网络存储原理与技术

周敬利　余胜生　等　编著

清华大学出版社

网络存储原理与技术

周敬利　余胜生　等　编著

清华大学出版社

内 容 简 介

本书对网络存储原理与技术进行了详细的研究，在全面介绍信息存储领域的各种实用技术，如SCSI、RAID、NAS、SAN和iSCSI的基础上，深入讨论了网络存储虚拟化、网络存储管理、网络存储中的备份与容灾，以及网络存储安全等问题，其中涉及的一些设计方案和技术来源于作者的研究成果。希望为信息存储技术向大容量、高性能、网络化以及可管理和低成本的方向发展提供一些解决方法。

本书可供相关领域的研究人员和专业技术人员参考，也可作为高校信息科学专业师生的教学参考书。

图书在版编目(CIP)数据

网络存储原理与技术/周敬利，余胜生等编著. —北京：清华大学出版社，2005.9

ISBN 7-302-11419-6

Ⅰ. 网…　Ⅱ. ①周…　②余…　Ⅲ. 计算机网络—信息存储　Ⅳ. TP393.0

中国版本图书馆CIP数据核字(2005)第082170号

出 版 者：清华大学出版社　　**地　　址**：北京清华大学学研大厦

http://www.tup.com.cn　　**邮　　编**：100084

社 总 机：010-62770175　　**客户服务**：010-62776969

责任编辑：黎　强

印 刷 者：北京市清华园胶印厂

装 订 者：三河市春园印刷有限公司

发 行 者：新华书店总店北京发行所

开　　本：185×260　**印张**：13.5　**字数**：324千字

版　　次：2005年9月第1版　2005年9月第1次印刷

书　　号：ISBN 7-302-11419-6/TP·7503

印　　数：1～4000

定　　价：28.00元

前言

PREFACE

随着计算机科学的迅速发展，目前，信息存储领域正朝着高速I/O通道、集群存储、大容量、高性能、可管理、高可靠性、高可用性和低成本的方向发展。为此，国内外的科研、教学部门和产业界都投入大量的人力和物力进行研究。

当今的存储技术，已由单纯的以服务器为中心（serve centric）的存储模式变为以数据为中心（data centric）的存储模式。这种存储模式导致了网络存储技术的飞速发展，基于FC-SAN技术、IP SAN技术和iSCSI-SAN技术的网络存储技术更是面临着巨大的发展机遇和诱人的前景。

本书主要反映作者近年在网络存储方面的理论研究和科研成果，并对目前信息存储领域的主要动向和国内外的主要研究成果进行了介绍，希望对从事信息存储的研究开发人员具有一定的参考价值。

全书共分为10章，第1章为概述，描述了当前网络存储技术及系统的基本概念和动态。第2章为SCSI技术，对其体系结构、工作原理及标准进行了阐述。第3章为RAID技术，主要论述了RAID级别、软/硬件实现方法和实际应用案例。第4章为NAS存储技术，详细地讨论了NAS的结构、文件共享协议、NAS的可靠性及现有的产品。第5章基于光纤通道的存储区域网络，主要讨论了SAN的特点、组件、拓扑结构、管理标准和基于FC-SAN的应用。第6章基于IP的存储区域网络，对iSCSI的实现、设备发现、安全策略等问题进行了详细的分析。第7章为网络存储管理，讨论了NAS、SAN的管理策略，并对NDMP进行了阐述。第8章的内容是介绍存储虚拟化技术，着重对存储虚拟化的实施、管理、安全等问题进行描述。第9章的内容是网络存储中的备份和容灾，着重对数据的备份、复制、容灾等关键技术进行了研究。第10章涉及的是网络存储系统的安全性，主要讨论网络存储系统的安全模型及安全策略。

本书第1章由周敬利编写，第2章由夏洪涛、周敬利编写，

第3章由夏涛编写，第4章由孙伟平、余胜生编写，第5章由孙伟平编写，第6章、第8章由曾东编写，第7章由郭辉和夏涛编写，第9章由朱建峰、周敬利编写，第10章由欧阳凯编写。全书由周敬利和孙伟平统稿。

