HZ BOOKS

Building Storage Networks, *Second Edition*

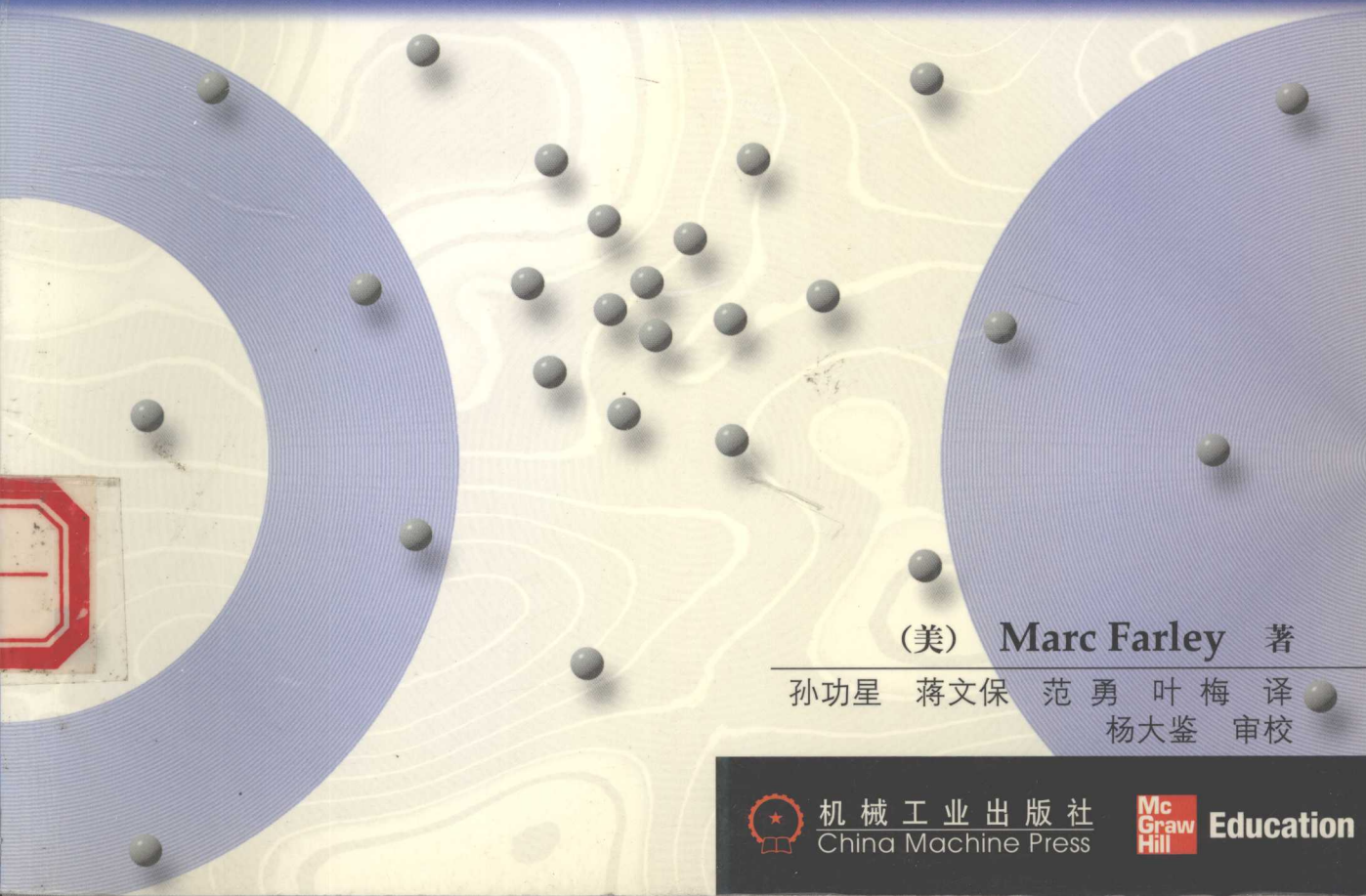
网络专业人员书库

"这不仅仅是一部参考书，SNIA之所以强力把它推荐给读者是因为这本书反映了最新的技术趋势，对于所有关注存储网络的人而言，这样一本书是绝对不容错过的。"

