Object-Based Network Storage

基于对象的网络存储

◎郭玉东　尹　青　编著

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

基于对象的网络存储

郭玉东　尹　青　编著

電子工業出版社
Publishing House of Electronics Industry
北京 • BEIJING

内 容 简 介

本书概述了网络存储体系结构的演变，介绍了网络存储的基本概念，论述了基于对象网络存储产生的必然性及其意义，并以基于对象的网络存储为主线，深入讨论了相关的协议标准，包括 SCSI 体系结构（SAM）、SCSI 基本命令（SPC）、Internet 上的 SCSI（iSCSI）、基于对象的存储设备（OSD）和并行网络文件系统（pNFS）标准等。最后，讨论了几种基于对象的网络存储文件系统，包括 Lustre、panFS、zFS 等。

本书重点讨论的是网络存储的基本协议，适用于对网络存储感兴趣的读者及从事网络存储工作的技术人员和管理工作者，可以作为大专院校本科生的高年级教材或硕士研究生教材，也可以作为相关专业的教学参考书。

图书在版编目（CIP）数据

基于对象的网络存储/郭玉东，尹青编著. —北京：电子工业出版社，2007.10
ISBN 978-7-121-05147-0

Ⅰ. 基… Ⅱ. ① 郭… ② 尹… Ⅲ. 计算机网络—信息存贮 Ⅳ. TP393.0

中国版本图书馆 CIP 数据核字（2007）第 154296 号

策划编辑：张 旭
责任编辑：宋兆武 沈德雨
印 刷：北京天宇星印刷厂
装 订：三河市皇庄路通装订厂
出版发行：电子工业出版社
北京市海淀区万寿路 173 信箱 邮编 100036
开 本：787×1 092 1/16 印张：18.25 字数：455 千字
印 次：2007 年 10 月第 1 次印刷
定 价：38.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

前 言

近年来，随着信息技术的普及，数据存储量每年约以80%的速度持续增长。关于信息增长的速度，图灵奖获得者Jim Gray提出了一个经验定律：网络环境下每18个月产生的数据量等于有史以来的数据量之和。由于数据量的持续增长，数据存储已经成了一个关键的问题。专家们断言，目前的信息技术已进入了以存储为核心的发展阶段。

网络存储结构的发展经历了几个阶段，大致包括DAS、SAN、NAS、带文件系统的SAN和OSD等。理想的存储结构应该能够提供强安全性、跨平台的数据共享、高性能和对存储设备与客户数量的高可伸缩性。

DAS（Direct Attached Storage）是最简单的一种存储结构。其主要缺点是存储设备扩展困难、共享困难，服务器容易成为瓶颈，而且系统的可靠性较差。对DAS的改进方法有两种：一是将存储设备从服务器中独立出来，组成单独的存储区域网SAN（Storage Area Network）；二是简化服务器的操作系统、网络协议、文件系统等，形成专用的存储服务器，称为NAS（Network Attached Storage）。NAS中存储的数据容易共享，但NAS的本质仍然是服务器，当它管理的容量过大时，仍然会成为系统瓶颈。SAN的扩展性很好，存储设备容易共享，但其中存储的数据却难以共享。为了方便共享SAN中的数据，可以在SAN上建立文件系统。带文件系统的SAN对外提供文件接口，而且容易扩展，但其安全性较差。为了提高存储系统的安全性，需要增加存储设备的处理能力。增加了处理能力以后的存储设备除了能够完成安全检查之外，还可以完成其他一些工作，例如：可以让存储设备自己负责存储空间的管理，包括存储空间的分配与回收等；可以让存储设备自己管理存储在其上的数据，如按照对象方式组织、管理数据和数据的属性等。

这种具有智能、能够自我管理、提供对象接口并有较高安全性的存储设备称为基于对象的存储设备（Object Based Storage Device，OSD）。相应地，以OSD为基础的网络存储就是基于对象的网络存储。OSD具有NAS和SAN的所有优点，而且有较高的安全性，是网络存储的主要发展方向。为此，INCITS（InterNational Committee for Information Technology Standards）的T10技术委员会开发了OSD标准。2005年1月，ANSI批准了OSD标准，目前的版本是2.0。

