卡 Realtek RTL8139; Windows 服务器使用 1000M 以太网卡 HP NC7782 Gigabit Server Adapter; Linux 服务器使用 1000M 以太网卡. HP NC7782 Gigabit Server Adapter.

5.2 功能测试

根据测试流程, 功能测试包括三个方面的内容: (2) 平台的统一, 即在 Windows 下能通过单一目录树方式访问多个存储节点, 功能与 Linux 下的 pvfs 相似; (2) 协议的统一, 即通过 Windows 的"计算机管理"和 Initiator 发起端(iSCSI 客户端)可以管理 FC-RAID 和 iSCSI Target 及普通的 NAS 设备, 并利用"动态磁盘机制"实现多种冗余; (3) 设备的统一, 即 iSCSI Target 通过和 initiator 配合, 使得该 Target 成为 NAS 系统中的一个存储设备.

5.3 性能测试

5.3.1 测试内容

采用第三方的 IOMETER 测试软件进行的测试. IOMETER 是 INTEL 公司专门开发的用于测试系统 I/O 性能的测试程序. 它的测试参数比较全面, 能非常全面的反映服务器的 I/O 性能. 为了说明 USN 存储系统的性能, 在相同条件下测试以下项目进行对比分析: (1) 对 USN 服务器本地硬盘读写性能测试; (2) 100M 以太网环境下 FC-RAID 盘读写性能测试; (3) 100M 以太网环境下远程 iSCSI 盘读写性能测试; (4) 100M 以太网环境下 FC-RAID 盘和远程 iSCSI 盘构建的各级 RAID 盘的读写性能测试; (5) 1000M 以太网环境下远程 iSCSI 盘读写性能测试; (6) 100M 以太网环境下 USN 系统的读写性能测试.

5.3.2 实验结果比较

本地 IDE 硬盘、100M iSCSI 硬盘、1000M iSCSI 硬盘、FC-RAID、FC-RAID 与 iSCSI 构成的 RAID0 及 USN 系统数据传输率性能比较如图 7 所示.

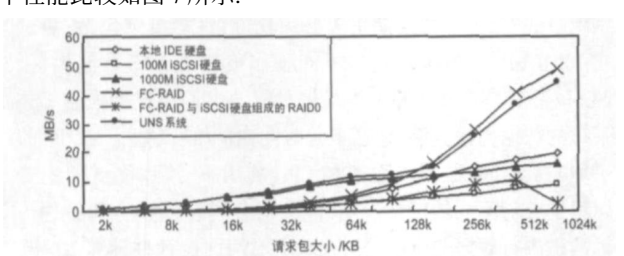

图 7 数据传输率性能比较

本地 IDE 硬盘、100M iSCSI 硬盘、1000M iSCSI 硬盘、FC-RAID 及 FC-RAID 与 iSCSI 构成的 RAID0, 以及 USN 的 IO/s 性能比较如图 8 所示.

本地 IDE 硬盘、100M iSCSI 硬盘、1000M iSCSI 硬盘、FC-RAID 及 FC-RAID 与 iSCSI 构成的 RAID0, 以及 USN 的平均响应时间性能比较如图 9 所示.