——Larry Krantz，存储网络行业协会（SNIA，Storage Networking Industry Association）主席

（第2版）

SAN 存储区域网络



（美）Marc Farley 著

孙功星 蒋文保 范勇 叶梅 译

杨大鉴 审校

机械工业出版社 China Machine Press

Mc Graw Hill Education

网络专业人员书库

SAN存储区域网络

（第2版）

（美） Marc Farley 著

孙功星 蒋文保 范 勇 叶 梅 译

杨大鉴 审校

机械工业出版社
China Machine Press

本书的第1版写得非常成功，在此基础上，作者除第6章和第7章外，对全书进行了彻底的更新。本书从I/O通道入手，分析了存储系统的各个成分，包括存储设备、I/O控制器和各种存储网络连接设备的应用特性，以及RAID、镜像、备份与恢复等技术。另外，还介绍了SAN和NAS的体系结构和各种应用，并专门讨论了Internet存储技术，对存储网络的基本功能（连接、存储和文件组织）进行了详细的技术分析，增加了InfiniBand和iSCSI等前沿技术方面的内容。

存储网络是一个正在兴起的全新领域，本书将成为读者的极有价值的参考资料。

本书版权登记号：图字：01-2001-3954

图书在版编目（CIP）数据

SAN存储区域网络（第2版）/（美）法利（Farley, M.）著；孙功星等译. -北京：机械工业出版社，2002.4

（网络专业人员书库）

书名原文：Building Storage Networks, Second Edition

ISBN 7-111-09979-6

Ⅰ.S… Ⅱ.①法… ②孙… Ⅲ.信息存储-计算机网络 Ⅳ.TP393.0

中国版本图书馆CIP数据核字（2002）第014270号

机械工业出版社（北京市西城区百万庄大街22号 邮政编码 100037）

北京第二外国语学院印刷厂印刷·新华书店北京发行所发行

2002年4月第2版第1次印刷

787mm×1092mm 1/16·26.25印张

印数：0 001-4 000册

定价：55.00元

新版译者序

在当今网络时代，存储技术正在发生着革命性的变化，并进入了一个全新的时代。SAN（存储区域网络）是近来业界十分引人注目的技术，它一方面能为网络上的应用系统提供丰富、快速、简便的存储资源，另一方面又能对网上的存储资源实施集中统一的管理，成为当今理想的存储管理和应用模式。

本书从I/O通道入手，分析了存储系统的各个成分，包括存储设备、I/O控制器和各种存储网络连接设备的应用特性，以及RAID、镜像、备份与恢复等技术。另外，还介绍了SAN和NAS的体系结构和各种应用，并专门讨论了Internet存储技术。本书作者在网络存储领域拥有十多年的工作经验，他曾经是一位系统工程师，精通存储技术，了解该领域的发展方向。2001年作者在第1版的基础上写出了该书的第2版，第2版除第6章、第7章内容外，对全书进行了彻底的更新，对存储网络的基本功能（连接（wiring）、存储（storing）和文件组织（filing））进行了详细的技术分析，并增加了InfiniBand和iSCSI等前沿技术方面的内容。

众所周知，具有深远意义的Web技术诞生于高能物理应用领域。同样，译者所在的中国科学院高能物理研究所，曾经开通了我国连入Internet的第一条专线，并设立了国内第一个Web服务器。实际上，高能物理领域对计算技术的需求历来走在时代的前列。目前，随着对微观世界认识的不断深入，我国高能物理发展对网络计算和存储环境提出了新的挑战。为了适应这种挑战，高能所计算中心目前正在进行“基于SAN的高速网络计算环境”课题的研究与实施。在该课题的研究过程中，我们深深感到了国内这方面资料的缺乏。因此，在华章公司的大力协助下，课题组在杨大鉴研究员的组织下，继2000年翻译了该书的第1版后，又翻译了该书的第2版。我们希望本书将有助于提高我国存储网络技术的应用水平。

本书主要由孙功星、蒋文保、范勇、叶梅具体翻译，其中孙功星翻译了第1～7章，蒋文保翻译了第14章和第16章，范勇翻译了第8～12章，叶梅翻译了第13、15章和第17章。全书由杨大鉴研究员负责统稿和审校。另外，在翻译过程中得到了高能所计算中心主任于传松研究员的大力支持，马梅、许冬、程耀东、田承钧等同志也提供了很大的帮助。

