# 第六章 外部存储器

6.1磁盘

磁读写机制

数据组织与格式化

物理特性

磁盘性能参数

6.2 RAID

RAID级别0

RAID级别1

RAID级别2

RAID级别3

RAID级别4

RAID级别5

RAID级别6

6.3固态驱动器

SSD与HDD比较

SSD组织

实践问题

6.4光学存储器

光盘

数字多功能磁盘

高清晰度光盘

6.5磁带

6.6关键术语、复习题和难题

**学习目标**

学习本章后，你应该能够：

* 了解磁盘的关键特性。
* 了解**磁盘**访问所涉及的性能问题。
* 解释**RAID**的概念并描述各个层次。
* 比较和对比普通硬盘和固态硬盘。
* 使用通用术语描述**闪存**的操作。
* 理解不同光盘存储介质之间的差异。
* 概述**磁带**存储技术。

本章研究一系列外部存储器设备和系统。我们先介绍最重要的设备，磁盘。磁盘是几乎所有计算机系统上外部存储器的基础。下一节将介绍如何使用磁盘阵列来获得更好的性能，具体研究被称为RAID（独立磁盘冗余阵列）的系统系列。在许多计算机系统中，扮演越来越重要角色的组件是固态硬盘，后续将对其进行讨论。然后，我们将研究外部光存储器。最后，介绍了磁带。

## 6.1 磁盘

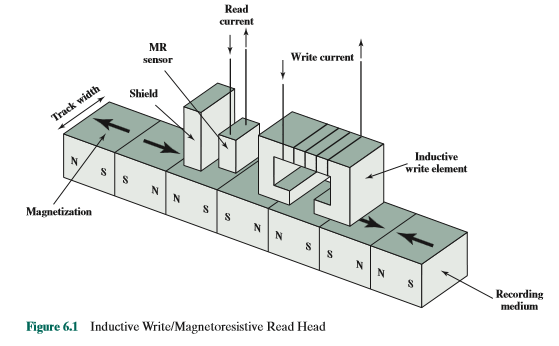
磁盘是一种由非磁性材料制成的圆盘，称为**基片（substrate）**，表面涂有可磁化材料。传统意义上，基片是铝或铝合金材料。最近，已经推出了玻璃基板，这种玻璃基板具有许多优点，包括如下：

* 改善了磁膜表面的均匀性，提高磁盘的可靠性。
* 显著减少了整体表面缺陷，以帮助减少读写错误。
* 支持较低飞行高度的能力(随后描述)。
* 更好的刚度，以减少磁盘动态。
* 更好的抵抗冲击和伤害的能力。

**磁读写机制**

数据通过被称为**磁头**的导电线圈记录在磁盘上，然后从磁盘中检索；在许多系统中，有两个磁头，一个读磁头和一个写磁头。在读或写操作期间，磁头是静止的，而磁盘在其下面旋转。

写入机制利用了流经线圈的电产生磁场这一现象。电脉冲被发送到写入头，并且结果磁性图案被记录在下面的表面，正负电流具有不同的图案。写入头本身是容易被磁化的材料，



其为矩形环形圈状，一侧有一个空隙，另一侧有几圈导线（图6.1）。导线中的电流在间隙中引起磁场，磁场又反过来使被记录介质的小区域磁化。反向电流的方向与被记录介质上的磁化方向相反。

传统的读机制利用了磁场相对于线圈运动会在线圈中产生电流的原理。当磁盘表面在磁头下旋转时，它会产生一个与已经记录的电流极性相同的电流。在这种情况下，用于读取的头部的结构基本上与用于写入的头部相同，因此相同的头部可用于两者。这种单磁头用于软盘系统和较老的硬盘系统。

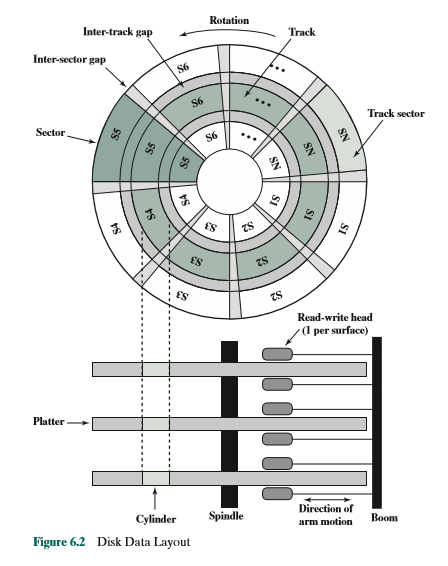
现代硬盘系统使用不同的读取机制，需要单独的读取头，为了方便，读取头定位在靠近写入头的位置。读取头由部分屏蔽的**磁阻（MR）**传感器组成。MR材料具有电阻值，电阻值取决于在其下移动的介质的磁化方向。电流通过MR传感器时，电阻的变化被检测为电压信号。MR的设计使得更高频率的操作成为可能，这等同于拥有了更大的存储密度和更快的操作速度。

**数据组织与格式化**

磁头是一个相对小的装置，能够读出或写入盘子下面转动的部分。这导致盘上的数据组织成一组同心的环，称为**磁道**。每个磁道的宽度与磁头的宽度相同。每个表面有数以千计的磁道。

图6.2显示了这个数据布局。相邻轨道通过**轨道间间隙（intertrack gaps）**分开。这可以防止或至少最小化由于磁头对不准或简单的磁场干扰引起的误差。数据在**扇区（sectors）**中传送到磁盘或从磁盘中传送出来。每个轨道通常有数百个扇区，这些扇区可以是固定长度的也可以是可变长度的。在大多数现代系统中，使用固定长度扇区，其中512字节是几乎通用的扇区大小。为了避免对系统施加不合理的精度要求，相邻扇区被轨道间间隙分开。

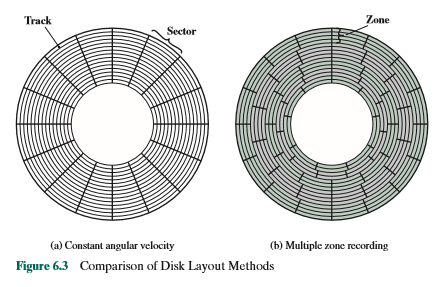
稍微靠近旋转盘中心的一个点通过固定点（如读写头）的速度比在外侧的一点要慢。因此，必须找到某种方法来补偿速度的变化，以便磁头能够以相同的速率读取所有位。这可以通过定义一个可变间隔来实现，此间隔用以记录在磁盘上信息位的间距，这种方法下，外层



磁道的扇区具有更大的间隔。然后，通过以固定速度旋转磁盘，可以以相同的速度扫描信息，该速度被称为**恒定角速度（CAV）**。图6.3a显示了使用CAV的磁盘布局。圆盘分成许许多多扇形的区域和一系列同心轨道。使用CAV的优势在于单个数据块可以被磁道和扇区直接寻址。要将磁头从其当前位置移动到特定地址，只需要将磁头短距离移动到特定轨道，并短暂等待适当的扇区旋转到磁头下。CAV的缺点是，可以存储在较长的外部轨道上的数据量仅仅和存储在较短的内部轨道上的数据量相同。

由于从最外层轨道向最内层轨道移动的过程中，信息**密度**不断增加，其单位为每线性英寸的比特数。所以在直接CAV系统中，磁盘存储容量受限于可以在最内层轨道上存储的最大记录密度。为了最大化存储容量，最好在每个轨道上具有相同的线性比特密度。这将会用到令人难以接受的复杂电路。现代硬盘系统使用了更简单的技术，它近似做到了每条磁道有相等的位密度，该技术被称为多区记录(MZR)，其中磁盘表面被分成若干同心区(典型为16个)。每个区域包含许多连续的轨道，通常数以千计。在一个区域内，每个磁道的位数是恒定的。远离中心的区域比靠近中心的区域包含更多的位（更多的扇区）。区域是这样定义的，即线性比特密度在磁盘的所有轨道上大致相同。MZR允许以稍微复杂一些的电路为代价，来获得更大的总体存储容量。当磁盘头从一个区域移动到另一个区域时，单个比特的长度（沿着轨道）发生变化，导致读写计时发生变化。

图6.3b是一个简化的MZR布局，其中15条轨道被分成5个区域。最里面的两个区域各有两条轨道，每条轨道有9个扇区；下一个区域有3条轨道，每个轨道有12个扇区；最外面的两个区域有4条轨道，每条轨道有16个扇区。



