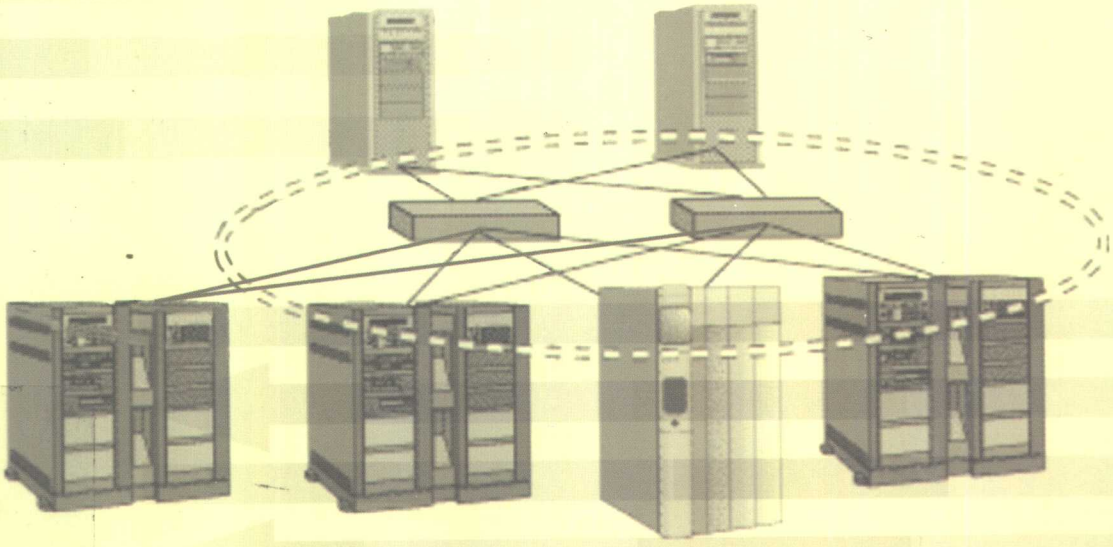
Massive Information Storage

海量信息存储

张江陵　冯　丹 著



科学出版社
www.sciencep.com

海量信息存储

张江陵　冯　丹　著

科学出版社

北京

内 容 简 介

本书是一部关于数据存储系统的著作.书中包含了作者大量的研究成果,并且是首次与读者见面.本书介绍了正在研究并即将流行的包括磁盘阵列、网络磁盘阵列、附网存储、存储局域网等海量存储系统,并对它们的系统结构、组成原理、软件和调度算法设计、容错处理、系统管理和性能测试等多个方面进行了详细、具体和深入的讨论.从理论分析考虑,本书提供了几种分析系统性能的新方法和应用实例;从工程实践考虑,又提供了多种存储系统的设计实例,说明了它们的设计方法、软件流程和调试技术.

本书可供对存储系统进行理论研究、工程实践、系统集成等科技人员使用,也可作为高等学校有关专业的研究生和本科生的教学参考书.

图书在版编目(CIP)数据

海量信息存储/张江陵等著.—北京:科学出版社,2003
ISBN 7-03-010851-5

Ⅰ.海… Ⅱ.张… Ⅲ.信息存储 Ⅳ.TP1842.0101

中国版本图书馆 CIP 数据核字(2002)第 076453 号

责任编辑:鞠丽娜 陈砺川 责任校对:宋玲玲
责任印制:吕春珉 封面设计:王 浩

科学出版社 出版
北京东黄城根北街16号
邮政编码:100717
http://www.sciencep.com
新蕾印刷厂 印刷
科学出版社发行 各地新华书店经销
*
2003 年 1 月第 一 版 开本:B5(720×1000)
2003 年 1 月第一次印刷 印张:17
印数:1—4 000 字数:321 000

定价:29.00 元

(如有印装质量问题,我社负责调换(环伟))

前　言

海量存储一词是从英文 Mass Storage 翻译过来的. 海量存储系统(Mass Storage System, MSS)原本是指与直接存取存储设备连接的后援存储系统,如 IBM3850 MSS 盒式磁带库、光盘库等系统. 这类号称海量的存储系统,其容量不过 1TB(太字节),远非海量,其速度很低,数传率小于 2～3MB/s(兆字节/秒). 本书所讨论的不是这种慢速的后援存储设备,而是由快速的磁盘驱动器组建的直接存取的存储系统,如:磁盘阵列、附网存储系统、存储局域网等. 它们的存储容量可以大到 10TB 甚至 100TB,存取的数传率可高达 400M/s. 目前,真正具备海量存储能力的只有这类系统.