由于存储网络是一个正在兴起的全新领域，译者涉足该领域的时间也不长，加上翻译时间仓促，错误在所难免，欢迎广大读者批评指正。

2001年12月

序

我们正处在信息时代，也是数据存储时代。数据存储量每年翻一番以上，它的“胃口”似乎还在不断地膨胀，而且没有马上会改变的迹象。存储数据类型、诉讼保护及新的规章要求方面的变化，驱动着数据存储的持续增长。

Internet正在改变着每一件事情。这虽然言过其实，但是对于数据存储的增长来说，它却切中要害。富媒体这种新型的数据已经诞生，它包括MP3、动画、彩色图像、PDF、PPT以及视频流等。富媒体的数据量相当于普通文本的3个量级（1000倍）。Internet让用户始料未及的是信息的及时可用性，这表明数据或文件不能被删除，一切都必须达到及时可用。

大型的电子商务和存储在关系数据库中的数据也驱使着数据存储的增长。IDC和Dataquest的报告显示，70%以上的数据管理使用了数据库方案，关系数据库就像倍增器一样，将数据增大到原来大小的3到8倍。

诉讼是导致数据存储增长的另一个原因，电子邮件在诉讼过程中起着越来越重要的作用，律师总是喋喋不休地告诉其代理机构要保存所有的电子邮件。

新的规章要求所有的病人资料需要保存长达7年之久，如美国的Health Insurance Portability and Accountability Act of 1996（HIPAA），因而，仅对MRI、PET和CAT扫描进行存储的需求便是巨大的。

对于大多数信息系统（IS）机构，管理呈指数增长的数据存储是项挥之不去的挑战。同时面临的挑战还有资质人才的严重短缺。IDC预计，到2003年，将有180万个IS位置虚位以待。随着经济增长趋缓，IS的预算被减少，用于存储管理的人员也更少。一个CIO感叹：“我们必须使用更少的人管理更多的存储，直到最终不用一兵一卒就能管理Yottabytes量级数据为止”。

那么，存储网络能够做什么呢？答案是每一件事情。存储的暴涨以及希望高效管理它们的愿望导致了SAN（存储区域网络）和NAS（网络连接存储）的诞生。

SAN的目的是增加IS的工作效率，用更少的人管理大量的存储增长。实现的办法是将存储与服务器分离，存储可以被多个服务器共享，而同时两者都能独立地扩容。分离和集中化存储可以简化管理、控制，也可以提高生产率。

NAS是为了简化基于文件的存储而设计的。它是使用符合工业标准的网络文件系统（NFS）、公共因特网文件系统（CIFS）、TCP/IP及以太网协议完成的，NAS经常也被称为“即插即用”存储。

像大部分简单的概念一样，造成混乱和复杂性的缘由常常是因为厂商的言过其实。因而，是实际的东西，还是广告宣传？让人很难判断，往往是很久以后才能看到结果。

在本书《SAN存储区域网络》第1版中，作者Marc Farley提供了令人难以置信的、深入的信息资源，帮助读者理解存储和存储网络，它消除了行业宣传带来的混乱。使用非专业用语，Marc Farley清楚地阐明并图示了存储及存储连网概念、缘由、实现等。

如果没有市场或没有技术，时间就会静止。自从两年前本书的第1版问世以来，出现了丰富的存储和存储网络技术，新的技术带来了新的行业术语、新的复杂性、新的缩写词，也带来了新的问题和新的混乱。

这些技术中有些蕴涵了光明的前景，有些将被抛弃。市场将决定谁是赢者，谁是输者。在本书的第2版中，Marc Farley论述了这些变化，并提供了前所未有的清晰解释。本书的第2版探讨并解释了存储管理、SAN管理、存储虚拟化（包括基于主机和基于网络的）、通过IP在千兆以太网上存储（包括iSCSI、FCIP、iFCP、mFCP），以及新的NAS/SAN文件系统（包括DAFS和SAFS）。

无论你是一个IS专职从业人员、存储工程师、网络工程师、产品经理、行业分析师，还是一个存储、SAN或者NAS公司的执行经理，你都应该购买本书。本书将成为你爱不释手的极有价值的资源，并伴随你度过数据存储时代。

