8.1 大容量存储器结构

一. 磁盘

磁盘（magnetic disk）为现代计算机系统提供了大容量的外存。从概念上来说，磁盘相对简单（见图 8.1）。每个磁盘片为扁平圆盘，如同 CD 一样。常用磁盘片的直径为 1.8～5.25 英寸。每个磁盘片的两面都涂着磁质材料。通过在磁片上进行磁记录可以保存信息。

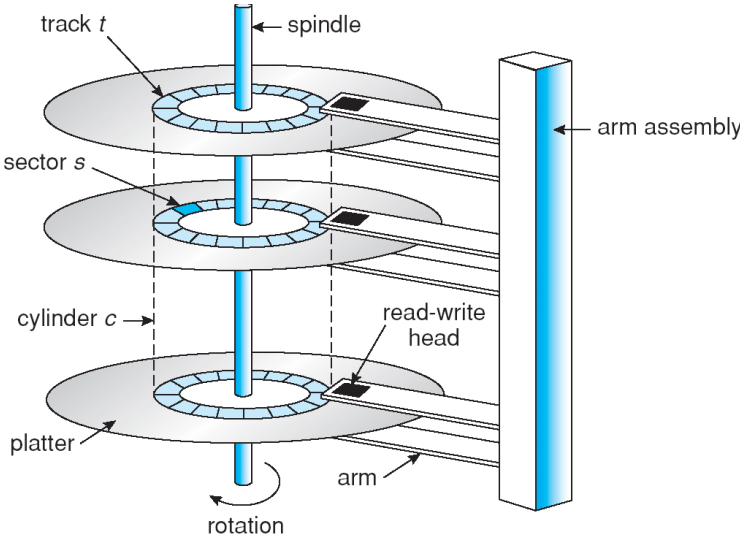

图 8.1 移动磁盘的结构

读写头移动在每个磁盘片的表面之上。磁头与磁臂(disk arm)相连，磁臂能将所有磁头作为一个整体而一起移动。磁盘片的表面被逻辑地划分成圆形磁道（track），磁道再进一步划分为扇区(sector)。位于同一磁臂位置的磁道集合形成了柱面(cylinder)。每个磁盘驱动器有数千个同心柱面，每个磁道可能包括数百个扇区。常用磁盘驱动器的存储容量是按 GB 来计算的。

当磁盘在使用时，驱动器马达会高速旋转磁盘。大多数驱动器每秒可转 60～200 圈。磁盘速度有两部分。传输速率（transfer rate）是在驱动器和计算机之间的数据传输速率。定位时间(positioning time)，有时称为随机访问时间（random access time），由寻道时间(seektime)（移动磁臂到所要的柱面所需时间）和旋转等待时间(rotational latency)（等待所要的扇区旋转到磁臂下所需时间）组成。典型磁盘能以每秒数兆字节的速率传输，寻道时间和旋转等待时间为数毫秒。

磁盘传输速率:正如计算的许多方面，磁盘发布的性能参数与现实中的性能参数是不一样。例如，磁盘所表现的传输速率总是低于有效的传输速率。传输速率是磁盘头从磁性介质读取比特速率，但这个不同于给操作系统传输块的传输速率

磁盘驱动器通过一组称为 I/O 总线(I/O bus)的线与计算机相连。有多种可用总线，包括 EIDE(enhanced integrated drive electronics)、ATA(advanced technology attachment)、串行 ATA（serial ATA，SATA）总线、USB（universal serial bus）、FC(fiber channel)以及 SCSI 总线。被称为**控制器**（controller）的特殊处理器执行总线上的数据传输。**主机控制器**(host controller)是计算机上位于总线末端的控制器。**磁盘控制器**(disk controller)位于磁盘驱动器内。为了执行磁盘 I/O 操作，计算机常常通过内存映射端口，在主机控制器上发送一个命令。

主机控制器接着通过消息将该命令传送给磁盘控制器，磁盘控制器操纵磁盘驱动器硬件以执行命令。磁盘控制器通常有内置缓存。磁盘驱动器的数据传输发生在其缓存和磁盘表面，而到主机的数据传输则以更快的速度在其缓存和主机控制器之间进行。

现代磁盘驱动器可以看做一个一维的逻辑块的数组，逻辑块是最小的传输单位。逻辑块的大小通常为 512 B，虽然有的磁盘可以通过低级格式化来选择不同逻辑块大小，如1024B。一维逻辑块数组按顺序映射到磁盘的扇区。扇区 0 是最外面柱面的第一个磁道的第一个扇区。该映射是先按磁道内扇区顺序，再按柱面内磁道顺序，最后按从外到内的柱面顺序来排序的。

通过映射，至少从理论上能将逻辑块号转换为由磁盘内的柱面号、柱面内的磁道号、磁道内的扇区号所组成的老式磁盘地址。事实上，执行这种转换并不容易，这有两个理由。第一，绝大多数磁盘都有一些缺陷扇区，因此映射必须用磁盘上的其他空闲扇区来替代这些缺陷扇区。第二，对有些磁盘，每个磁道的扇区数并不是常量。

二．磁盘连接

计算机访问磁盘存储有两种方式。一种方式是通过 I/O 端口(或主机附属存储(host-attached storage))，小系统常采用这种方式。另一方式是通过分布式文件系统的远程主机，这称为网络附属存储（network-attached storage）。