近十年来,虽说数字信息存储技术发展很快,但离"无纸"或不使用书面材料的时代还相距很远. 书面信息和数字信息在相当长的时间内仍将共存共荣,有关支持两者的技术创新都将获得发展. 但是人们已感觉到有用的数据量显现爆炸式增长,对数字信息存储系统的存储容量和速度的要求与日俱增. 特别反映在以下的技术领域,如:书面资料的数字化(数字城市,数字流域,数字地球,数字图书馆,……);信号采集数字化(卫星图像,海洋探测,导弹试验,……);信息交互数字化(电子政务,电子商务,远程教育,……)以及影视数字化等领域. 这些技术领域无一不要求巨大的容量和极高的速度. 同时,一些频繁重复使用或需永久保存的数据还要求很高的可用性(数据不丢失,全天持续服务).

社会需求是推动技术进步的永恒动力,它将推动整个信息存储技术,包括原理、器件、设备、系统和应用多层面的发展. 由于磁记录、量子存储等技术的进步,单台存储设备的容量将超过 1TB,存取速度接近目前的半导体存储器;大规模的磁电阻存储芯片的出现,将兼有 RAM 和 Flash Memory 的功能,并在容量、速度上超过后者. 利用这些设备、器件的新进展,在高速处理器、高速总线和适配器的支持下,存储系统将更上一层楼. 未来的存储系统将迈向三个无限的境界:无限的存储容量、无限的存取带宽和无限的处理能力,即我们称之为 3"I"(Infinite capacity, Infinite bandwidth, Infinite processing power)的技术. 它与计算机的 3"T"(Teraflops computing power, Terabytes main memory, Terabytes I/O bandwidth)技术相互呼应,以满足人类对数字技术的无限需求.

考虑到上述前景,本书选择能体现 3"I"技术发展的数据存储系统作为主要内容,从系统结构、组成原理、软件与算法、可用性与容错、性能分析与评测等各方面给读者以存储系统的专门知识. 内容取自同行学者的学术论著和作者及其合作伙伴的论文(包括博士学位论文). 本书拟在以下方面形成特色:第一,新颖,翔实. 内

容大多取自近十年发表的论文，有些内容还是第一次在书中出现；第二，源于实践．书中程序、算法、分析方法等等大多取自作者及合作者的研究成果，这些研究已应用于实际，并获得过国家级的奖励；第三，重视理论分析．期望通过解析方法引导实践，以求系统品质的改善．

本书共分 12 章，其中第 1、2、3、4、7、8、9、10 章由张江陵撰写，第 5、6、11、12 章由冯丹撰写，全书由张江陵做文字和内容编排的统一处理．在成书过程中得到了华中科技大学计算机学院领导以及师生们的热心支持，书中使用了大量的所在研究组的资料，在此谨致衷心的谢意．

由于作者水平有限，书中难免存在不足之处，敬请读者批评指正．

作　者

2002 年 8 月于武汉

目　录

第 1 章　存储系统导论

本章将对信息存储作一简单扼要的概述，同时对系统的硬件组成与软件成分作较为详细的介绍.

1.1　存储系统概述

存储系统包括两部分：一部分放置在计算机系统内，一般以局域总线与 CPU 连接，除主存外还包括一级或两级高速缓存，它的存储容量较小，而速度很高. 另一部分则放置在计算机系统之外，以外部设备总线连接，它包括直接存取的存储器与后援存储器，它的存储容量很大，而速度相对较低. 本书所称的存储系统主要是指存储容量极大的存储系统. 大容量存储系统，如磁盘阵列(RAID)、网络存储(NAS)、存储局域网(SAN)，已具有相对于主机的独立性，并且它本身也包含了 CPU、Cache 和 Main Memory. 在技术上将包括计算机的内存系统.

存储设备的品种很多，由其组成的系统也不少. 本书将讨论的存储系统主要是指磁盘阵列系统、网络存储系统和存储局域网系统等直接存取的存储系统. 由于这些存储系统不论是集中存储，抑或分布存储都具有大容量、快速和直接存取的特点，其容量甚至可大到 100TB，而以往的海量存储系统主要是指一种联机辅助存储系统，如自动盒带库、光盘库、螺旋扫描的自动磁带系统等，其容量不超过 TB 级，较之由当今出现的大容量磁盘存储器组成的上述存储系统已望尘莫及. 因此，称本书的将要讨论的存储系统为海量信息存储似觉名实相符.