Dragon Slayer咨询公司总裁

Marc Staimer

前　言

在决定事物是否应该产生的时刻，我更加懂得了，当今的需求是简单、统一。人们勉强地撤出了自身的迷宫，千年的神秘却又迫使他们止步不前。

——摘自Rene Char的《凋落的花瓣》

在刚刚结束《SAN存储区域网络》（第1版）的写作时，我就想改变它。在这个领域里，我了解了许多新的东西和新的前景，我需要把这一切都写下来。但在清楚地表达它们之前，我还需要一点时间来理出头绪。在过去两年里，网络存储行业发生了巨大的变化，因此，我很迫切地想在书中分析这个新的技术。

尽管本书第1版与第2版中的许多材料相同或类似，但与其说它是建立在第1版的基础上，还不如说它是把第1版作为起点。第2版的改动是非常大的，虽然保持着类似的结构，但从第1章开始，就有了彻底的更新。仅存的没有彻底改写的是第6章和第7章，它们分别是关于RAID和网络备份的内容。其余的章节都填充了许多新的材料。

首先，本书围绕着三个基本的存储网络功能——连接、存储和文件组织。尽管这些主题在第1版中以不同的形式出现过，但是并不明确，也没有作为一种方法来分析产品和技术的功能。对于一种新兴的行业，设计一种新的分析方法带有某种危险。但是，我强烈地感到需要做些什么，帮助读者理解网络存储带给我们的混乱。

本书的另一个目标是大大地增加网络的内容，特别是在InfiniBand和以太网/TCP/IP存储方面，如iSCSI。对于前沿的和没有实际产品可供分析的技术，写书是一件相当困难的事情，但是，这就是前沿技术的性质。今天描绘的许多幻想也许明天有些就会变成现实。希望本书的一些描述能够给读者带来更多的启示，更少的误导。

与第1版一样，本书主要是强调对问题的解释，以帮助读者理解技术和体系结构，并能够评估行业的发展趋势。为了达到这个目的，这一版在每章的结尾都附加了练习，以帮助读者检验对所讨论的主要体系结构概念的理解程度。练习适用于不同程度的读者，能满足读者自己进一步研究的需要。

本书的第1版是非常成功的，许多读者对本书的反映良好。对于我来说，没有什么能够比听到这些更好。我真诚地希望本书能够有益于读者理解这个迷人的领域。

Marc Farley

2001年4月23日

目　录

第三部分 存储区域网络

第四部分 连接技术

第一部分　网络存储概述

第1章　存储网络意义

正如许多新的技术一样，网络存储技术勾画出一幅新的前景——这并非基于它现有的能力，而是基于它未来的潜力。正因为如此，许多IT的从业人员虽然正在进行存储网络产品的评估和比较，但有时对其真正的意义却不尽知晓。本书的目的就是要为读者填补这个空白，本章首先将区分三个明显不同的功能元素：连接（wiring），存储（storing）和文件组织（filing），这些将为分析和设计存储网络奠定基础。

1.1　作为商业财富的数据角色的变化

过去20年里，计算领域发生了很大的变化。20年前，许多工作必须由主机系统才能完成，现在这些工作却可以由价格低廉但功能强大的计算机和计算机集群完成。尽管如此，计算机所处理和产生的数据的重要性却没有改变，因为一旦数据丢失，所有的计算能力都变得毫无价值。这就给数据存储工业提出了新的挑战，即必须在廉价的网络上提供对数据进行24×7操作的可靠性和保护。

虽然这个想法看起来很简单，但实现起来却并非易事。经验业已证明，在网络环境下实现存储管理存在着普遍的困难。一个企业的站点可能拥有不同的硬件/软件平台，而每一种平台都有与之相配套的实用管理程序。对于系统管理员来说，要想正确无误地处理所有这些差异，是极端困难的。

网络环境下的存储管理大约有两个办法。其一，利用由服务器提供的机制来实现管理；其二，通过存储产品的直接接口来管理。后者是大部分存储工业公司经常采用的办法。这就导致了一个有趣的结论：数据是有价的独立实体，与访问它的计算机是分离的，因而，也要求它的管理系统与主机系统是分离的。图1-1显示了计算机和它所处理的数据之间的差别，它恰当地反映了本书的主题。