由于网络存储技术仍处在迅速发展的阶段，因此还有不少的新技术和新文献不能在书中反映，且因为作者的学术水平及实际能力有限，书中错误与不妥之处，敬请读者批评指正。

在编著本书的过程中，参考和引用了许多国内外有关的书刊和文献资料，同时得到作者所在单位及科研课题组的同事，尤其是博士生的大力支持，特此向提供这些书刊和文献资料的作者及所有支持和帮助过我们工作的朋友表示衷心的谢意。

周敬利
于华工园
2005年2月

目录

CONTENTS

第1章

CHAPTER 1

概 述

在计算机网络技术、计算机软/硬件技术及计算机应用技术的迅速发展过程中，IT 技术经历了三个阶段的发展过程。第一个阶段是以处理器为核心，它促进了计算机的普及和应用；第二个阶段是以传输技术为核心，它带动了计算机网络的使用和普及，使得数字化信息的应用席卷全球，并因此导致数字化信息的爆炸性增长，从而引发了第三个阶段——存储技术的发展。因此信息存储系统已成为国内外研究的重点和新的经济增长点。

网络技术的发展和应用对信息存储系统提出了更为广泛的要求，如高可靠性和高可用性的 356×7 在线存储服务、基于内容的数据存储、数据远程容灾和快速灾难恢复、异构平台的数据共享、存储管理的复杂度、系统的开放性、可扩展性、可靠性以及自主性等，以便提高存储系统的响应时间、吞吐率、可扩性和可靠性。

目前主流的存储技术有 DAS(direct attached storage)、NAS(network attached storage)、SAN(storage attached network)、CAS(content addressable storage)等。

DAS 是一种以服务器为中心的存储结构，服务器通过总线与各种存储设备相连，所有的数据传送和客户端的请求都需要

通过服务器，因此当客户数增加时，受主机带宽和内存限制，服务器成为整个存储系统的瓶颈；同时，由于在 DAS 环境下的各服务器是相互独立的，因此所有的存储数据将形成“孤岛”而不能共享，所以 DAS 不能满足现代网络应用对存储系统的要求。

NAS 是一种以数据为中心的存储结构，可向网络用户提供跨平台的文件级海量数据共享的功能。它通过多种文件共享协议（如 NFS，CIFS，Apple Talk 等）为不同的异构平台客户端提供文件共享服务。

NAS 设备通常具备三个特点：①附加大容量的存储；②内嵌操作系统；③专门针对文件系统进行重新设计和优化以提供高效率的文件服务。NAS 一般直接接入基于 TCP/IP 的局域网或广域网中，并使用特定的文件访问/共享服务（比如 Unix 系统的 NFS 及 Windows 系统的 CIFS）向服务器或客户机提供基于文件的服务。这种连接技术主要应用在基于文件的共享环境中，比如：文件服务器、打印服务器，因而 NAS 特别适合替代一些中、小企业的文件共享服务器。但由于 NAS 采用传统的基于文件的备份方式，因此扩展性、灾难恢复能力不足。

SAN 是近年来流行的另一种基于“块”数据访问的存储解决方案，是以网络为中心的存储结构，它采用高速的光纤通道（fibre channel，FC）作为传输媒介，将存储系统网络化，从而实现高速的共享存储，具有可扩展、高可用性、集中管理的优势，但是投资高，不能实现异构平台共享。2000 年 IBM、Cisco、Intel 等推出了 iSCSI 网络存储协议，即 Internet SCSI 协议，目前该协议得到广泛的研究和使用，iSCSI 是把 SCSI 命令封装到 TCP 包中在 TCP/IP 网络上传输，是存储技术和网络技术的融合，具有利用成熟的 IP 技术实现远程“块”数据访问的能力。该协议的推出得到了存储网络行业协会（SNIA）的大力推广，并在 2003 年 2 月被 IETF（internet engineering task force）批准为 IETF 标准，目前基于 iSCSI 的 SAN 受到了广泛的关注和研究。