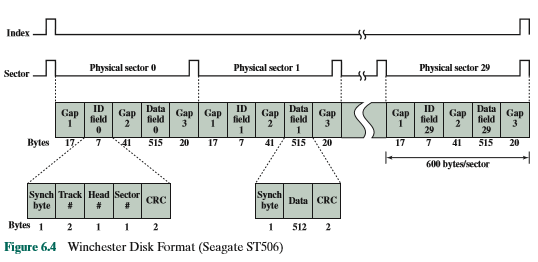
需要找到一些方法来定位轨道内的扇区位置。显然，在轨道上必须有一些起点，以及用以确定每个扇区的开始和结束的方法。这些需求可以通过记录在磁盘上的控制数据来实现。因此，磁盘被格式化为带有一些仅能被磁盘驱动器使用而无法被用户访问的额外数据。

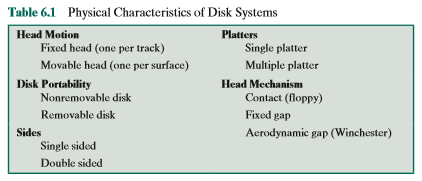
磁盘格式化的一个例子如图6.4所示。在这种情况下，每个轨道包含30个固定长度扇区，每个扇区600字节。每个扇区包含512字节的数据外加对磁盘控制器有用的控制信息。ID字段是用于定位特定扇区的唯一标识符或地址。SYNCH字节是一种特殊位模式，用于限定字段的开头。轨道号标识磁盘表面的轨道。磁头编号标识磁头，由于这个磁盘具有多个表面（目前已解释）。ID和数据域都包含错误检测代码。

**物理特性**

表6.1列出了区别不同类型磁盘的主要特征。首先，头部可以相对于盘片的径向方向固定或移动。在**固定磁头磁盘**中，每个磁道有一个读写磁头。所有的磁头都装在一个横跨所有轨道的磁臂上，这样的系统现在很少见。在**可移动磁头磁盘**中，只有一个读写磁头。同样的，磁头安装在磁臂上。因为头部必须能够定位到任何轨道之上，所以臂可以伸展或缩回。

磁盘本身安装在磁盘驱动器中，磁盘驱动器由磁臂、旋转磁盘的转轴以及输入和输出二进制数据所需的电子设备组成。**不可移动磁盘**永久地安装在磁盘驱动器中；个人计算机中的磁盘是不可移动磁盘。可移动磁盘可以被移除并被替换成另一个磁盘。后者的优点是，有限数量的磁盘系统可以获得无限量的数据。此外，这样的磁盘可以从一个计算机系统移动到另一个计算机系统。可移动磁盘的典型例子有软盘和ZIP盒式磁盘。





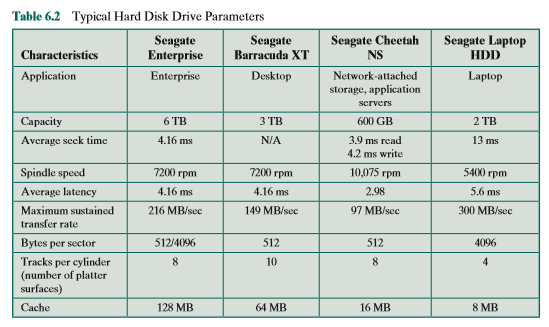
对于大多数磁盘，磁化涂层被涂在到盘的两侧，因而被称为**双面（double sided）**。一些较便宜的磁盘系统使用**单面（single-sided）**磁盘。

一些磁盘驱动器可以容纳垂直堆叠在一起几分之一英寸高度的**多盘磁盘（multiple platters）**。提供多个磁臂（图6.2）。多盘磁盘采用可移动的磁头，每个盘表面都有一个读写磁头。所有的磁头都是机械固定的，所以所有的磁头离圆盘的中心都有相同的距离，并且一起移动。因此，在任何时候，所有的磁头都位于距磁盘中心等距离的轨道上。在盘片上处于相同相对位置的所有轨道组称为一个**圆柱**，如图6.2所示。

最后，磁头机制将磁盘分为三种类型。传统上，读写磁头位于磁盘上方的固定距离，允许有气隙。在另一种极端的磁头机制下，读或写操作期间磁头可以与介质有实际物理接触。这种机制与**软盘**一起使用，软盘是一种小而灵活的磁盘，同时也是所有磁盘类型中成本最低的。

为了理解第三种磁盘，我们需要讨论下数据密度与空气间隙大小之间的关系。磁头必须产生或感测到足够大的电磁场，以便正确地写入和读取。磁头越窄，它就必须越靠近磁盘表面才能发挥作用。更窄磁头的意味着更窄的磁道，因此需要更大的数据密度，这是可取的。然而，磁头越靠近磁盘，由于杂质或缺陷而产生错误的风险就越大。为了进一步推动这项技术，开发出了温彻斯特磁盘。温彻斯特磁头使用了几乎无污染的密封驱动组件。它们被设计成比传统的刚性磁盘头更靠近磁盘表面来操作，从而可以实现更大的数据密度。头部实际上是一个空气动力学箔片，当盘片静止时，轻轻地被搁置在盘片表面上。旋转圆盘产生的气压足以使箔片上升到表面之上。由此开发的非接触系统可以使用比传统刚性盘头更靠近盘片表面的窄磁头。

表6.2给出了当代典型高性能磁盘的磁盘参数。

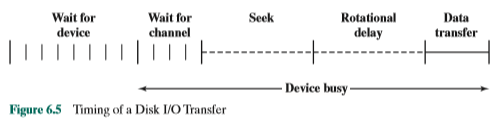


**磁盘性能参数**

磁盘I/O操作的实际细节取决于计算机系统、操作系统和I/O通道和磁盘控制器硬件的性质。磁盘I/O传输的一般时序图如图6.5所示。

当磁盘驱动器工作时，磁盘以恒定速度旋转。要读或写，必须将磁头定位在所需要的磁道上，并在磁道上所需扇区的起始处。寻道包括可移动磁头系统中的磁头移动或固定磁头系统中电子地选择磁头。在可移动磁头系统中，将磁头定位到磁道上所需的时间称为**寻道时间(seek time)**。在任一种情况下，一旦选择了磁道，磁盘控制器即开始等待合适的扇区旋转到与磁头对齐。扇区的开头旋转到达磁头所需的时间称为**旋转延迟**或旋转潜伏期。寻道时间（如果存在）和旋转延迟的总和等于**访问时间（access time）**，即进入读或写位置所需的时间。一旦磁头就位，读或写操作便随扇区在磁头下方移动而执行；这是磁盘操作中的数据传输部分；数据传输所需的时间是**传输时间**。

除了访问时间和传输时间之外，通常还有几个与磁盘I/O操作相关的排队延迟。当进



程发出I/O时请求，它必须首先在队列中等待设备可用。在那时，设备被指派给进程。如果设备与其他磁盘驱动器共享单个I/O通道或一组I/O通道，则可能存在额外的等待该通道可用的时间。此时，执行寻道来开始磁盘访问。

在一些用于服务器的高端系统中，使用了一种称为旋转位置感知（RPS）的技术。其工作原理如下：当发出寻道命令时，通道被释放以处理其他I/O操作。当寻道完成时，设备确定数据何时将旋转到磁头下方。当扇区接近头部时，设备试图重新建立到主机的通信路径。如果控制单元或通道正忙于另一I/O，则重新连接尝试失败，此时设备必须旋转一整圈后才能尝试重新连接，这被称为RPS丢失。这是一个额外的延迟因素，必须添加到图6.5的时间线上。

**寻道时间** 寻道时间是将盘臂移动到所需轨道所需的时间。事实表明很难确定这个时间。寻道时间由两个关键部分组成：初始启动时间，以及访问磁臂达到速度后穿越必要的磁道所花费的时间。不幸的是，穿越时间不是关于轨道数量的线性函数，它还包括稳定时间(将头部定位在目标轨道上后，直到轨道识别确认花费的时间)。

许多改进来自更小和更轻的磁盘组件。几年前，一个典型的圆盘直径是14英寸（36厘米），而今天最常见的尺寸是3.5英寸（8.9厘米），这减少了磁臂必须移动的距离。当代在硬盘上典型的平均寻道时间小于10ms。