OSD标准的制定宣布了网络存储新时代的到来。它已经对网络存储的发展产生了影响，而且将继续产生深远的影响。因而有必要对OSD及其相关标准进行深入研究。事实上，国际上许多大学和科研机构都在开展OSD的研究，目前的主要成果是基于OSD的分布式文件系统，如Cluster公司的Lustre、Panasas公司的panFS、IBM研究室的zFS、美国加利福尼亚大学存储系统研究中心的Ceph等。研究结果表明，与普通的DAS、NAS、SAN相比，基于

OSD 的存储系统具有高性能、高安全性和高可伸缩性等许多优点，能够提供跨平台的数据共享，是一种先进的网络存储结构。本书第 1 章讨论了网络存储体系结构的演变及 OSD 的特点。

OSD 标准是 SCSI 协议族的一个成员，是一种特殊的 SCSI 命令集。要想了解 OSD，就不得不研究 SCSI。SCSI 协议族大致分成四层，从下到上依次是物理接口层、传送协议层、基本命令集和专有命令集，每一层都由若干标准组成。为了描述 SCSI 的基本概念和组成关系，在 SCSI 协议族中还专门定义了一个标准，称为 SCSI 体系结构模型（SCSI Architecture Model，SAM）。SAM 是 SCSI 协议的基础文档，它描述 SCSI 协议的基本概念，是理解 SCSI 协议的关键。因而本书第 2 章讨论 SCSI 体系结构。

SCSI 基本命令集（SCSI Primary Commands，SPC）是所有 SCSI 设备都应该实现的命令集。其中定义了 SCSI 命令描述块（CDB）的基本格式和最基本的 SCSI 命令，是理解 SCSI 命令的基础。因而，本书第 3 章讨论 SCSI 基本命令集。

一个 SCSI 系统中必须包含一个服务传送子系统，用来将不同的 SCSI 设备连接起来并实现它们之间的通信。SCSI 服务传送子系统可以使用不同的传送协议，如 FCP、SPI、SSA 等。当前最受关注的是 iSCSI，即 Internet 上的 SCSI。2003 年 2 月 11 日，国际互联网工程任务组（Internet Engineering Task Force，IETF）通过了 iSCSI 协议标准。iSCSI 标准的发布标志着廉价网络存储时代的到来。本书第 4 章讨论 iSCSI 协议。

OSD 标准是专门为 OSD 设备定义的命令集。在 OSD 标准中，定义了 OSD 的基本概念、OSD 对象的属性和 OSD 命令的基本格式。OSD 标准是设计和使用 OSD 设备的基础，也是基于对象的网络存储的基础。因而，本书第 5 章讨论 OSD 标准。

虽然 OSD 具有更高的智能，但它本质上仍然是存储设备。在一个基于 OSD 的存储系统中可能存在多个 OSD 设备，因而需要对它们进行统一管理，这就是基于 OSD 的分布式文件系统。在众多的基于 OSD 的分布式文件系统中，只有 pNFS 正趋于标准化，很有可能成为标准的基于 OSD 的分布式文件系统。因而，本书第 6 章讨论 pNFS 标准。

除了 pNFS 之外，还有几个很有特色的基于 OSD 的分布式文件系统，如 panFS、Lustre、zFS 等，这些文件系统管理基于 OSD 的网络存储系统，向用户提供标准的文件系统接口，是目前市场上正在使用的基于对象的网络存储文件系统。另外，OSD 设备中也需要一个管理系统，专门用于管理其中的对象，OBFS 是这类管理系统的一个代表。本书第 7 章讨论上述几个文件系统。

本书的写作，先后花费了一年多的时间。其间，有很多同事、朋友给予作者极大的帮助，借此机会一并表示感谢，感谢信息工程大学信息工程学院为本书的出版提供了支持。

由于作者水平有限，错误之处在所难免，敬请读者批评指正。指正意见请发邮件至 ydguo621@163.com。

编著者
于河南郑州信息工程大学信息工程学院
2007 年 6 月 6 日

读者意见反馈表

谢谢您关注《基于对象的网络存储》一书！烦请填写本表。您的意见对我们出版优秀教材、服务教学十分重要。如果您认为本书有助于您的教学工作，请您认真地填写表格并寄回。**我们将定期给您发送我社相关教材的出版资讯或目录，或者寄送相关样书。**