1.1.1　技术特点

信息以数据为载体，对于计算机和网络而言，信息存储就是指二进制数据的存储. 任何能表示和保持两种不同状态的物理现象都可以用于二进制数据的存储. 例如：触发器的两种状态；不同极性的磁化翻转；金属材料的不同金相组织；磁电阻的不同磁化状态；不同折射率的光反射；甚至表面的凸凹形状都能用于二进制数字的存储. 目前广泛使用的只是半导体存储器、磁盘(带)存储器和光盘存储器，并由它们组成存储系统.

信息存储是多学科交叉研究的领域. 就其整体而言，它包括原理、器件、设备、系统和应用五个层面. 存储器的原理和物理特性主要体现在媒体(介质)和器件之中，因涉及的物理现象不同，各种存储器的特性大相径庭. 例如：以大规模集成电路工艺将大量的相同功能的存储单元电路组成阵列，加上外围控制电路所构成的芯

片，它的存储媒体是单元电路，读、写操作和地址译码则通过外围电路实现，因此显示出速度快、容量小、可随机存取、每位信息存储的价格高等特性. 以表面溅射了磁性薄膜的磁盘做媒体，用磁头(MR，GMR)及读写电路读写，通过机械装置寻找存储单元的磁盘存储系统则显示出速度慢、容量大、可直接存取、每位信息存储成本低的特性. 磁盘存储器与光盘存储器有许多类似之处，但他们的存储机理相去甚远. 前者以每次磁化翻转来存储一位信息，因而速度快、翻转密度(记录密度)高，但因需要一定的磁道宽度以便提供足够的回读信号，因而道密度不高；后者用微小的光斑改变存储媒体的光学性质以记录一位信息，因需积聚一定的热能而使速度不快，但其光斑的直径极小，故能获得极高的道密度，而位密度则不及磁化翻转可能做到的程度. 两者在特性上存在诸多差别，所以磁盘存储器可作为直接存取存储器，而光盘存储器只能作为计算机的后备存储. 存储机理及与其密切相关的媒体和读、写头往往依赖物理学和其他学科的基础，从事这类研究的大多是计算机科学与技术之外的学科.

存储设备由精密机械与电子电路组成，按计算机对数据处理的规范与计算机匹配，并要求严格符合接口协议. 近年来，磁盘驱动器和光盘驱动器的机械部分已逐步精简，并采用标准化的部件，而以微处理机为核心的电子电路则功能不断扩展. 大容量存储系统则无论是系统结构和组成的软硬件，或采用的算法和调度策略都与计算机系统相似. 至于大容量存储系统的应用则遍布于计算机、通信等许多方面. 因此，存储设备、存储系统及其应用技术的研究必须建立在计算机科学技术的基础上，并成为计算机科学技术的一个重要分支.

1.1.2　载体属性

信息的载体是多种多样的，从传输和存储的角度看，可以概括为模拟的与数字的两类. 本书讨论的是数字信息的存储与存取. 因此首先对作为信息载体的数据的性质作一点描述.

在计算机科学技术领域中，数据是指具有离散值的数或符号，而信息则是这些具有离散值的数和符号所包含的意义. 可见数据的性质和功能与信息有一定的关系. 通常数据有以下的性质和属性：

独立性　数据是独立的实体. 它是自由存在的，不属于任何特定的系统. 它像有价证券和珍藏的书卷、画卷一样，也是一种独立的资产，并可供社会共同享用.

价值的相对性　数据是有价值的. 其价值与所载信息相关，但其相对性很难精确地确定. 例如，一份因特网语音压缩、解压缩的源代码，一个实时操作系统或一个开发系统，他们的定价是明确的. 但是，在网上发布的一份资料或广告，它的价值则是不确定的. 除此之外，有些数据如病毒，则不仅无益，反而十分有害.

流动性　数据在网上是可以流动的. 流动时它所显示的特性如同流体一样，可以用速度和流量等指标衡量. 例如，对于音、视频数据流，可以用数据传输率、帧率、

传送字节数等参数计量.

可重用性　数据可以重复使用,可以拷贝,可以多次读取,如同一份书面资料可以复印,可以多次翻阅,甚至同时提供给多处使用.

数据的这些性质和功能对数据的存储有很大的影响.为满足或发挥这些性能,要求数据存储系统具有独立、共享、大容量、高速度、容错、拷贝和迁移等等功能.