CAS（content addressable storage）是 EMC 公司提出的一种新的存储解决方案，其特点是面向大量用户对一次写入多次读出的存储应用需求，通过对存储内容的分类和索引，实现基于内容的存储检索。

在网络存储领域，除了上面阐述的研究工作外，国内外研究机构还进行了大量的工作，目前比较有影响的存储系统有：

Ocean Store 是加州大学伯克利分校的 John Kubiatowicz 等人提出的全局存储体系结构，其特点是实现数据的全局存储表示和全局惟一名字来实现任意存储，因此结构复杂，实现困难，管理成本和复杂性较高。

GFS（global file system）是明尼苏达大学 Steven R. Soltis 等人提出的一种应用于光纤通道存储系统中的全局文件系统，它允许多个 Linux 客户机通过网络共享存储设备。每一台机器都可以将网络共享磁盘看作是本地磁盘，而且 GFS 自己也以本地文件系统的形式出现。如果某台机器对某个文件执行了某种操作，则后来访问此文件的机器就会读到写以后的结果，GFS 中采用了设备锁来实现不同 Linux 客户机的高速缓存（cache）数据一致性。

NASD（network-attached secure disk）是卡内基·梅隆大学的 Garth A. Gibson 提出的基于智能存储设备的集成安全存储系统，该系统在 NAS 存储设备智能化基础上进

行扩充，实现基于文件的安全访问。

自主计算(autonomic computing)是IBM研究中心首先提出来的概念，其灵感来自人体复杂的自主神经系统，能以同样的方式预测系统的需求和清除故障——在无需人工干预的情况下自主运行。通过自管理来帮助解决日益复杂的计算环境中所面临的管理与成本问题。哥伦比亚大学的Gail Kaiser教授等在Autonomizing Legacy systems中对存储系统的自主特性进行了有效的研究。

在国内，不少大学和科研机构也在网络存储技术方面进行了相应的研究并取得了一定的成果。

综上所述，虽然在网络存储方面做了大量的工作，但上述研究均是针对不同的应用来解决各自的问题，如何针对不同应用的存储服务质量需求，建立I/O资源分配与调度策略以及存储系统的自主管理机制等问题的研究还有待加强。

1.1 外存储器

本节就物理存储设备及存储原理、有关技术参数和存储网络技术的问题进行讨论。

1.1.1 存储器的类型

根据存储器的物理特性，可以将存储器分为以下三种类型。

1. 半导体存储器

半导体存储器是一种用集成电路技术构成的存储器。它品种繁多，主要作为计算机的内存使用，因此本书不作详细介绍。

2. 磁面存储器

这是一种用介质表层磁性材料的磁化状态来记录信息的存储设备，包括磁带机、磁盘机以及其他一些设备。其中磁带机具有容量大、性能可靠等优点，但却存在数据传输率低的致命弱点，目前仅用在需要大量备份的情况，本文亦不再赘述。对磁盘机而言，原有软盘机和硬盘机之分，但由于技术的发展，软盘机已由U盘所代替，因此下面仅就硬盘机进行描述。

磁盘机是目前用得最多的主要的外存储设备，从20世纪50年代发展至今，其存储密度、存储速度每年都以成倍以上的速度增长。它经历了固定头磁盘机、可换盘磁盘机到固定盘磁盘机的发展阶段，当前所用的磁盘，绝大多数都是固定盘磁盘机。这种磁盘机由主轴、固定在旋转轴上的盘片、磁头和磁头臂以及定位系统、电路等构成。它的规格从14英寸、8英寸、5.25英寸开始，一直发展到目前用得最多的3.5英寸。此外，市场上还有微型磁盘机，为2.5～1.8英寸。以3.5英寸磁盘为例，其存储容量可达80GB，120GB，160GB等。