**旋转延迟** 磁盘，除了软盘，其旋转速度从3600rpm（用于手持设备，如数码相机）到写本文时达到的20000rpm不等；在20000rpm的速度下，每3ms旋转一次。因此，平均旋转延迟为1.5ms。

**转移时间** 传输到磁盘或从磁盘传输出来的传输时间取决于磁盘的旋转速度，遵循以下公式：



其中：

T = 传输时间

b = 要传输的字节数

N = 轨道上的字节数

r = 旋转速度，即每秒转数

因此，总平均读或写时间Ttotal可以表示为：



其中T是平均寻道时间。注意，在分区驱动器上，每个轨道的字节数是可变的，这使计算变得复杂化1。

注释：

1 将前面的两个公式与 公式（4.1）比较

**一个计时比较** 通过定义上述参数，让我们来看看两种不同的I/O操作，它们说明了依赖于平均值的危险之处。考虑一个磁盘，它标识的平均寻道时间为4ms，转速为1500rpm，有512字节扇区，每个磁道有500个扇区。假设我们希望读取一个由2500个扇区组成的文件，总共1.28M字节。我们想估算一下传输的总时间。

首先，我们假设文件被尽可能紧凑地存储在磁盘上。也就是说，该文件占据了5个相邻轨道上的所有扇区（5个轨道 × 500扇区 / 轨道数 = 2500个扇区）。这就是所谓的顺序组织。现在，读取第一条轨道的时间如下：

平均寻道时间 4ms

平均旋转延迟 2ms

读取500个扇区 4ms

10ms

假设剩下的轨道现在基本上不需要寻道时间就可以被读取。也就是说，I/O操作能跟得上来自磁盘的数据流。然后，我们最多需要处理四个剩余轨道的旋转延迟。因此，每个连续的磁道以2 + 4 = 6 ms读取。读取完整个文件，

总时间＝10＋（4×6）＝34 ms＝0.034秒。

现在让我们计算使用随机访问而不是顺序访问的方法来读取相同数据所需的时间；也就是说，对扇区的访问在磁盘上随机分布。对于每个扇区，我们有

平均寻道时间 4ms

旋转延迟 2ms

读取1个扇区 0.008ms

6.008ms

总时间 ＝ 2500 × 6.008 ＝ 15020毫秒 ＝ 15.02秒

很明显，扇区从磁盘被读取的顺序对I/O性能有巨大的影响。对于读取或写入多个扇区的文件访问，我们可以控制数据扇区的部署方式。然而，即使在文件访问的情况下，在多程序环境中，也会有多个I/O请求竞争同一个磁盘的情况。因此，值得研究如何提高磁盘I/O的性能，使其优于纯粹随机访问的磁盘。这引起了人们对磁盘调度算法的思考，磁盘调度算法属于操作系统的领域，超出了本书的范围（相关讨论，请参阅[STAL15]）。



## 6.2 RAID

正如前面所讨论的，二级存储性能的改善率远远低于处理器和主存储器的速率。这种不匹配可能使得磁盘存储系统成为提高整个计算机系统性能的主要关注点。

与计算机性能的其他领域一样，磁盘存储设计人员认识到，如果只能将一个组件改进到目前的水平，那么通过使用多个并行组件将获得额外的性能增益。在磁盘存储的情况下，这指导了独立并行工作的磁盘阵列的开发。对于多个磁盘，可以并行处理单独的I/O请求，只要所需的数据驻留在单独的磁盘上。此外，如果要访问的数据块分布在多个磁盘上，则可以并行执行单个I/O请求。

通过使用多个磁盘，可以采用多种方式来组织数据，并且可以通过添加冗余来提高可靠性。这可能使开发可在多个平台和操作系统上使用的数据库模式变得困难。幸运的是，业界已经达成了多磁盘数据库设计的标准化方案，被称为RAID（独立磁盘冗余阵列）。RAID方案由七个级别2组成，从零到六。这些级别并不意味着层次关系，而是指明了共享三个共同特征的不同设计体系结构，这三个特征为：

1. RAID是一组被操作系统视为单个逻辑驱动器的物理磁盘驱动器。
2. 在以后的描述中，数据按照一种被称为条带化的方案，分布在阵列的物理驱动器上。
3. 冗余磁盘容量用于存储奇偶校验信息，这保证了磁盘故障时的数据可恢复性。

对于不同的RAID级别，第二条和第三条特性的细节是不同的。RAID 0和RAID 1不支持第三条特性。

术语RAID一词最初是由加州大学伯克利分校的一组研究人员在一篇论文中创造的[PATT88]3。这篇论文概述了各种RAID配置和应用，并介绍了目前仍在使用的RAID级别的定义。RAID策略采用多个磁盘驱动器，并以能够同时访问来自多个驱动器的数据的方式分发数据，从而改进I/O性能，同时使得磁盘容量的增加更容易。

注释:

2一些研究人员和一些公司已经定义了其他级别，但是本节中描述的七个级别是普遍认可的级别。

3在那篇论文中，缩写RAID代表廉价磁盘冗余阵列。术语“廉价(inexpensive)”用于将RAID阵列中相对较便宜的小磁盘与单个大型昂贵磁盘（SLED）进行比较。SLED本质上是过去使用的磁盘，类似的磁盘技术被同时用于RAID和非RAID配置。因此，业界已经采用术语独立（independence）来强调RAID阵列创造了显著的性能和可靠性增益。

RAID提案的独特贡献是有效地解决冗余的寻址需求。尽管让多个磁头和执行器同时操作可以获得更高的I/O和传输速率，但是使用多个设备增加了故障发生的概率。为了弥补这种可靠性的降低，RAID使用存储的奇偶校验信息，使得能够恢复由于磁盘故障而丢失的数据。

现在我们来研究每个RAID级别。表6.3对七个层次进行了粗略的介绍。在该表中，I/O性能在数据传输能力或移动数据的能力以及I/O请求率或满足I/O请求的能力两个方面得到展示，这些RAID级别相对于这两个度量固有地具有不同的性能。每个RAID级别的优点通过深色阴影突出显示了出来。图6.6说明了七种RAID方案的使用，用以存储需要四个没有冗余的磁盘的数据容量。这些图表突出显示用户数据和冗余数据的布局，并指出了各个级别的相对存储需求。在下面的讨论中，我们将参考这些图表。在所描述的七个RAID级别中，通常只使用四个：RAID级别0、1、5和6。

**RAID级别0**

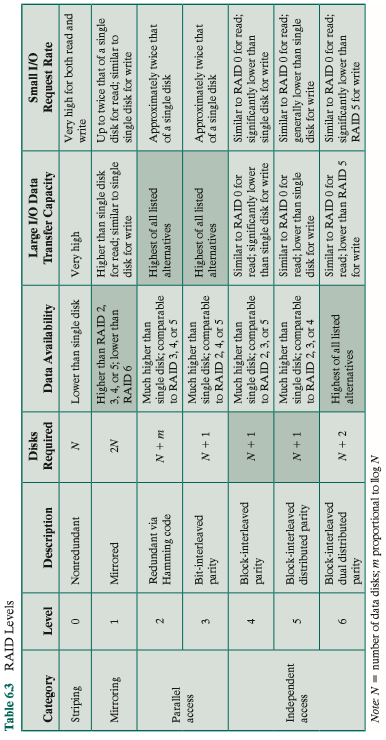
RAID级别0不是RAID家族的真正成员，因为它不包括用以提高性能的冗余。然而，存在一些应用，例如在超级计算机上，其中性能和容量是首要考虑的问题，同时降低成本比提高可靠性更重要。

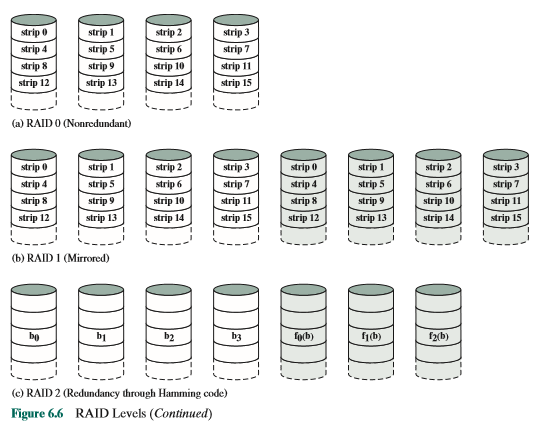
对于RAID 0，用户和系统数据分布在阵列中的所有磁盘上。这比使用单个大磁盘具有明显的优势：如果不同的I/O请求等待两个不同的数据块，那么所请求的块很可能位于不同的磁盘上。因此，两个请求可以并行地发出，从而减少了I/O排队时间。