为使数据有序地存储和被用户使用,它的特性常以一定的格式予以标识.例如,对于一组数据,标志了它的记录长度、记录格式、数组名称、存储装置的类型、媒体的类型、数据产生的日期等等,也有人称这类标识为数据的属性.

由于数据具有上述的性质和功能,因而对数据的需求十分广泛和迫切,由此导致对信息存储系统的需求与日俱增.

1.1.3　社会需求

对存储数据的需求是十分广泛的,就像人们需要容器装载物质和需要纸张记载文字一样,需要存储设备存储数据,而且对容量和速度的要求越来越高.20 世纪 80 年代中期,天气预报和大型科学计算所需要的容量不过 1GB 级,运行速度也较低,约为 1Gflops.时至今日,有关人类基因,全球气候变化,各种科学计算等挑战性问题需要内存的存储容量超过 1TB,运算速度大于 1Tflops,I/O 带宽也要求在 TB 左右,即所谓计算机系统的"3T"性能目标.因此,对外存储系统提出了相应的要求,下面列举部分课题对数据存储容量的要求:

图书馆数字化	＞ 3TB
卫星数据采集(地面站)	＞ 1TB
雷达信号数字化	＞ 500GB
网络教育	＞ 1TB
电视的非线性编辑	＞ 1TB
电视台的电视播放	1～3TB
深海激光探测	＞ 1TB
气候模型诊断与比较	～ 1TB
核聚变建模与分析	＞ 1TB
电子商务、电子邮件等	0.5～3TB

其中,存储空间大多数耗费在图像存储上.以存储一幅 11 英寸×8.5 英寸的彩色图片为例,若扫描仪的分辨率为 300 点/英寸,每个象素的 R、G、B 分量分别为 1 字节,经数字化后一帧的存储空间需要 25.245MB.若视频流的帧速率为 30 帧/秒,一帧图像的象素为 512×480,每个象素用三个字节表示,则此视频流的数据传输率为 512×480×3×30B/s = 21.1MB/s.按照 MPEG-2 的规范,未经压缩和压缩后的数据传输率和每个小时数据量如表 1.1 所示:

表 1.1　MPEG-2 图像规范

级别、参数		数据传输率(MB/s)	每小时数据量(GB)
高级	1920×1080×30	124.416	447.898
	1920×1152×25	110.592	398.131
高-1440 级	1440×1080×30	93.312	335.923
	1440×1152×25	82.944	298.598
基本级	780×480×29.79	22.306	80.302
	780×480×25	20.736	74.649
低级	352×228×29.79	4.782	17.215

由表 1.1 可知，用于 HDTV(高清晰度电视)的高级和高-1440 级，无论是数据传输率，或者每小时数据量都达到了惊人的地步. 这种要求将促进数据存储系统的新发展.

人们经常喜闻乐见的电影、电视节目，在其制作过程中都必须经历多次审查，包括艺术的、思想的和政治的审查. 每次审查过后，随之而来的是无数次的修改. 如果每次都使用数据压缩方法，减少存储容量，则在解压缩后，由于压缩和解压缩带来的损失，图形将有一定的失真表现. 反复压缩和解压缩的结果是影视片在艺术、思想和政治方面虽已令人满意，但片子的清晰度却难以令人接受. 因此在电视、电影节目制作中，艺术家们希望使用无压缩的非线性编辑系统，这就是存储系统追求高速度、大容量的重要原因之一.

另一方面，网络技术的发展，使得在网上流动的数据量大增. 数据来自连接在网上的数据源，也就是来自数据存储系统. 通常网上的数据源包括分布式数据库、文件服务器、Web 服务器等等，尤以 E-mail 和 Internet 文件传输的数据量最大. 此外，一些公司发布、更新软件的信息量也很庞大. 基于数据安全和使用方便的目的，人们通常一再备份，因而耗费的存储空间非常之多，这也是需求容量急剧暴涨的另一个重要原因.

科学计算和仿真，如飞行动力学、超导建模、以及核爆炸仿真、虚拟现实等，所需的存储容量更是大到惊人的程度，也许这些领域的需求将成为数据存储系统的科技工作者们积累毕生精力也难以满足的.