3. 光盘存储器

光盘存储器是利用激光器产生的激光使光记录介质产生变化，从而达到记录数据的设备。写入时以高度集中的能量照射在介质上使其发生物理或化学变化，以达到数据记

录的目的。读出时，激光器产生的能量以小于写入时的激光光束照射在介质上，经过光的反射(或透射)率的变化，实现对数据的读取。用于计算机外存储的光盘机有以下三类。

(1) 只读型光盘机

这种光盘机媒体上的信息在出厂时已由生产厂家写入，用户只能读出而不能改写。它可用于大容量的数据库和应用软件库，也可供各种专家系统、文字和图像存储之用。由于其盘片易于复制，价格低廉，因而得到广泛的应用，是目前广播电视、商业、金融、计算机、军事、航空等技术或领域中不可缺少的部分。

(2) 只写一次型光盘机

这种光盘机用户可以写入一次并可多次读出，但已写入的信息同样不能抹除或重写，它称为 CD-R(CD-recordable)，适用于文件、档案、资料、图像等的存储。

(3) 可擦除多次重写光盘机

这是一种可多次读写的光盘机。根据读/写原理的差异，可分为磁光盘、相变光盘等类型。光盘机存储容量很大，每一位信息存储的成本很低，近年来在大容量外存储设备中占有很大的优势。

1.1.2 存储器的主要技术参数

1. 存储密度

(1) 道密度

道密度是指在垂直于信息道即盘片的半径方向上单位长度所记录的磁(光)道数。道密度 D_t 可按下式求得：

$$D_t = \frac{1}{W+G} = \frac{1}{P_t}(\text{道 / 毫米或道 / 英寸，记作 Tpmm 或 Tpi}) \tag{1-1}$$

其中，W 为信息道宽度(毫米或英寸)，G 为道间距或沟槽间距(毫米或英寸)，P_t 为道距(毫米或英寸)。

(2) 位密度

对于磁盘或光盘机，位密度是指信息道单位长度上所记录的二进制数的位数(或称比特数，记作 b)。位密度 D_b 可按下式计算：

$$D_b = \frac{ft}{D_{min}\pi}(\text{位 / 毫米或位 / 英寸，记作 bpmm 或 bpi}) \tag{1-2}$$

其中，D_{min} 为最内圈信息道直径(毫米或英寸)，t 为每转时间(秒)，f 为数据传输率(位/秒)。

2. 存储容量

存储容量是指存储器所能容纳的二进制数码的总量(以位数或字节数计量)。设 m 为记录面数，n 为每面磁道数或柱面数，则磁盘存储器非格式化的存储容量 C_d 为

$$C_d = f\ t\ m\ n(\text{位}) \tag{1-3}$$

格式化后的容量为

$$C_d = n\ m\ n_s\ B_{secto} \tag{1-4}$$

其中，n_s 为每道区段数，B_{secto} 为每区段的字节数。光盘存储器以块(block)作单元。

3. 存取时间

存取时间是指读、写头从所在位置到达要求的某一位置开始读/写时为止的一段时间。它包括两部分，一部分是寻找磁盘(或光信息道)所需的寻道时间 t_s，另一部分是等待所需写入或读出的区段旋转到读/写头的下方时的等待时间 t_w。其中，寻找相邻信息的寻道时间最短，称为最小寻道时间；而从首道寻找末道(或相反)的寻道时间最长，称为最大寻道时间。两者的平均值称为平均寻道时间。

平均存取时间 t_a 可以用下式粗略地计算：

$$t_a = \frac{1}{2}[(t_s + t_w)_{min} + (t_s + t_w)_{max}] \tag{1-5}$$

4. 数据传输率

数据传输率(传输速度)是指在单位时间内外存储设备向主存储器传送数码的位数或字节数(记作 B)。数据传输率正比于记录的位密度 D_b 和媒体相对读、写头的移动速度 v。数据传输率 $D_{r.d}$ 可表示为：

$$D_{r.d} = D_b v \text{ (b/s 或 B/s)} \tag{1-6}$$

对于任意信息道上的数据，数据传输率是相等的。但是，对于磁盘机，不同信息道上的实际位密度与切线速度都不相同。因此(1-6)式中的位密度与切线速度均应按最内圈信息道计算。若为等位密度，则磁盘机与光盘机的数据传输率，因不同信息道而改变。