个人资料

姓名_____年龄_____电话_____（办）_____（宅）_____（手机）

学校_____专业_____职称/职务_____

通信地址_____邮编_____E-mail_____

您校开设课程的情况为：

本校是否开设相关专业的课程　□是，课程名称为_____□否

您所讲授的课程是_____课时_____

所用教材_____出版单位_____

本书可否作为您校的教材？

□是，会用于_____课程教学　　□否

影响您选定教材的因素（可复选）：

□内容　□作者　□封面设计　□教材页码　□价格　□出版社

□是否获奖　□上级要求　□广告　□其他_____

您对本书质量满意的方面有（可复选）：

□内容　□封面设计　□价格　□版式设计　□其他_____

您希望本书在哪些方面加以改进？

□内容　□篇幅结构　□封面设计　□增加配套教材　□价格

可详细填写：_____

_____

您还希望得到哪些专业方向教材的出版信息？

_____

谢谢您的配合，请将该反馈表寄至以下地址。如果需要了解更详细的信息或有著作计划，请与我们直接联系。

通信地址：北京市万寿路 173 信箱 基础教育分社

邮　　编：100036

电　　话：010-88254511；88254518

E-mail：jichu@phei.com.cn

http://www.hxedu.com.cn

反侵权盗版声明

举报电话：（010）88254396；（010）88258888

传　　真：（010）88254397

E-mail：　dbqq@phei.com.cn

通信地址：北京市万寿路 173 信箱

　　电子工业出版社总编办公室

邮　　编：100036

目　录

第 1 章　网络存储概论

专家们认为，在计算机网络技术、计算机软/硬件技术及计算机应用技术迅速发展的过程中，IT 技术经历了三个主要发展阶段。第一个阶段是以处理器为核心的阶段，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的阶段，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，并因此导致了数字化信息的爆炸性增长；第三个阶段是以存储为核心的阶段，它主要研究存储系统的可靠性、可用性、开放性、可扩展性，以及存储数据的容灾与恢复、共享与安全等。因此，信息存储技术已成为国内外研究的重点和新的经济增长点。

存储结构的发展也经历了几个阶段，大致包括 DAS、SAN、NAS、带文件系统的 SAN 和 OSD。理想的存储结构应该能够提供强安全性（Strong security）、跨平台的数据共享（Data sharing across plateforms）、高性能（High performance）和对存储设备与客户数量的高可伸缩性（Scalability in terms of the number of devices and clients）。

1.1　直接附属存储 DAS

直接附属存储 DAS（Direct Attached Storage）是最简单的一种存储结构。

事实上，DAS 是一种以服务器为中心的存储结构，服务器通过总线（SCSI、ATA/IDE 等）与存储设备相联，客户机与服务器之间通过 IP 网络相联。

图 1.1 是一个 DAS 系统的结构示意图。

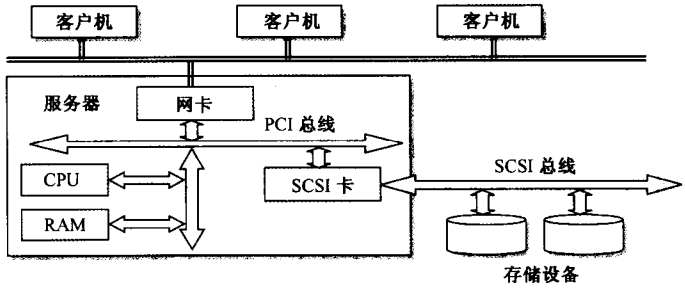

图 1.1　DAS 系统结构

WhatIs.com对 DAS 的定义是："Direct-attached storage (DAS) is computer storage that is directly attached to one computer or server and is not, without special support, directly accessible

to other ones"。也就是说，DAS 是与计算机或服务器直接相联的存储设备，在没有特别支持的情况下，其他机器不能直接存取它。对一般的计算机用户来说，硬盘是最常见的 DAS。

对服务器来说，存储设备是它的本地块设备，服务器和存储设备之间以块为单位交互。

为了使用方便，在服务器上通常安装有数据库管理系统（如 Oracle、SQL Server 等）或网络文件系统（如 NFS、CIFS 等）。服务器上的数据库管理系统或网络文件系统负责存储设备的管理与共享，客户机按照下列步骤读取存储设备中的数据：

（1）客户机通过网络将读数据命令发送到服务器。

（2）服务器查询本地缓存，若客户机所请求的数据在缓存中，则经网卡将其直接发送给客户机；否则，将读请求翻译成对本地存储设备的读命令，并将其发向与服务器直接相连的存储设备。