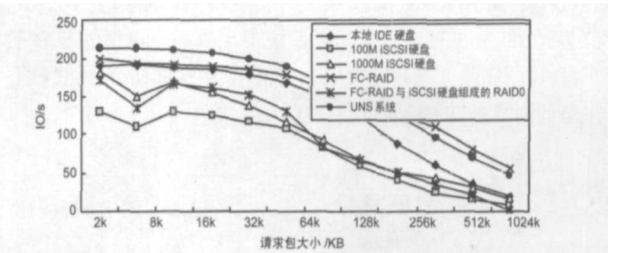

图 8 IO/s性能比较

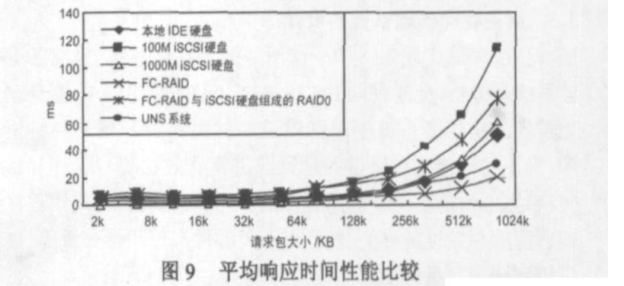

图 9 平均响应时间性能比较

本地 IDE 硬盘、100M iSCSI 硬盘、1000M iSCSI 硬盘、FC-RAID 及 FC-RAID 与 iSCSI 构成的 RAID0, 以及 USN 的 CPU 占用率比较如图 10 所示.

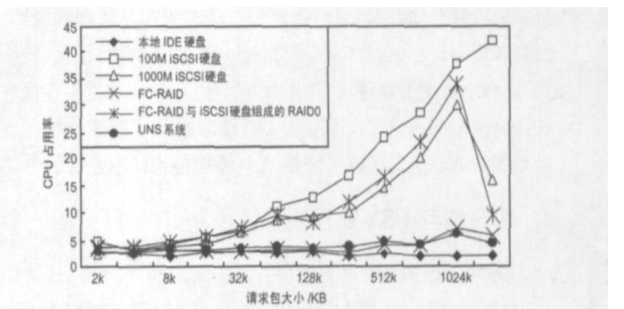

图 10 CPU 利用率比较

5.4 实验结果分析

5.4.1 请求文件或数据块大小对存储系统性能的影响

从图 7、图 8 和图 9 中单条曲线的走势可以看出, 当请求文件或数据块较大时, 从目的盘或系统上读写数据耗费的时间长, 通过网络传输的时间也相应增加, 所以: 小包的平均响应时间< 大包的平均响应时间, 小包的 IOps> 大包的 IOps. 请求包大时, 针对一个请求包所进行的额外操作较请求包小时少, 连续的读写所耗费的时间小于小包读写所耗费的时间, 因此: 小包的 MBps< 大包的 MBps.

服务器端 iSCSI 盘的各项性能表现趋势在 100M 以太网和千兆以太网环境中不同请求包大小的情况下符合上述规律, 本地 IDE 硬盘、FC-RAID 和 USN 系统也符合上述述规律.

5.4.2 性能分析

从图 7、图 8 和图 9 可以看出, I/O 请求在 1k~128kB 时, USN 系统的 I/O 请求响应速度比本地 IDE 硬盘、FC-RAID、100M 远程 iSCSI 硬盘和 1000M iSCSI 硬盘快的多. 当 I/O 请求大于 128kB 时, USN 系统的 I/O 请求响应速度比 FC-RAID 的

I/O 请求响应速度略慢,比其它存储子系统的速度快的多,最高速度可达 45MB/s. 其原因是我们在 USN 的服务器端除加载了 GMPFS(支持使用多种访问协议用户)和 ASA(提供服务器通道和附网高速通道)的同时,还加载了我们实验室以前开发的智能预取、硬盘缓存技术(DCD)、负载均衡和零拷贝系统或软件模块,所以,不论是大 I/O 请求还小 I/O 请求,都能提供极好的 I/O 请求响应性能. 而 FG-RAID 由于自身的数据校验等时延等特性,对小的 I/O 请求响应速度较慢,对越大的 I/O 请求响应速度越快.

对于 USN 的 iSCSI 盘存储子系统,从实验结果可以看出,当请求数据块较小时,100M 网络环境下的性能和 1000M 网络环境下的性能差别不明显,随着请求块或文件逐步增大,两者 IOps 和 MBps 的差距越来越大. 请求数据块为 1024K 时,仅更换网络传输中的数据链路层和物理层,从 100M 网络环境提升到 1000M 网络环境,磁盘数据传输率得到较大的提高,后者约是前者的 3 倍.