5. 无故障间隔时间(*MTBF*)

无故障间隔时间是指正常工作的平均持续时间，即两次故障的间隔时间。它表示设备工作的可靠性程度。通常，对于磁记录设备 *MTBF* 约在 800～20000h。与此相关，还有一个称为“平均修复时间”的指标。平均修复时间一般标明为 1～0.5h，实际上大于此值。

6. 误码率

外存储设备的误码率是指向设备写入一批数据并回读后，所检出的错误位数与这一批数据总位数的比值。通常，有所谓“软错误”与“硬错误”之分。软错误是指经过重新测试可以纠正的错误，硬错误则是虽经重试也不能消除的错误。

1.1.3 磁存储设备的记录原理

磁存储设备的记录原理包括写/读和抹除的过程。这些过程均是利用电磁转换原理来完成的。它又可分为纵向磁记录和垂直磁记录两种。

(1) 抹除过程：抹除过程的工作是在记录之前，通以交流或直流电使磁头产生抹去磁场，将已记录的信息或残余的杂散信息抹去。磁带机采用直流抹除的方法，磁盘机通常采取重写覆盖而不使用抹除。

(2) 写入过程：写入时，记录媒体匀速地通过磁头下方，同时将磁头线圈通以一定的编码规则调制的脉冲电流，在磁头上产生磁场使介质磁化，从而记录相应的数据。当磁头线圈离开介质时，介质上的剩磁不再变化，从而能保存数据。

(3) 读出过程：根据法拉第和楞次定律，当磁化的记录介质通过磁头时，磁头线圈中

即感应出电压，此电压经过处理，即可还原出原来记录的信息。

1.1.4 光存储设备的记录原理

尽管光存储设备有很多，但其读/写的基本原理相似。在写入时，由激光器产生的衍射光斑，以高度集中的能量照射在介质上，使其发生物理或化学变化，达到数据记录的目的。读出时，仍由激光器产生能量较写入时小一些的激光光束，照射在介质上，经过光的反射率(或透射率)的变化，实现对数据信息的读取。

1.2 存储体系结构及其演变

1.2.1 传统的存储系统

当前应用最广泛的企业级存储产品仍然是磁盘阵列、磁带机、磁带库、光盘塔或光盘库等几大类。这些存储设备多以并行 SCSI 总线连接至某一特定的主机，存储设备只能被该主机直接访问和控制，其他主机需访问存储设备中的数据时，必须经该服务器的存储和转发。这样的“以服务器为中心”的存储结构被称为直接附属存储(direct attached storage，DAS)或附属于服务器的存储(server attached storage，SAS)。

如图 1.1 所示，在 DAS 结构中，客户端访问共享数据的步骤是：①通过网络将请求命令发至服务器；②服务器查询缓冲区，若数据在缓冲区中就经网络适配器发送数据给客户机，否则就将请求翻译成本地数据访问命令，然后发向与服务器相连的存储设备；③存储设备在收到命令后将数据拷贝到服务器的系统缓冲区；④数据再由系统缓冲区拷贝到网络适配器的数据缓冲区；⑤数据最后通过网络发向客户端。