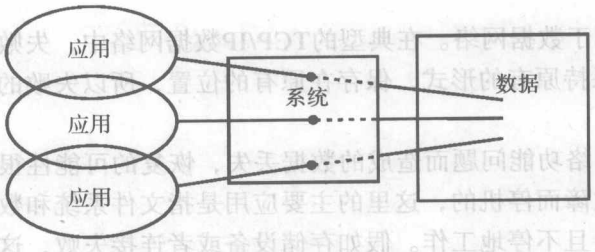


图1-1　计算机及其所处理的数据间的分别

现在人们越来越认识到：数据是一个自由存在的、不必属于任何特定系统的实体，就像资本或智力资产一样，数据也是一种可以共同享用的财富，需要加以保护和存储。同样，作为数据保护和存储管理的平台，存储网络产品和体系结构也提升到战略高度，需要像系统和软件一样，要求定期召开有关会议以制定规划和预算。

1.2　存储网络的基本概念

假想存储网络能在未来提供最好的访问方法，那么一个明显的问题是：存储网络到底是什么？答案可能比预想的要复杂。因为正在开发和营销的产品不止一类，而且每一类产品都不止一种，所以这不仅使用户感到迷惑不解，也使身在其中的存储网络从业人员感到迷惑不解。

1.2.1　存储网络的特点

在着手进行技术讨论之前，我们首先探讨一下存储网络的存在环境，以及它想解决的一些问题。

1. 灵活性

在相对较短的计算历史中，Internet计算可能算是最大的、也是最富变化的环境。在未来的几年内，变化将成为一个永恒谈论的词语。商业机构将通宵达旦地重新产生它们的图像和产品信息，并将这些发布在WWW上，以建立一个全新的、畅通的商业运营模式。

存储网络必须能快速地满足这样的需求：改变系统配置、支持经常的设计以及Web站点的内容变化。任何阻碍这种快速变化的存储实现，都将被Internet数据中心所抛弃。

2. 可扩展性

最常见的系统变化之一是系统的扩展，在最为快速增长的环境中，增长的大小取决于所期望的系统记录和处理信息的能力。因为商业机构正在搜集日益增长的信息，以支持它们运营的各个方面的需要，还需要维护对数据可访问性。因此，存储网络必须提供系统可扩展的解决方案，允许在数据存储容量扩展的同时，服务并无间断。

3. 可用性和访问

事实上，用户可能随时地请求存储的数据。无论是客户服务应用数据、用于研究的数据、销售分析数据，还是任何其他类型的数据，数据的可用性和数据的可访问性都将是保证机构高效地运行的必要条件。因此，存储网络必须提供数据的可访问性，同时，还可以通过重新路由访问到次级存储位置，而防止对本地存储服务的威胁。

4. 可靠性

存储网络完全不同于数据网络。在典型的TCP/IP数据网络中，失败的数据传输并不是一种灾难，因为数据仍然保持原有的形式，保存在原有的位置，所以失败的传输可以在此后的重新传输时而得以恢复。

相反地，因存储网络功能问题而造成的数据丢失，恢复的可能性很小。问题在于主要的应用是如何处理因存储故障而停机的，这里的主要应用是指文件系统和数据系统。这些应用总是认为存储总是可用的，且不停地工作。假如存储设备或者连接失败，这些应用经常会崩溃，而且连同它们运行的系统一同崩溃。因而，存储网络除了提供一般数据网络所拥有的功能外，还

必须能够提供最高级的可靠性。

1.2.2 存储网络的主要成分

当一个新的技术出现在高度竞争的市场上时——存储网络就属于这类情况，构成这类技术的元素总是不太清楚。导致这种模糊性的原因之一是，在尝试确认各种不同方法和应用时所产生的市场信息量。本书所做的分析独立于现有的市场，把存储网络分解成三个主要的成分，包括：

- 连接。
- 文件组织。
- 存储。

这里所提供的分析不同于常见的方法，它们对存储区域网络（storage area network，SAN）和网络连接存储（network attached storage，NAS）技术进行分别处理。SAN和NAS并不是毫无关联的，事实上，它们拥有一些共同的特征，与其说它们互相竞争，毋宁说它们互为补充。使用连接、文件组织和存储三个成分，就可以避免许多产品分类所造成的混乱。此外，当你对存储网络的基本成分以及它们之间的相互作用有了一个清晰的了解之后，设计一个存储网络将变得更容易。

1. 连接

简单地说，连接就是用于存储设备和系统及其他设备相连接的有关的连接性技术。它包括各种各样的技术，如网络布线、主机适配器、网络交换机和集线器，以及诸如网络流控制、虚拟网络、连接聚集和网络安全等逻辑成分。简而言之，它是涉及在存储网络上传输的任何事物。图1-2给出了连接的逻辑和物理成分。