但是RAID 0，如同所有级别的RAID一样，不仅仅将数据分布在磁盘阵列上：数据在可用磁盘上分条显示。通过参照图6.7可以更好地理解这一点。所有用户和系统数据都被视为存储在逻辑磁盘上。逻辑磁盘被划分为条带；这些条带可以是物理块、扇区或其他单元。条带被循环映射到RAID阵列中连续的物理磁盘上。一组将唯一的一个条映射到每一个阵列成员的逻辑连续条组被称为一个**条带**。在一个n盘的阵列中，每个磁盘上的前n个逻辑条带被物理存储为第一个条，从而形成第一个条带；之后的n个条带被分配为每个盘上的第二个条带；以此类推。这种布局的优点是，如果单个I/O请求由多个逻辑上连续的条带组成，那么最多可以并行处理n个用于该请求的条，从而大大减少了I/O传输时间。

图6.7显示了使用阵列管理软件来在逻辑磁盘空间和物理磁盘空间之间进行映射。该软件可以在磁盘子系统中或主机中执行。

高数据传输能力的*RAID 0*任何级别的RAID的性能都非常依赖于主机系统的请求模式和数据的布局。这些问题可以在RAID 0中被清晰地看到，其中冗余的影响不干扰分析。首先，



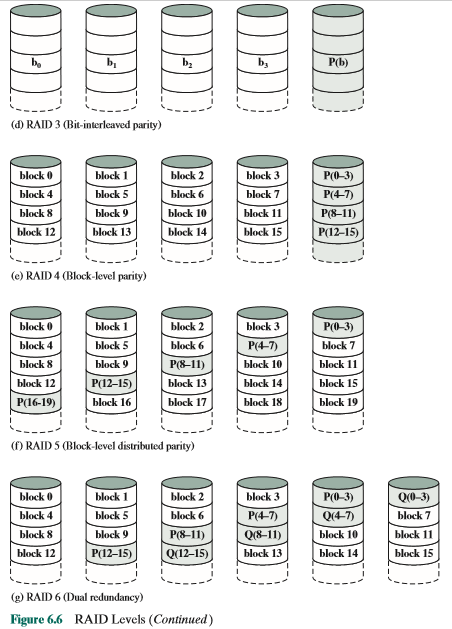


让我们考虑使用RAID 0来实现高速率数据传输。对于体验高传输速率的应用程序，必须满足两个要求。首先，在主机内存和单个磁盘驱动器之间的整个路径上必须存在高传输容量。这包括内部控制器总线、主机系统I/O总线、I/O适配器和主机存储器总线。

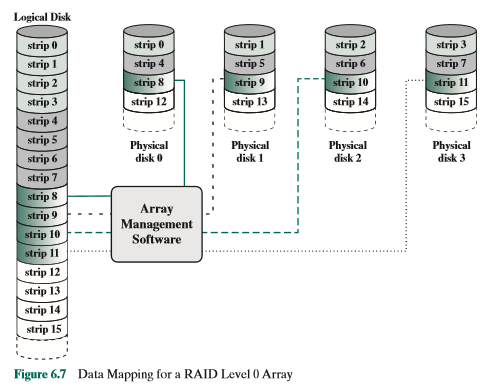
第二个要求是应用程序必须能发出可以有效驱动磁盘阵列的I/O请求。如果与条带的大小相比，典型的请求是针对大量逻辑上连续的数据，则满足此要求。在这种情况下，单个I/O请求涉及从多个磁盘并行传输数据，与单磁盘传输相比，提高了有效传输速率。

针对高I/O请求率的*RAID* 在面向事务的环境中，用户通常更关心响应时间而不是传输速率。对于传输少量数据的单个I/O请求，I/O时间主要受磁盘头的运动(寻道时间)和磁盘的运动(旋转延迟)影响。

在事务环境中，每秒可能有数百个I/O请求。磁盘阵列可以通过在多个磁盘上平衡I/O负载来提供高的I/O执行率。只有存在多个未完成I/O请求时，才能实现有效的负载均衡。



反过来，这意味着存在多个独立应用程序或单个面向事务的应用程序，它们能够执行多个异步I/O请求。性能也会受到条带大小的影响。如果条带尺寸相对较大，是的单个I/O请求仅涉及单个磁盘访问，则可以并行处理多个等待中的I/O请求，从而减少每个请求的排队时间。



**RAID级别1**

RAID 1与RAID级别2到6的不同之处在于其实现冗余的方式。在这些其他RAID方案中，使用某种形式的奇偶校验计算来引入冗余，而在RAID 1中，冗余是通过复制所有数据来简单实现的。如图6.6b所示，在RAID 0中使用了数据条带。但是在这种情况下，每个逻辑条都映射到两个独立的物理磁盘，因此阵列中的每个磁盘都有一个包含相同数据的镜像磁盘。RAID 1也可以在没有数据条带的情况下实现，尽管这不太常见。

使用RAID 1组织有许多优点：

1. 读取请求可以由包含所请求数据的两个磁盘中的任何一个来服务，其读取时间都是最小寻道时间加上旋转延迟时间。
2. 写请求需要更新两个相应的条带，但更新可以并行进行。因此，写入性能由两个写入操作中的较慢写入（即，寻道时间加上旋转延迟时间较大的写入操作）决定。然而，RAID 1没有“写入惩罚”。RAID级别2到6都需要使用奇偶校验位。因此，当更新单个条带时，阵列管理软件必须首先计算和更新奇偶校验位，同时更新所讨论的实际条带。

3. 从失败中恢复是很简单的。当一个驱动器故障时，仍然可以通过第二驱动器访问数据。

RAID 1的主要缺点是成本；它需要它所支持的逻辑磁盘的两倍磁盘空间。因此，RAID 1配置可能仅限于支持存储系统软件和数据以及其他高度重要文件的驱动器。在这些情况下，RAID 1提供所有数据的实时副本，以便在发生磁盘故障时，仍然可以立即获得所有关键数据。

在面向事务的环境中，如果大量请求被读取，RAID 1可以实现高的I/O请求率。在这种情况下，RAID 1的性能可以接近RAID 0的两倍。然而，如果大部分I/O请求是写请求，那么在RAID 0上可能没有显著的性能增益。RAID 1还可以为具有高读取率的数据传输密集型应用程序提供比RAID 0更好的性能。如果应用程序可以分割每个读取请求，以便两个磁盘成员都参与，则性能会继续改进。

**RAID级别2**

RAID级别2和3使用并行访问技术。在并行访问阵列中，所有成员磁盘都参与每个I/O请求的执行。通常，各个驱动器的主轴是同步的，因而每个磁盘头在任何给定时间都处在每个磁盘上的相同位置。

与其他RAID方案一样，数据条带被使用。在RAID 2和3中，条带都非常小，通常与一个字节或一个字一样小。在RAID 2中，使用每个数据磁盘上的对应位计算出纠错码，在多重奇偶校验磁盘中，纠错码的比特被存储在对应的比特位置上。通常，使用汉明码，可以纠正单比特错误并检测双比特错误。

虽然RAID 2比RAID 1需要更少的磁盘，但是它仍然相当昂贵。冗余磁盘的数量与数据磁盘数量的对数成比例。在一次读取时，同时访问所有磁盘。请求的数据和相关的错误纠正码被传递给阵列控制器。如果存在单比特错误，控制器可以立即识别并纠正错误，从而不减慢读取访问时间。在单次写入时，写入操作必须访问所有数据磁盘和奇偶校验磁盘。

RAID 2只有在出现磁盘错误较多的环境中才是有效的选择。鉴于单个磁盘和磁盘驱动器的高可靠性，RAID 2有点大材小用，没有得到应用。

**RAID级别3**

RAID 3以类似RAID 2的方式组织。不同之处在于，无论磁盘阵列有多大，RAID 3只需要一个单独的冗余磁盘。RAID 3采用并行访问，数据通过小条带分布。RAID 3针对所有数据磁盘上同一位置的一组单个位计算出一个简单的奇偶校验位，而不使用纠错码。