（3）存储设备在收到读数据命令之后，将数据复制到服务器的缓存中。

（4）服务器通过网卡将数据发送给客户机。

客户机按照下列步骤将数据写入存储设备：

（1）客户机通过网络将写命令及要写的数据发送到服务器。

（2）服务器将数据暂时保存在本地缓存中，并向客户机返回写操作成功的标志。

（3）当需要刷新缓存数据时，服务器生成对本地存储设备的写命令，并将其发向与服务器直接相联的存储设备。

（4）存储设备收到写命令后，将数据写入存储设备。

由此可见，不管是读存储设备还是写存储设备，数据都需要经过服务器的存储转发，服务器的负荷较重。研究表明，无论怎样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服务的瓶颈。因此，DAS 结构很难满足数字化时代海量数据存储与传输的实时性要求。

DAS 的优点：简单、便宜，容易安装、部署和管理，存储设备的安全性也容易得到保障。

DAS 的缺点：移植性差。对存储设备的使用和管理依赖于服务器操作系统；扩展困难。一台服务器上能够挂接的存储设备的数量是有限的，如一个 ATA/IDE 接口只能支持两块硬盘，一条 16 位的 SCSI 总线也只能支持 16 块硬盘；共享困难。存储设备由一台服务器所私有，其他机器无法直接访问；服务器容易成为瓶颈。对存储设备的所有访问都需要经过服务器转发；可靠性差。如果服务器发生故障，那么所有的数据访问均会受到影响，甚至会造成系统瘫痪。

注意，这里说的共享有两个层次的含义：

（1）在服务器层次上，一台服务器能否直接访问另一台服务器上的存储资源，或者能否直接利用另一台服务器上的剩余存储资源。如果能够直接访问，则说明在服务器层次上的存储资源是可以共享的。

（2）在客户机层次上，客户机能否访问存储在服务器上的数据，或者说，多台客户机能否同时访问存储在一台服务器上的数据。

对 DAS 来说，客户机层次上的共享是可以做到的（通过数据库管理系统或网络文件系统），但服务器层次上的共享是难以做到的，虽然存在两台服务器共用一套磁盘阵列的现象（该磁盘阵列不能被第三台服务器直接访问）。

1.2　存储区域网 SAN

为了解决 DAS 的扩展性、可靠性和共享性问题，人们提出了存储区域网 SAN（Storage Area Network）的概念。

在 DAS 中，存储设备直接与服务器的总线相联，一个存储设备只属于一台服务器，不能被别的计算机直接访问，其存储资源不容易被共享；服务器总线的能力有限，不能在其上挂接过多的存储设备，其存储容量不容易被扩展。上述两点限制了 DAS 结构的应用范围。

如果能够将存储设备从与之相联的服务器总线上摘下，组成一个专用的存储区域网络，那么，该网络上的存储设备就可以被与之相连的所有计算机直接访问，其存储资源就可以被多台服务器共享；存储网络上可以连接的设备数不受限制，其存储容量可以无限扩展。所以说，存储区域网（SAN）能够有效地解决 DAS 的扩展性问题。

SAN 是一种高速的、专门用于存储操作的网络，通常独立于计算机局域网（LAN）。SAN 将主机和存储设备连接在一起，能够为其上的任意一台主机和任意一台存储设备提供专用的通信通道。事实上，连接到 SAN 上的任意一台主机都可以看见该 SAN 上的任意一台存储设备，可以像访问本地磁盘一样访问 SAN 上的任意一台存储设备，也就是说，SAN 上的任意一台存储设备都可以被该 SAN 上的任意一台主机直接访问；对 SAN 上的任意一台主机来说，整个 SAN 上的存储设备都是它的外部存储设备。

SAN 将存储设备从服务器中独立出来，实现了服务器层次上的存储资源共享。SAN 将通道技术和网络技术引入存储环境中，提供了一种新型的网络存储解决方案，能够同时满足吞吐率、可用性、可靠性、可扩展性和可管理性等方面的要求。SAN 的提出使服务器和存储设备之间的连接方式发生了根本性的变化。