从图 10 可以看出,100M 的 iSCSI 存储子系统的 CPU 占用率最高,原因是在响应用户的 I/O 请求,要求服务器不断的对 iSCSI 的协议数据单元进行封装和解封装. 本地的 IED 硬盘 CPU 占用率最低,USN 系统的服务器端 CPU 占用率次之,原因是 USN 系统中小的 I/O 请求直接经过服务器处理,而大的 I/O 请求经过附网高速通道由存储设备自身处理.

6　结论和展望

我们提出、设计和实现的统一存储网络系统,全部采用 IP 互联设备,价格比光纤通道低得多,在管理软件的开发实现上以及系统的使用维护上,都具有多得多的资源和经验. 并且,千兆以太网技术比光纤通道技术发展迅速,10Gbps 以太网交换机已经推出并在市场上热销,其性能前景也比光纤通道交换机好得多. 所有这些为统一存储网络的产品化打下了坚实的基础.

目前,我们已经从理论、结构和实践上实现了统一存储网络原型系统,现在,我们正在开发和完善多用户、多功能、多种平台支持的 iSCSI 设备,设计和实现新的安全和高可用文件系统,以便为统一存储网络系统产品化后能真正为广大企业,尤其是为广大中小企业提供开放性、性能、可展性、性/价比都更好的海量存储系统.

参考文献:

[1] 韩德志,谢长生,等. 一种基于 iSCSI 的附网存储服务器系统的设计与实现[J]. 计算机研究与发展, 2004, 41(1): 208 - 213.
Han De-zhi, Xie Chang-sheng, et al. Design and implementation of an iSCSI-based network attached storage server[J]. Journal of Computer Research and Development, 2004, 41(1): 208-213. (in Chinese)

[2] 韩德志, 谢长生, 等. 一种新的附网存储集群系统的研究与设计[J]. 通信学报, 2005, 30(5): 1- 8.
Han De-zhi, Xie Chang-sheng, et al. Study and design of a new network attached storage cluster[J]. Journal Computer Research and Development, 2005, 30(5): 1- 8. (in Chinese)

[3] 谢长生, 傅湘林, 等. 一种基于 iSCSI 的 SAN 的研究与实现[J]. 计算机研究与发展, 2003, 40(5): 747- 751.
Xie Chang-sheng, Fu Xiang-lin, et al. The study and implemerrtation of a new iSCSI-based SAN[J]. Journal of Computer Research and Development, 2003, 40(5): 747- 751. (in Chinese)

[4] Telikepalli R, Drwiega T, et al. Storage area network extension solutions and their performance assessment[J]. Communications Magazine, IEEE, 2004, 42(4): 56- 63.

[5] Auspex whitepaper: NAS-SAN convergence today: new trends in enterprise storage[EB/OL]. www. internetnews. com/storage/article. php/, 2005-12-01.

[6] LeftHand Networks Inc. White Paper: Leveraging iSCSI SANs for Disaater Recovery[EB/OL]. http://www. lefthandnetworks. com, 2005-08-19.

[7] Mike Mesnier, Gregory R. G, et al. Object-based storage[J]. IEEE Communications Magazine. 2003, 41(8): 84- 90.

作者简介:

韩德志　男, 1966 年生于河南信阳,中山大学博士后. 主要研究方向: 高可用、高可扩展的海量网络存储系统,网络存储系统的安全管理和性能优化. E-mail: han_ dezhi88@ tom. com

余顺争　男, 1958 年生于江西南昌,中山大学教授,博士生导师,主要研究方向: 网络行为与网络安全、信号处理、无线 Internet.

谢长生　男, 1957 年生于湖北襄樊,华中科技大学教授、博士生导师,主要研究方向: 高精度光存储设备及系统,海量网络存储系统.