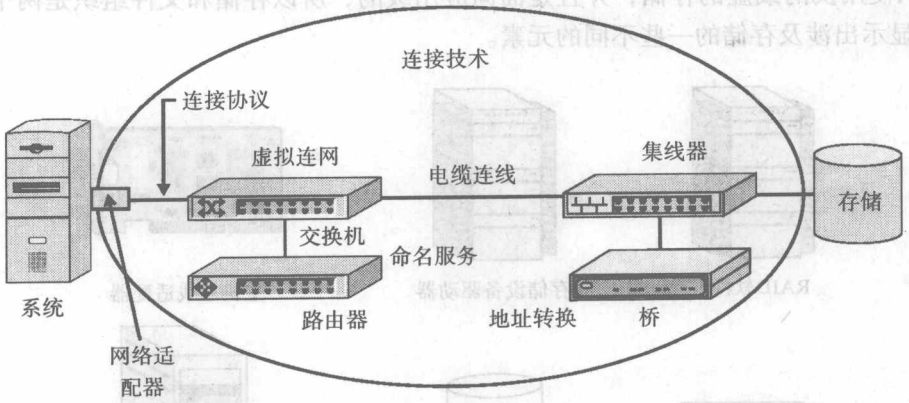


图1-2 连接的共同成分

虽然连接的物理成分容易识别，但逻辑成分辨识起来却相当困难。虚拟网络（或称VLAN）和定义VLAN交换操作的802.1Q是连接技术，低级的网络控制逻辑以及存储网络适配器的驱动程序也是连接技术。然而，更高级的存储协议驱动程序并不是连接技术，如，管理应用对存储的请求和明确描述存储网络通信的存储内容的技术等。

2. 文件组织

文件组织就是组织存储数据的智能过程。一般说来，文件组织由文件系统和数据库系统完成，文件系统和数据库系统确定数据如何被存储和还原，那些额外信息应该与数据存放在一起以描述数据（称做元数据，metadata），以及存储的数据是如何提供给应用和用户，等等。文件组织功能在本质上是逻辑的，换句话说，它并不依赖于硬件。尽管NAS产品也能够包含完美的连接和存储技术，但文件组织是NAS的主要功能。图1-3给出了当前的一些典型的文件组织的实现技术。

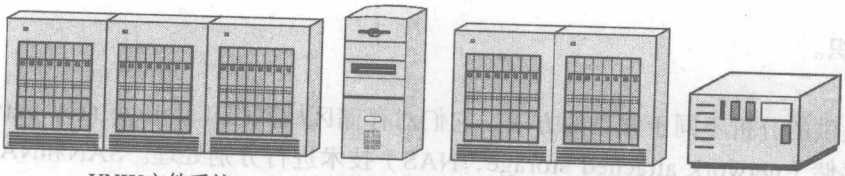


图1-3　存储网络的文件组织的实现

3. 存储

存储技术提供了一个稳定的、非易失的、可靠的保存数据的地方，保证数据能够被重复使用。存储技术既拥有物理成分，也拥有逻辑成分。物理成分包括磁盘驱动器、电源、冷却设备和连接等；逻辑成分包括RAID、镜像、卷管理软件等，卷管理软件的目的是把多个磁盘驱动器转映射成单一的虚拟设备。逻辑成分也包括存储网络适配器的应用级的驱动程序，它用于表示通过存储网络在计算机和存储设备及子系统间传送的命令和数据。因为存储是面向设备级的，而文件组织使用实的或虚的存储，并且是面向应用级的，所以存储和文件组织是两个不同的概念。图1-4显示出涉及存储的一些不同的元素。