1.1.4　发展趋势

近几年来，存储设备特别是磁盘存储器的存储容量成百倍地增加，存取速度也提高了一个数量级以上，存储容量从每台驱动器的 270MB 增加到了 160GB，增加了 600 倍；磁盘转速(它影响旋转等待时间和数据传输率)从 3600rpm 提高到了 10000rpm(甚至 15000rpm)，提高了近 3 倍；数据传输率(以下简称数传率)也显著提高，以 SCSI(Small Computer System Interface)接口为例，峰值数传率从 5MB/s

(异步方式)提高到了 160MB/s,不久 320MB/s(甚至 640MB/s)的 SCSI 总线适配器也将上市.磁盘驱动器的性能的大幅提升主要源于使用了磁电阻(MR)和巨磁电阻(GMR)磁头,以及局部响应最大似然(PRML)通道技术.若再将改进了的磁记录介质(如垂直取向的记录介质等)技术加以使用,则单台驱动器的容量将再增加几倍,300～500GB 的磁盘驱动器将指日可待.在这种超大容量的磁盘驱动器出现之后,必然将取代目前使用多台磁盘驱动器组成存储系统的部分市场,人们不禁要问是否还需要存储系统呢?答案是肯定的.由设备组成的系统将长期存在.正如上世纪 70 年代,当半导体存储器单片容量不断增长的时候,曾有人预言以半导体存储芯片构造的存储系统将取代磁盘驱动器.但是,由于寻址逻辑过分庞大,这种设想未能实现.10 年前,当光盘记录以其道密度高的优势曾使人们以为光盘存储器将取代磁盘存储器,结果是光盘存储器只能作为后备存储应用于某些低速的场合,目前尚无法作为直接存取的存储设备用于计算机系统.存储系统之所以将长期存在的理由是以下几点:

第一,任何一种存储设备,当其接入计算机、服务器或直接接入因特网络时,必须具有适配器和控制器,并由计算机、服务器或网络中的系统软件通过设备驱动程序的操作才能对数据进行存取.计算机、服务器或网络的种类十分繁杂,其中的操作系统和总线结构又多种多样,因此接口适配器和控制器十分复杂.在数据的存取过程中,只有由这些软、硬件组成的系统在性能上与存储设备同步发展,才能收到相得益彰的效果.

第二,某些功能,如镜像、容错、抗毁、拷贝、迁移、增速等,只有在系统中才能有效地实现.当然,也有单台存储设备具有这些功能,例如具有可将数据直接拷贝到光盘的磁盘驱动器;具有自管理能力、面向对象存储(OBS)的磁盘驱动器(它利用文件属性进行属性管理,能实现纠错、分配存储空间,备份、镜像/复制、数据移动等诸多功能);基于网络连接安全装置(NASD)的智能磁盘驱动器等等.但是,这些设备的形成主要是嵌入了由系统实现的功能,将系统与设备相加集成为一种产品的结果,并非取代了系统本身的工作.历史上,为了充分利用驱动器上微处理器的富余功能,曾将接口、控制器和驱动器三者结合在一起.目前的市售磁盘存储器就是三者相结合的产品.今后仍将出现各种不同功能的存储设备,它们之中嵌入了许多系统功能,以满足用户的需求.

第三,由于数据的独立性,存储设备与访问它的计算机是分离的,计算机与存储设备将沿着各自的目标发展,两者之间的性能差距和联结方式主要由存储系统来处理.通过存储系统既可以弥补差距,又能规范存储设备与计算机(或网络)的连接.从这个意义上看,介于计算机与存储设备之间的管理、连接和控制的系统是永恒地存在的.

第四,虽然驱动器的容量有了大幅增长,但是对容量与速度的需求也日益增长.当要求的容量超过单台设备的容量,要求的数据传输率超过单台设备的数据传

输率时，或者存储设备处于分布状态时，只有采取组成存储系统的形式才能提供具有单一 I/O 空间的数据存储.

目前，存储系统中的通道技术、并行存取技术、网络存储技术以及智能化都是研究的热点，其追求的目标不外乎扩大存储容量、提高存取速度、保证数据的完整性和可靠性、加强对数据(文件)的管理和组织，因此在系统结构上反映出以下一些特点.

独立性　系统的独立性表现在两方面：一是降低对主机的依赖，通过标准接口与各种形式的主机相连，实现开放的体系结构；二是允许不同厂商、不同品种、不同规格的设备接入系统，只要符合接口协议便可，即实现与设备无关的结构特性. 独立性还有其更新的含义. 既然数据被看作重要的财富，而且要求提供频繁的访问，从网络存储考虑，对于数据的存取又要求具有通信的信息短、传输的数据量大且连续的时间长等特点，因此宜于使专用的后端网络从数据的前端网络独立出来. 由此，信息存储的独立性便对系统结构产生了重大的影响.