**冗余** 如果驱动器发生故障，则访问奇偶校验驱动器，并从其余设备重建数据。一旦更换了发生故障的驱动器，就可以在新驱动器上恢复丢失的数据并恢复操作。

数据重构很简单。考虑一个五个驱动器的阵列，其中x0到X3包含数据，X4是奇偶校验磁盘。第i位的奇偶校验计算过程如下：

x4（i）＝ x3（i）⊕x2（i）⊕x1（i）⊕x0（i）

其中⊕是异或函数。

假设驱动器X1已失败。如果我们把X4(i)⊕X1(i)加到前面方程的两边，我们得到

x1（i）＝ x4（i）⊕x3（i）⊕x2（i）⊕x0（i）

因此，X1上的每个数据条带的内容可以从阵列中剩余磁盘上的对应条带的内容再生。这个原理对于RAID级别3到6都是正确的。

在磁盘出现故障的情况下，所有数据在缩减模式中仍然可用。在这种模式下，对于读取，可使用异或计算动态地重新生成丢失的数据。当数据被写入一个缩减的RAID 3阵列时，为了以后的再生，必须保持奇偶校验的一致性。返回完整操作需要替换故障磁盘，并在新磁盘上重新生成故障磁盘的全部内容。

**性能** 因为数据被划分成非常小的条带，所以RAID 3可以实现非常高的数据传输速率。任何I/O请求都可以从所有数据磁盘并行传输数据。对于大型传输，性能改进尤其显著。另一方面，一次只能执行一个I/O请求。因此，在面向事务的环境中，性能会受到影响。

**RAID级别4**

RAID级别4至6使用独立访问技术。在独立访问阵列中，每个成员磁盘独立操作，从而可以并行满足单独的I/O请求。因此，独立访问阵列更适合于需要高I/O请求速率的应用程序，并且相对地更不适合于需要高数据传输速率的应用程序。

与其他RAID方案一样，仍然使用数据条带。从RAID 4到RAID 6，条带相对较大。对于RAID 4，需要在每个数据磁盘的对应条带上逐位计算奇偶校验条，并且奇偶校验位存储在奇偶磁盘上的对应条带中。

RAID 4在执行小尺寸的I/O写请求时会遇到写惩罚(write penalty)。每次发生写入操作时，阵列管理软件不仅必须更新用户数据，而且还必须更新相应的奇偶校验位。考虑一个由五个驱动器组成的阵列，其中X0到X3包含数据，X4是奇偶校验磁盘。假设执行只涉及磁盘X1上的条带的写入。最初，对于每个比特i，我们有以下关系：

x4（i）＝x3（i）⊕x2（i）⊕x1（i）⊕x0（i） （6.2）

在更新之后，利用由素符号指示的可能改变的比特：

x4’（i）＝ x3（i）⊕x2（i）⊕x1’（i）x0（i）

= X3(i) ⊕X2(i) ⊕X1′(i) ⊕X0(i) ⊕X1(i) ⊕X1(i)

= X3(i) ⊕X2(i⊕)X1(i) ⊕X0(i) ⊕X1(i) ⊕X1′(i)

= x4（i）⊕x1（i）⊕x1’（i）

前面的一组方程推到过程如下所述。第一行显示X1的改变也会影响奇偶磁盘X4。在第二行中，我们添加辅助项⊕x1（i）⊕x1（i）。因为任何量和本身的异或是0，所以这不影响等式。但是，通过重新排序，可以方便地推导出第三行。最后，使用方程(6.2)可将前四项替换为X4(i)。

为了计算新的奇偶校验条带，阵列管理软件必须读取旧的用户条带和旧的奇偶校验条带。然后它可以用新数据和新计算的奇偶校验来更新这两个条带。因此，每个条带写入包括两个读操作和两个写操作。

在涉及磁盘上条带的更大尺寸I/O写操作的情形中，奇偶校验很容易仅通过使用新数据位被计算出来。因此，奇偶校验驱动器可以与数据驱动器并行更新，并且没有额外的读或写。

在任何情况下，每个写操作都必须涉及奇偶校验磁盘，因此奇偶校验磁盘可能成为瓶颈。

**RAID级别5**

RAID 5以类似RAID 4的方式组织。不同之处在于，RAID 5将奇偶校验条分布在所有磁盘上。典型的分配是循环分配，如图6.6f所示。对于一个n磁盘阵列，奇偶校验条带在与前n个条带不同的磁盘上，然后重复该模式。

奇偶校验条在所有驱动器上的分布避免了RAID 4中潜在的I/O瓶颈。

**RAID级别6**

RAID 6是由伯克利的研究人员在随后的论文中提出的[KATZ89]。在RAID 6方案中，执行两种不同的奇偶校验计算，并将其存储在不同磁盘上的不同块中。因此，用户数据需要N个磁盘的RAID 6阵列由N+2个磁盘组成。

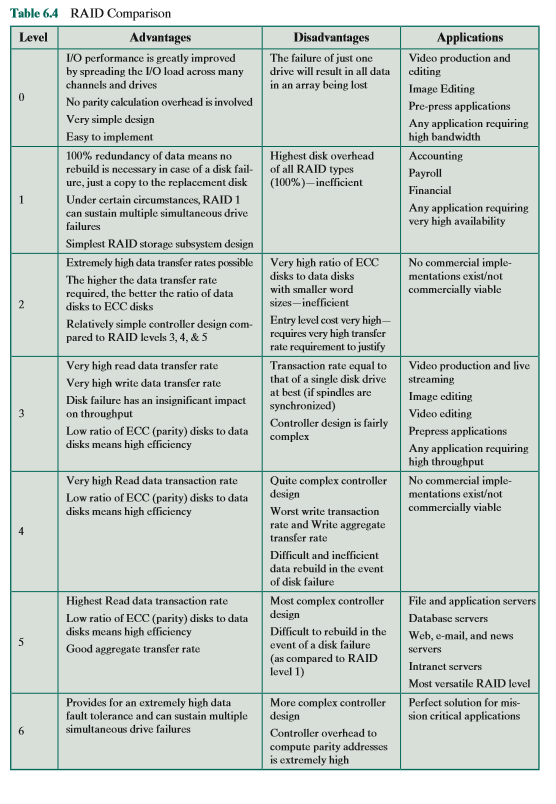
图6.6g说明了该方案。P和Q是两种不同的数据校验算法。其中之一是RAID 4和5中使用的异或计算。另一种是一种独立的数据校验算法。这使得即使包含用户数据的两个磁盘出错，也可以重获数据。

RAID 6的优点在于它提供了极高的数据可用性。三个磁盘只有在MTTR（平均修复时间）间隔内失败，才会导致数据丢失。另一方面，RAID 6会引起大量的写入惩罚，因为每个写入操作会影响两个奇偶校验块。性能基准[EISC07]显示了RAID 6控制器与RAID 5实现相比，在总体写入性能上可能会产生超过30%的下降。RAID 5和RAID 6的读取性能是相当的。

表6.4是七个层次的比较总结。

## 6.3 固态硬盘

近年来，作为计算机体系结构中最重要的发展之一，固态硬盘（SSD）的使用不多增多，补充甚至取代了**硬盘驱动器（HDD）**，它们既被作为内部存储器，又被作为外部的次级存储器。



术语固态是指用半导体构建的电子电路。SSD是一种由固态部件制成的存储器设备，其可以用作硬盘驱动器的替代品。现在市场上和线上使用的SSD使用NAND闪存，在第5章中有关于此的相关描述。

**SSD与HDD的比较**

随着基于闪存技术的SSD的成本下降，性能和比特密度的增加，SSD与HDD的竞争日益激烈。表6.5显示了撰写本文时的典型比较度量。

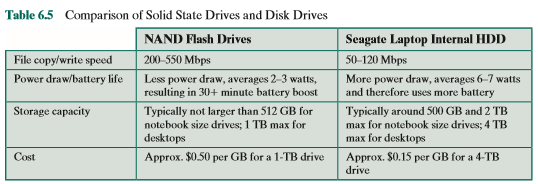
SSD比起HDD，具有以下优点：

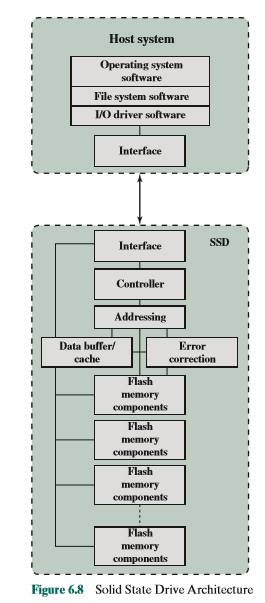
* **高性能每秒输入/输出操作(IOPS)**：显著提高了I/O子系统的性能。
* **耐久性**：受物理冲击和振动的影响小。
* **寿命更长**：SSD不易受到机械磨损。
* **低功耗**：SSD比相同尺寸的HDD功耗要小。
* **更安静和发热更少的运行能力**：所需空间更少，能源成本更低，企业更环保。
* **更短的访问时间和延迟率**: 比旋转的磁盘快10倍以上。：