1.2.1　SAN 的结构

图 1.2 是 SAN 的结构示意图。

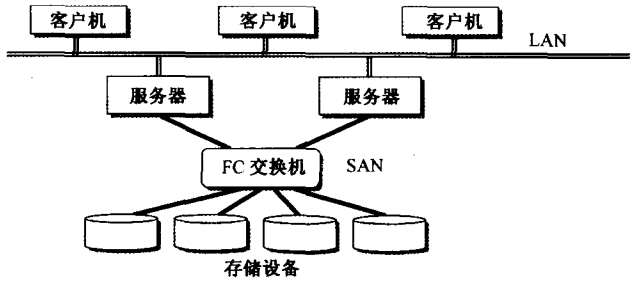

图 1.2　SAN 的系统结构

由图 1.2 可见，在存储区域网络中，服务器上通常配置两个网络接口适配器：与 IP 网络连接的普通网卡，服务器通过该网卡与客户机交互；与 SAN 连接的主机总线适配器 HBA（Host Bus Adaptor），服务器通过该适配器与存储设备交互。

HBA 是服务器内部 I/O 通道与存储系统 I/O 通道之间的物理连接。最常用的内部 I/O 通道是 PCI 和 Sbus，它们是服务器 CPU 和外围设备的通信协议，在主机主板上实现了这种通信协议。最常见的存储系统 I/O 通道是 IDE、SCSI 和光纤通道（FC），它们各自采用自己的协议实现存储系统与主机之间的通信。存储设备上通常有控制器，控制器可实现一种或几种通信协议，可以实现从 IDE、SCSI、FC 等存储协议到物理存储设备的操作协议之间的转换。服务器内部需要一种设备（扩展卡或主板上的集成电路）来实现内部通道协议（PCI、Sbus 等）与存储系统通道协议（IDE、SCSI、FC 等）之间的转换，这种设备就是 HBA。图 1.3 是存储通道结构的示意图。

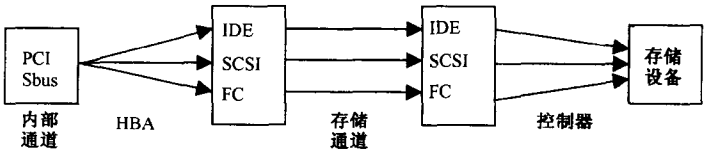

图 1.3　存储通道

内部通道到 IDE 的转换器通常集成在主板上，不需要专门的适配器。内部通道到 SCSI 的转换器就是 SCSI 卡，它是一种常见的 HBA。内部通道到 FC 的转换器就叫做 HBA，它实现了 FC 协议中的 FC-0、FC-1 和 FC-2 层的功能。

SAN 的主体是存储网络连接设备。它的作用是连接服务器和各种存储设备。通常使用的连接设备有光纤通道（Fibre Channel，FC）集线器、交换机或路由器。光纤通道集线器类似于局域网中的 HUB，所有的设备都连接到集线器的端口上，集线器内部把它们连接成环形。因此，由集线器连接的光纤通道网络实际上是一个仲裁环。不管仲裁环上有多少设备，在同一时刻只能有两个设备占有环而进行通信，相当于所有设备共享一个光纤通道带宽。光纤通道交换机避免了共享带宽的问题，在同一时刻，连接到交换机上的设备可以两两通信，因此，交换机能让任意两个设备都拥有一个光纤通道带宽。存储路由器用于连接存储网络和基于 IP 的计算机通信网络，可以实现存储协议与 IP 协议之间的转换。存储路由器可以支持多种协议，如 FC、IP、iSCSI 等。

当然，也可以使用其他形式的技术，如 Infiniband、SSA、SAS、iSCSI 等，构造存储区域网络。Infiniband、SSA、SAS、iSCSI 等都已经成为了标准，但 FC 是事实上的标准。

SAN 中的存储设备有很多种，如盘阵（RAID）、盘堆（JBOD）、光盘库、磁带库等，一般来说，这些设备都有比较大的存储容量，比较好的存取性能，比较高的可靠性。

要构造一个 SAN，除了上述那些硬件设备之外还需要一定的软件。SAN 管理软件用于管理 SAN 中的各种设备，提供存储系统和应用程序之间的编程接口以及存储系统和管理人员之间的人机界面。存储软件的研究热点是存储虚拟化（Virtualization）和可视化（Visualization）。

对 SAN 上存储设备的使用是比较简单的。只要管理员允许，SAN 上的任何一台主机都可以直接访问其上的任何一台存储设备，可以直接存取其上的任何一个数据块并可以在其上建立文件系统。对 SAN 上的主机来说，SAN 是一个扩充了的本地总线，其上的所有存储设备都是它的块设备。