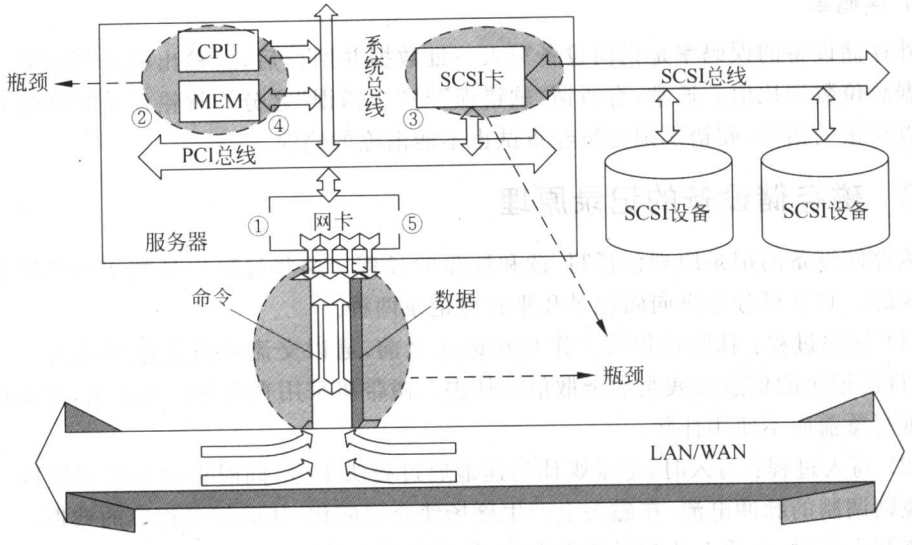

图 1.1 传统服务器中的瓶颈

这样，网络上的数据须经服务器的存储和转发，服务器的负荷较重，服务能力受到影响，在服务器中存在着 SCSI-IP 的协议变换，效率低下，实时性差，服务器中存储 I/O、网络 I/O 以及 CPU 和内存容易成为系统的瓶颈，很难满足数字化时代海量数据的存取和传输的实时需要。

在传统的存储系统结构 DAS 下，无论如何提高服务器和存储设备的性能，在大量客户机请求的情况下，服务器都将成为数据访问的瓶颈。卡内基·梅隆大学的研究和实验结果表明，这种瓶颈效应有可能会导致存储系统的资源利用率只有 3%。

1.2.2 网络存储的优势

受各种技术进步和应用的推动，现代企业和个人的数据信息量正在持续地爆炸式地增长。由于新 Moor 定律所揭示的现代信息技术发展的迅速和不平衡，传统的存储系统结构 DAS 已无法满足需求，存储系统学科面临着巨大的压力：一方面，数据量每年翻番，不断挑战信息服务系统的性能；另一方面，存储设备的可用性已成为威胁数据安全和信息服务连续性的重要原因之一，这个问题由于数据的重要性日益提高而表现得越来越突出。

通常，缓解高速网络环境下的 I/O 瓶颈采用两种途径：①改善存储设备和网络接口的性能；②协调存储 I/O 和网络 I/O 之间的关系。前者主要致力于不断提高单个存储设备的存储密度和数据传输速度，如采用垂直磁记录技术，不断提高主轴电机转速，采用性能更高的磁头、存储介质和新的伺服定位技术，增大缓存等。但是，存储设备本身的性能改进总是远远落后于 CPU 和光通讯的改进，无法从根本上解决 I/O 瓶颈问题。因此需将设备按并行、可扩展、易管理、经济等原则组成多级并行存储系统，从系统结构的角度对存储系统的性能加以改进来满足 I/O 需求。

为此，采用高速网络把存储设备连接起来，通过系统软件进行存储资源合理的调度和分配，使系统的整体性能达到最优，以一个系统的协同工作来满足众多单个用户的需求的方案已被提到当前的存储领域，这种新型的存储系统结构就是网络存储（network storage）。网络存储技术的出现，响应了网络发展的三种重要趋势，即：网络正成为主要的信息处理模式；需要存储的数据大量增加；数据作为取得竞争优势的战略性资产，其重要性在增加。

与传统的存储结构 DAS 相比较，网络存储系统具有以下的技术优势[21]：

（1）良好的可扩展性。在传统的存储系统中，由于受服务器 CPU 处理中断速度、I/O 通道带宽和存储设备有限的编址空间等限制，不能无限制地扩展存储系统。而网络存储则很容易随系统负载的增加而进行“无限制”的扩展，系统的存储容量、带宽等技术指标同时得到了几乎是线性的提升。

（2）数据共享。全球的用户都可以通过网络系统借助于多种操作系统获得所需的数据。

（3）更好的可管理性。网络存储设备及其中的数据均可通过专门的管理软件统