目前，HDD具有成本优势和容量优势，但是这些差异正在缩小。

**SSD组织**

图6.8展示了与SSD系统相关联的公共体系结构系统组件的一般视图。在主机系统上，操作系统调用文件系统软件来访问磁盘上的数据。文件系统反过来调用I/O驱动程序软件。I/O驱动程序软件提供对特定SSD产品的主机访问。图6.8中的接口组件指的是建立在主机处理器和SSD外围设备之间的物理和电气接口。 如果设备是内部硬盘驱动器，则公共接口是PCIe。如果是外部设备，则公共接口是USB。





除了与主机系统相连的接口之外，SSD还包含以下组件：

* **控制器**：提供SSD设备级接口和执行固件。
* **寻址**：在闪存组件上执行选择功能的逻辑。
* **数据缓冲器/高速缓存**：用于速度匹配和增加的数据吞吐量的高速RAM存储器组件。
* **纠错**：错误检测和纠错的逻辑。
* **闪存组件**：个人NAND闪存芯片。

**实践问题**

存在两个SSD会遇到，而HDD不会遇到的特有的实际问题。第一个问题，随着设备的使用，SSD性能有降低的趋势。为了理解其原因，我们需要知道文件是以页面集合的形式存储在磁盘上，通常长度为4KB。这些页面作为一组连续的页面集合存储在磁盘上，其实并无必要而且实际上并不典型。这种组织的原因我们将在第8章对虚拟存储器的讨论中进行解释。然而，闪存是以块为单位访问的，典型的块大小为512KB，因此每个块通常有128页。现在考虑一下在闪存中写入页面时必须做那些工作。

1. 整个块必须从闪存中读取并放在RAM缓冲区中。然后更新RAM缓冲区中的合适的页。
2. 在块可以被写回闪存之前，整个闪存块必须被擦除----只擦除一页闪存是不可能实现的。
3. 缓冲区的整个块现在被写回闪存。

现在，当闪存驱动器相对为空，同时一个新文件被创建时，该文件的页面被连续写入驱动器，从而影响一个或多个块。然而，随着时间的推移，由于虚拟内存的工作，文件变得碎片化，页面分散在多个块上。随着对驱动器的占用越来越多，碎片也变得越来越多，因此写入新文件可能会影响多个块。因此，从一个块中写入多个页会变得越慢，磁盘的占用就越充分。制造商已经开发了各种技术来补偿闪存的这种特性，例如将SSD的一大部分留出来作为用于写操作的额外空间（称为过度供应），然后在用于磁盘碎片整理的空闲时间删除不活动的页。另一种技术是TRIM命令，它允许操作系统通知SSD哪些数据块不再被考虑使用并且可以在内部擦除4。

闪存驱动器的第二个实际问题是闪存经过一定数量的写入后会变得不可用。当闪存单元受到压力时，它们就失去了记录和保存数值的能力。一个典型的上限是100000次写[GSOE08]。延长SSD驱动器寿命的技术包括，使用缓存前端处理闪存以延迟写操作和分组写操作，使用均匀分布在单元块上的写操作的损耗均衡算法，以及复杂的坏块管理技术。此外，供应商正在部署RAID配置中的SSD，以进一步降低数据丢失的可能性。大多数闪存设备也能够估计它们自己的剩余寿命，因此系统可以预测故障并抢先采取行动。

注释4：

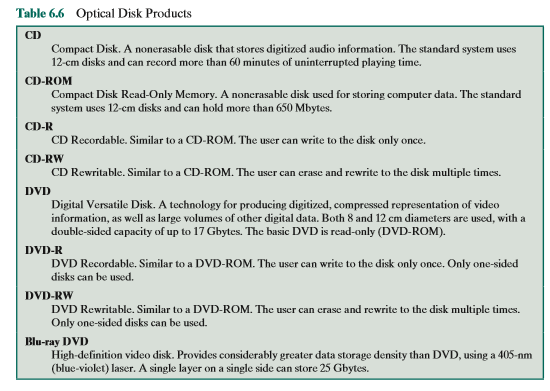
虽然TRIM经常用大写字母拼写，但它不是首字母缩写；它只是一个命令名。

## 6.4 光学存储器

1983年，人们推出了有史以来最成功的消费产品之一：光盘（CD）数字音频系统。CD是一个不可擦除的磁盘，可以在一边存储超过60分钟的音频信息。CD在商业上的巨大成功促进了低成本光盘存储技术的发展，使计算机数据存储发生了革命性的变化。已经提出了各种各样的光盘系统（表6.6）。我们简要回顾一下每一个系统。

**光盘**

*CD-ROM* CD音频和CD-ROM（光盘只读存储器）使用了相似的技术。其主要区别在于，CD-ROM播放器更加坚固，并且具有纠错设备，可以确保数据从磁盘正确地传输到计算机。这两种类型的磁盘是通过相同的工艺制成的。该磁盘是使用一种树脂，例如聚碳酸酯制成。数字记录的信息（音乐或计算机数据）被印在聚碳酸酯表面的一系列微型凹坑中。首先，使用一个精准聚焦的高强度激光来制作主盘。主盘又反过来用于制作模具，将模具复制压印到聚碳酸酯上。然后用高反射面（通常是铝或金）涂覆有坑的表面。这个闪光的表面使用表层明亮的丙烯酸涂层来防灰尘和划痕。最后，可以在丙烯酸树脂上丝网印刷标签。



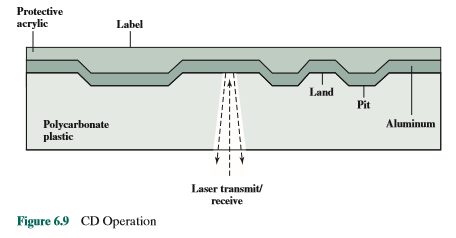
信息是通过低功率激光从CD或CD-ROM中读取的。激光被安装在光盘播放机或驱动单元中。当电机旋转着磁盘通过激光时，激光打在明亮的聚碳酸酯上（图6.9）。当激光遇到**凹坑（pit）**时，其反射光的强度会发生变化。特别地，如果激光束落在表面有些粗糙的凹坑上，激光发生散射并且以低强度反射回光源。坑之间的区域称为**平地（land）**。陆地是一个光滑的表面，它以较高的强度反射激光。凹坑与平地之间的变化被光传感器检测并转换成数字信号。传感器定时测试表面。凹坑的开始或结束位置记录为1；当间隔之间没有高度变化时，记录为0。

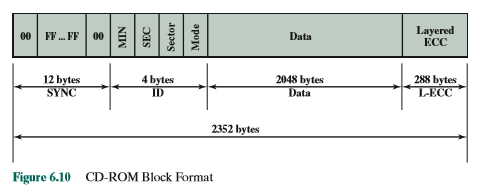
回想一下，在磁盘上，信息是通过同心轨道记录的。对于最简单的恒角速度（CAV）系统，每条轨道的比特数是恒定的。通过**多区域记录**可以实现密度的增加，其中表面被分成多个区域，其中远离中心的区域比靠近中心的区域包含更多的比特。虽然这种技术提高了容量，但仍然不是最佳的。

为了获得更大的容量，CD和CD-ROM不在同心轨道上组织信息。不同于磁盘，光盘有一个单一的螺旋轨道，从中心附近开始，然后螺旋向外绕到磁盘的外边缘。靠近磁盘外部的扇区与靠近磁盘内部的扇区长度相同。因此，信息以相同大小的段均匀地存储在磁盘上，通过可变速度旋转磁盘但以相同速率扫描这些信息。然后激光以**恒定线性速度（CLV）**读取凹坑。磁盘在访问接近外边缘的数据时，比访问靠近中心的数据时旋转的更慢。因此，对于靠近盘外边缘的位置，轨道的容量和旋转延迟都会增加。CD-ROM的数据容量约为680 MB。