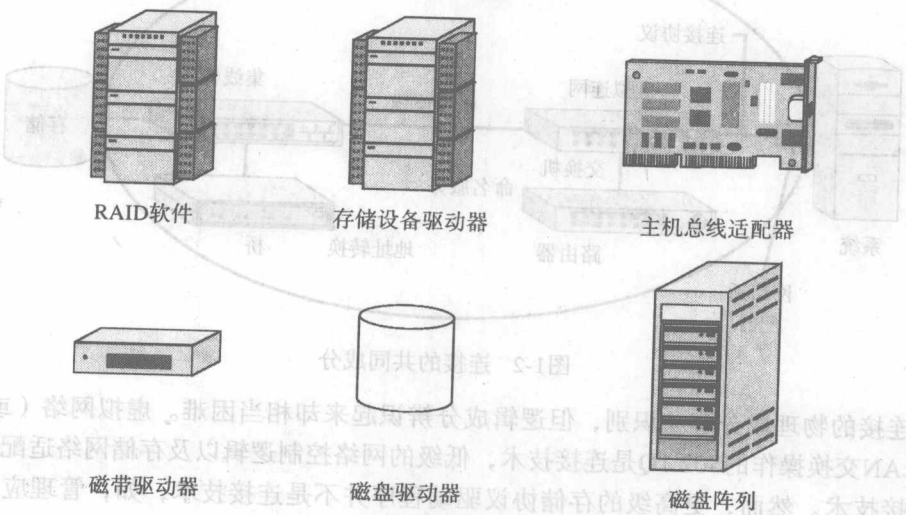


图1-4　存储网络的存储技术

4. 连接、文件组织和存储的组合

图1-5给出了连接、文件组织及存储等在系统中通常的实现位置。传统上说，文件组织功能位于主机计算机系统，但是在NAS环境中，它一般位于某种类型的文件服务器中。理论上讲，文件组织功能可以位于存储网络的几个不同位置，作为一种数据共享技术，在本书中将进行详细的讨论。

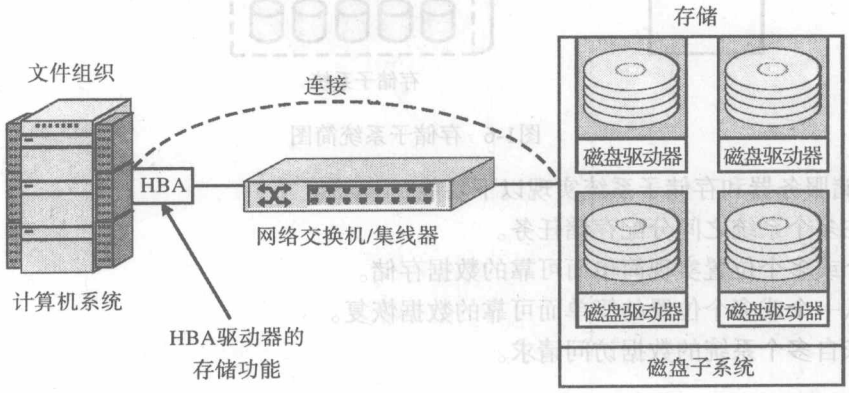


图1-5 文件组织、存储及连接的简单模型

连接功能由网络适配器、线缆以及交换器/集线器所组成，本质上，在计算机系统和磁盘子系统之间的任何东西都可以看做其组成部分。在SAN中，网络适配器被称做主机总线适配器（HBA）；而在NAS环境中，使用网络接口卡（NIC），它用于控制各自网络的操作，因而，也被认为是连接功能的一部分。连接成分还包括控制NIC和HBA网络通信功能的设备驱动程序，这些驱动程序运行在主机系统上。

存储功能主要与磁盘设备和子系统相关，但事实上，它的功能是分散的，其中一个重要的部分是作为设备驱动程序运行在主机计算机系统上。这个设备驱动程序代码是作为一个应用运行在连接功能之上，管理存储命令和数据交换。我们知道，存储功能是使用连接提供的网络服务完成任务，因而，有理由认为存储功能是一个更高级的网络应用。

1.3 传统开放系统的存储方法概述

为了理解网络存储的新技术，有必要首先了解一下已有的传统存储系统，因为网络存储的许多概念是来源于几十年前的传统技术。

本书的“开放系统”是指由UNIX和PC系统所组成的产品，它们支持和接收广泛的工业参与。由于开放系统的计算平台已经成为当今网络工业的主流，因此，由它们来确定网络存储环境也是很正常的。在开放系统领域中，客户/服务器网络计算使数据和处理能力的共享成为可能，这促进了应用开发产业的发展。

网络存储是建立在客户/服务器计算基础上的，它将管理存储和文件系统的负担分摊在计算机系统和存储设备之间，计算机负责数据的处理，而存储设备或子系统负责数据的存储。本书所用的“存储子系统”是指处理器和存储设备的集成，它能提供增加的容量、性能及方便的使