可扩展性　存储系统的可扩展性主要是指容量的扩展. 但是，简单地增加容量不是可扩展性的最佳体现，应当是在扩展容量的同时，其他性能也得到了提高或不致降低，即可扩展性也应包括与存储有关的其他性能的扩展.

并行性　并行性反映在两个方面：多个独立的请求可由一组磁盘驱动器并行服务，减少 I/O 请求的排队等待时间，从而提高系统的吞吐量；大块数据的请求可划分为多块，由多台磁盘驱动器共同服务，从而提高系统的数据传输率. 并行性还表现在数据的检、纠错处理和数据的恢复等方面.

实时性　实时响应是所有存储系统的重要性能. 除了硬件的系统结构、器件性能之外，采用嵌入式实时操作系统对读写请求做最优化的调度和管理是大幅度提高系统速度的重要措施. 软件的实时化技术无疑对存储系统的性能改进有重要的意义.

可靠性　由于数据是非常珍贵的资产，为提高其利用率，使用非常频繁. 某些场合，如视频流，要求不间断地存取，通常都是 24 小时不间断地工作. 因此，考虑连续工作的存储系统，其系统结构要比一般的系统有较大的区别.

可维护性　存储系统的价值是相当昂贵的，它是服务器价格的 1～3 倍，因此维修的难易程度是用户评价的重要指标. 在线维护，即允许在线更换已被损坏的磁盘驱动器，并能一边继续工作一边恢复数据方式是很受用户欢迎的. 提供备份盘(Spare 盘)的方式，即当磁盘驱动器失效时，备份盘立即投入使用替代失效盘同时进行数据重建的方式，能自动地维护系统的可靠工作. 此外，单一板卡式的设计可使系统结构紧凑、装卡方便、牢靠，但是它的任何一处微小的故障，可使全系统报废，造成用户的全部投资丧失. 分板的集成方式可以减轻损失的程度，但也存在其他的缺点.

共享性　即系统(网络)提供多个接口，并与多台主机(服务器)连接，达到所存

储的信息被共享的目的. 系统结构影响共享性和可扩展性.

智能性　人们可以从多个方面赋予存储系统以智能，如前面提到的 OBS 磁盘系统、NASD 磁盘系统. 通过自动获取存储设备数据分布状况，实现 I/O 负载与设备状况的自动匹配的自适应存储系统，也是智能性的具体体现.

若赋予存储系统以上特性，必将派生出诸多的高性能存储系统. 因此，以上论述的特点也可视为未来存储系统的发展方向.

1.2　存取途径的硬件组成

数据的存取途径是指从数据源到目的地的数据和命令传输的路径. 数据源和目的地通常是存储器或存储设备. 介于两端之间的物理器件便组成了存取途径的硬件系统，如图 1.1 所示的例子.

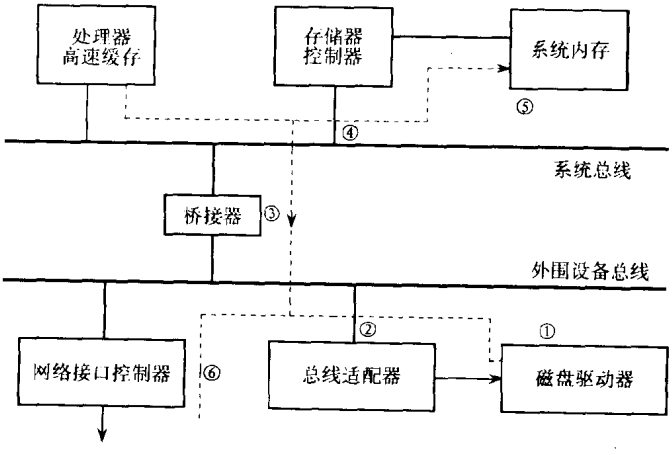


图 1.1　存取途径的硬件成分

一次从磁盘驱动器读取数据的操作，数据需经由总线适配器，外围设备总线，桥接器，系统总线，在存储器控制器的控制下才能写入系统内存，即存取路径由①→②→③→④→⑤组成. 如果这批数据需从服务器上载到网络，供远程用户使用，则在 CPU 处理后再由系统总线，桥接器和网络接口控制器上网，即数据上网的路径由⑤→④→③→⑥组成. 这里将讨论涉及到的硬件环节包括：

系统总线

外围设备总线

总线适配器

网络接口控制器及其他

系统的处理器与高速缓存无疑对 I/O 有很大的影响，但它们不是存取途径的必需部分. 如果 CPU 用作 I/O 控制器，则其影响是直接的. 一般服务器中不采用这