在CD-ROM上的数据被组织成一个块序列。典型的块格式如图6.10所示。它由以下字段组成：

* **同步**：同步字段标识块的开始。它由全0的一个字节、全1的10个字节和另一个全0的字节组成。
* **报头**：报头包含块地址和模式字节。模式0表示空白数据字段；模式1表示使用纠错码





和2048字节的数据；模式2表示没有纠错代码的2336字节用户数据。

* **数据**：用户数据。
* **附加字段**：在模式2中表示添加额外的用户数据。在模式1中，这是一个288字节的纠错码。

随着CLV的使用，随机访问变得更加困难。定位特定地址的过程包括将磁头移动到通用区域、调整转速和读取地址，然后进行小范围调整找到并访问特定扇区。

CD-ROM适合向大量用户分发大量数据。由于初始写入过程的开销，它不适合个性化应用程序。与传统磁盘相比，CD-ROM有两个优点：

* 光盘连同其上存储的信息可以被低成本大量复制-----而不像磁盘。磁盘上的数据库的复制必须通过使用两个磁盘驱动器，而且一次复制一个磁盘。
* 光盘是可移动的，这使得光盘本身可用于档案存储。大多数磁盘是不可移动的。磁盘驱动器/磁盘用于存储新信息之前，必须首先将其中的信息复制到另一存储介质中。

CD-ROM的缺点如下：

* 它是只读的，不能被更新。
* 它的访问时间比磁盘驱动器的访问时间长得多，多达半秒。

可记录*CD* 为了适应只需要一个或少量数据拷贝的应用程序，已经开发出了一次写，多次读的CD，被称为**可录制CD（CD-R）**。对于CD-R，磁盘被制备成方便随后用适中强度的激光束写入。因此，与CD-ROM相比，使用稍微贵一点的磁盘控制器，客户既可以写一次磁盘，也可以读磁盘。

CD-R与CD或CD-ROM相似，但不完全相同。对于CD和CD-ROM，信息通过媒介表面的凹痕被记录下来，这些凹痕可以改变反射率。对于CD-R，媒介有一个染料涂层，这种染料被用来改变反射率，并由高强度激光驱动。制成的磁盘可以在CD-R驱动器或CD-ROM驱动器上读取。

CD-R光盘在存储文档和文件方面很有竞争力。它可以永久记录大量的用户数据。

可重写*CD* CD-R可重写光盘可以像磁盘一样重复地写和覆盖。虽然已经尝试了很多方法，但是唯一被证明有吸引力的纯光学方法叫做**相变(phase change)**。相变盘使用在两种不同相位状态下具有两个显著不同的反射率的材料。存在一种无定形状态，其中分子呈现出弱反射能力的随机取向；也存在一种晶体状态，其具有光反射能力好的光滑表面。一束激光可以把材料从一个相位改变到另一个相位。相变光盘的主要缺点是材料最终会永久地失去其所期望的属性。目前的材料可使用500000到1000000个擦除周期。

CD-RW与CD-ROM和CD-R相比具有一个明显的优势，那就是它可以被重写，从而被用作真正的二级存储。因此，它与磁盘竞争。光盘的一个主要优点是，光盘的工程误差远不及大容量磁盘的误差那么严重。因此，它们具有更高的可靠性和更长的使用寿命。

**数字通用磁盘**

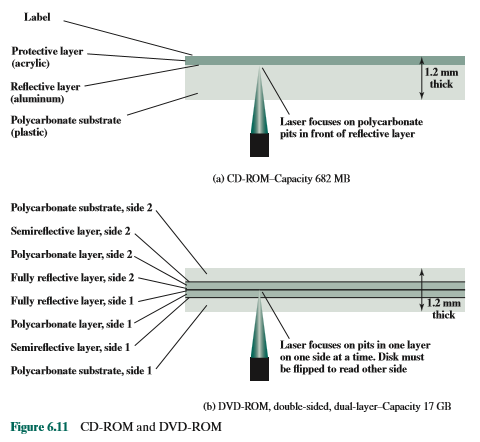
有了大容量的**数字多功能光盘（DVD）**，电子工业终于找到了可接受的模拟VHS录像带的替代品。DVD已经取代了录像机（VCR）中使用的录像带，对于本讨论来说更为重要的是，它取代了个人计算机和服务器中的CD-ROM。DVD把视频带进了数字时代。它可以存储具有令人惊艳的图像质量的电影，并且可以像音频CD一样被随机访问，可以被DVD播放器播放。大量的数据可以堆放在磁盘上，目前的存储量是CD-ROM的7倍。由于DVD的巨大存储容量和生动的存储质量，PC游戏变得更加真实，教育软件可以包含更多的视频。随着这些技术的发展，互联网和公司内部网的流量出现了新的高峰，正是因为DVD被纳入到Web网站。

DVD更大的容量来自于它和CD的三个不同点（图6.11）：

1. DVD上数据位的排布更加紧密。CD上的螺旋线圈之间的间距为1.6μm，沿螺旋的凹坑之间的最小距离为0.834μm。

DVD使用波长较短的激光器，实现了0.74μm的线圈间距和0.4μm的凹坑之间的最小距离。这两项改进的结果是容量增加了大约7倍，达到了大约4.7GB。

1. DVD在第一层之上使用了第二层的凹坑和平地。双层DVD在反射层的顶部有一个半反射层，通过调整焦距，DVD驱动器中的激光器可以分别读取每一层数据。这种技术几乎使磁盘的容量翻了一番，达到大约8.5GB。第二层的低反射率限制了它的存储容量，从而不能实现完全倍增。



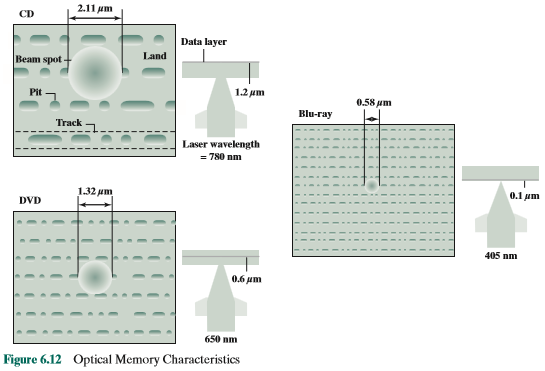
1. DVD-ROM可以是两面的，而数据只记录在CD的一面上。这使总容量达到17 GB。

与CD一样，DVD有可写和只读两种版本（表6.6）。

**高清晰度光盘**

高清晰度光盘被设计成存储高清晰度视频，并且与DVD相比可提供明显更大的存储容量。使用波长较短（在蓝紫光范围内）的激光器可获得更高的比特密度。由于激光波长较短，在高清晰度光盘上的构成数字1s和数字0s的数据凹坑比DVD上更小。

两个相互竞争的磁盘格式和技术最初为了被市场接受而相互竞争：HD DVD和**蓝光**DVD。蓝光方案最终实现了市场支配地位。HD DVD方案可以在单侧的单层上存储15GB。蓝光将磁盘上的数据层定位在更靠近激光的位置（图6.12中每个图的右侧所示）。这使得聚焦更紧密，失真更小，因此坑和轨道更小。蓝光可以在单个层上存储25 GB。有三个版本可用：只读（BD-ROM）、可记录一次（BD-R）和可重新记录（BD-RE）。



## 6.5 磁带

磁带系统使用与磁盘系统相同的读取和记录技术。介质使用的是柔性聚酯（类似于某些服装中使用的材料）并使用可磁化材料涂覆。涂层可以由特殊粘合剂中的纯金属颗粒或汽镀金属膜组成。磁带和磁带驱动器类似于家庭磁带录音机。磁带的宽度从0.38厘米（0.15英寸）到1.27厘米（0.5英寸）不等。胶带过去被包装成敞开的卷轴，必须被穿织在第二主轴上才能使用。今天，几乎所有的磁带都装在暗盒里。