1．主机附属存储

主机附属存储是通过本地 I/O 端口访问的存储。这些端口使用多种技术。典型的台式计算机使用 I/O 总线结构，如 IDE 或 ATA。这种结构允许每条 I/O 总线支持最多两个端口，而 SATA 是一种新的简化了电缆连接的类似协议。高端工作站和服务器通常采用更为复杂的 I/O 结构，如 SCSI 或 FC(fiber channel)。

SCSI 是个总线结构，其物理介质通常为带状电缆，具有大量电线（通常 50 或 68）。SCSI 协议在一根总线上可支持 16 个设备。通常这些设备包括主机的一个控制器卡(SCSI 引导器)和 15 个存储设备（SCSI 目标）。SCSI 磁盘是个典型 SCSI 目标，但是协议给每个 SCSI 目标提供访问 8 个逻辑单元的能力。逻辑单元寻址的典型使用是向 RAID 阵列的成员或移动介质库的成员（如 CD 自动换片机向介质切换机制或一个驱动器发送命令）直接发送命令。

FC 是个高速串行结构。该结构可在光纤上或 4 芯铜线上运行。它有两种方式：一是大的交换结构，具有 24 位地址空间。这种方式可望在将来流行，是存储区域网络（SAN）的基础。由于大地址空间和通信的交换特性，多主机和存储设备可以附属到光纤上，获得更灵活的 I/O 通信。另一种是裁定循环（FC-AL），可以访问 126 个设备（驱动器和控制器）。

有多种存储设备可用于主机附属存储。它们包括硬盘驱动器、RAID 阵列、CD、DVD 和磁带驱动器。向主机附属存储设备发出数据传输的 I/O 命令是针对特定存储单元的逻辑数据块的读和写（例如总线 ID、SCSIID 和目标逻辑单元）。

2．网络附属存储

网络附属存储（network-attached storage，NAS）设备是数据网络中远程访问的专用存储系统（如图 8.2）。客户通过远程进程调用接口来访问 NAS，如 UNIX 系统的 NFS 或 Windows 系统的 CIFS。远程进程调用(RPC)可通过 IP 网络（通常为向客户传输所有数据的局域网）的 TCP 或 UDP 来进行。

网络附属存储系统的缺点之一是存储 I/O 操作需要使用数据网络的带宽，因此增网络通信延迟。这一问题对于大客户一服务器环境可能尤为明显——客户与服务器间的和存储设备与服务器间的通信互相竞争。

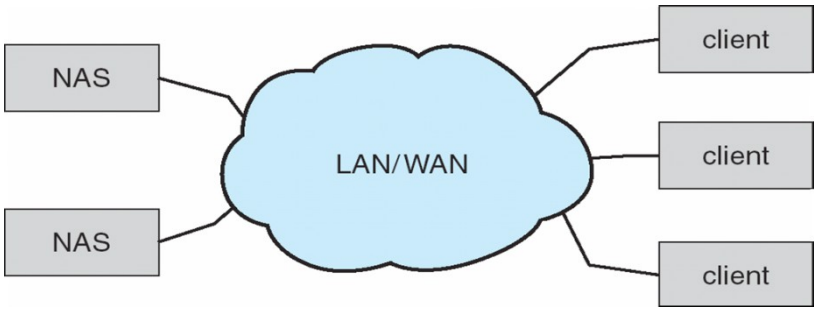

图 8.2 网络附加存储

3．存储区域网络

存储区域网络（storage area network，SAN）是服务器与存储单元之间的私有网络（采用存储协议而不是网络协议），如图 8.3 所示。SAN 的优势在于其灵活性。多个主机和多个存储阵列可以附加在同一 SAN 上，存储可以动态地分配给主机。SAN 的一个开关可以用来允许或者阻止主机和存储之间的访问。举个例子，如果一个主机缺少磁盘空间，那么可以配置 SAN 为该主机分配更多存储。SAN 可以使服务器集群共享同一存储，也可以使存储阵列与多个主机直连。与存储阵列相比，SAN 具有更多数量的端口，以及更少昂贵的端口。 FC 是一种最常见的 SAN 互联。

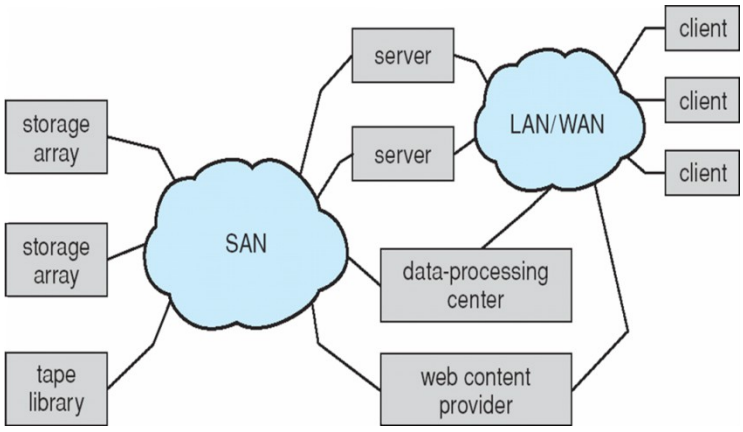

图 8.3 存储域网络