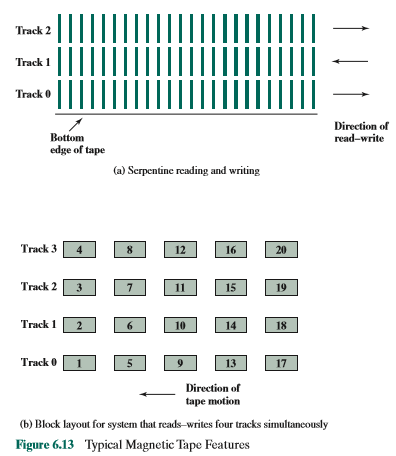
磁带上的数据按长度方向排列为一些平行磁道。早期的磁带系统通常使用九个磁道。这使得一次存储一个字节的数据成为可能，其中第九个轨道是附加的奇偶校验位。接着是使用18或36个磁道的磁带系统，对应于一个字或双字。这种形式的数据记录称为**并行记录(parallel recording)**。大多数现代系统改为使用**串行记录(serial recording)**，其中数据按位序列沿着每个磁道排列，就像磁盘一样。与磁盘一样，数据在磁带上以称为物理记录的连续块读取和写入。磁带上的块通过被称为记录间隙的间隙分开。与磁盘一样，磁带通过格式化来帮助定位物理记录。

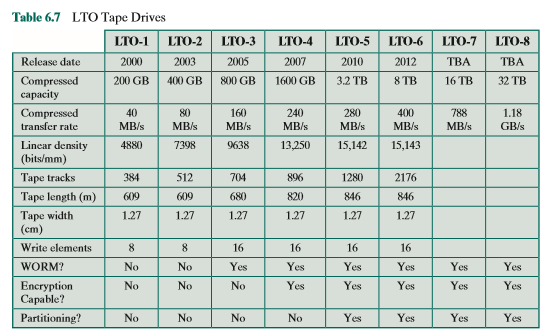
在串行磁带中使用的典型记录技术称为**蛇形记录**。在这种技术中，当数据被记录时，第一组比特位沿着磁带的整个长度被记录。当磁带被读完时，磁头被重新定位来记录新的磁道，磁带再次在其整个长度上记录信息，这次是在相反的方向上。这个过程一直来回进行直到磁带写满为止（图6.13a）。为了提高速度，读写头能够同时读写大量相邻的磁道（通常是两到八个磁道）。数据仍然沿单个轨道串行记录，但是顺序排列的块会被存储在相邻轨道上，如图6.13b所示。

磁带驱动器是一个顺序存取的设备。如果磁带头位于记录1，那么为了读取记录N，必须从物理记录1依次读取到记录N-1。如果磁头当前位置在期望读的记录的后面，则必须将磁带倒带一定距离再向前读取。与磁盘不同，磁带仅在读或写操作期间处于运动状态。

与磁带不同，磁盘驱动器被称为直接访问设备。磁盘驱动器不需按要顺序读取完磁盘上的所有扇区来获得所需的扇区。它只需要在一个轨道内等待中间的扇区，并且可以连续访问任何轨道。

磁带是第一种二级存储器。它依然作为存储器层级结构中成本最低、速度最慢的存储设备被广泛使用。





当今主导的磁带技术是被称为线性磁带打开（LTO）的盒式磁带系统。LTO是在20世纪90年代末，作为市场上各种专有系统的开源替代品开发的。表6.7显示了LTO各个版本的参数。详情见附录J。

## 6.6 关键术语，复习题和难题

**关键术语**

|  |  |  |
| --- | --- | --- |
| 存取时间 | DVD-RW | 光存储器 |
| 蓝光 | 固定磁头磁盘 | 凹坑 |
| CD | 闪存 | platter |
| CD-R | 软盘 | RAID |
| CD-ROM | 间距（gap） | 可移动磁盘 |
| CD-RW | 硬盘驱动器（HDD） | 旋转延迟 |
| 恒角速度（CAV） | 磁头 | 扇区 |
| 恒线速度（CLV） | 平地 | 寻道时间 |
| 圆柱(cylinder) | 磁盘 | 蛇形记录 |
| DVD | 磁带 | 固态硬盘（SSD） |
| DVD-R | 磁阻的 | 条带数据 |
| DVD-ROM | 活动磁头磁盘 | 基底 |
|  | 多区记录 | 轨道 |
|  | 不可移动磁盘 | 传输时间 |
|  |  |  |

**复习题**

6.1 使用玻璃基板来做磁盘的优点是什么？

6.2 如何将数据写入磁盘？

6.3 数据是如何从磁盘读取的？

6.4 解释简单CAV系统和多区域记录系统之间的区别。

6.5 定义术语轨道、*cylinder*和扇区。

6.6 典型的磁盘扇区大小是多少？

6.7 定义术语寻道时间、旋转延迟、访问时间和传输时间。

6.8 所有RAID级别有哪些共同特征？

6.9 简要定义七个RAID级别。

6.10 解释镜像这个术语。

6.11 如何在RAID系统中实现冗余？

6.12 在RAID环境中，并行访问与独立访问的区别是什么？

6.13 CAV和CLV的区别是什么？

6.14 高密度光盘的高比特密度是如何实现的？

6.15 顺序存取设备的缺点是什么？

问题

6.1 假设有一个由九个驱动器构成的阵列，其中X0到X7包含数据，X8是奇偶校验磁盘。给出一个描述数据和奇偶位之间的关系的方程。

6.2 假设有一个磁盘，其中N个磁道编号为从0到（N-1），假设待请求的扇区随机均匀地分布在磁盘上。我们想要计算所穿越的轨道的平均数量。

a. 首先，当头部当前位于轨道t上时，计算长度为j的搜索的概率。提示：所有轨道最终被寻到的可能性都是相同的，这个问题其实是找出所有的组合情况数。

b. 接下来，计算寻道长度为K的概率。提示：这其实是对移动距离为K的所有可能组合求和。

c. 使用下面的期望值的公式来计算一次寻道穿过的平均轨道数



提示：请使用公式：

d.证明：对于大数N，寻道穿过的平均轨道数接近N/3。

6.3 为磁盘系统定义以下内容：

ts = 寻道时间；磁头放置到轨道上的平均时间

r = 磁盘的旋转速度，单位为每秒转数

n = 每个扇区的比特位的个数

N = 轨道的容量，单位为比特

tsector = 访问扇区的时间

使用其他几个量导出tsector的表达式。

6.4 考虑一个磁盘驱动器，它有4个表面，每个表面256个磁道，每个磁道128个扇区。扇区大小为0.5 kB。平均寻道时间为10ms，轨道对轨道访问时间为2ms，驱动器以7200rpm的速度旋转。圆柱体中连续的轨道可以在磁头不动的情况下读取。求：

1. 磁盘容量是多少？
2. 平均访问时间是多少？假设该文件存储的起始位置为磁柱I，轨道0，扇区0，且文件存储在连续磁柱的连续轨道和扇区上。
3. 估算传输一个9-MB文件和一个12-MB文件所需的时间。
4. 什么是突发传输率？

6.5 考虑具有以下参数的单盘磁盘：转速3600rpm；盘一侧的轨道数量：45000；每条轨道的扇区数量：800；寻道时间：每100条轨道花费5ms。假设磁盘接收到访问随机轨道上的随机扇区的请求，并且假设磁盘头从轨道0开始。

a. 平均寻道时间是多少？

b. 平均旋转延迟是多少？

c. 一个扇区的传输时间是多少？

d. 满足一个请求的总平均时间是多少？

6.6 物理记录和逻辑记录之间有区别。**逻辑记录**是作为概念单元处理的相关数据元素的集合，独立于信息的存储方式或位置。**物理记录**是由存储设备和操作系统的特性定义的存储空间的连续区域。假设一个磁盘系统，其中每个物理记录包含40个60字节的逻辑记录。如果磁盘具有固定的扇区，且满足512字节/扇区、100扇区/磁道、每表面125个磁道，同时又10个可用表面，那么需要多少磁盘空间（用扇区、磁道和表面数表示）来存储175000条逻辑记录。忽略任何文件头标记录和轨道索引，并假设记录不能跨越两个扇区。

6.7 考虑一个以每分钟3600转的速度旋转的磁盘。在相邻轨道之间移动磁头的寻道时间是2ms。每个轨道有32个扇区，它们以从扇区0到扇区31的线性顺序存储。从磁头看，扇区呈升序排列。假设读/写磁头位于磁道8上扇区1的起始位置。有一个足够大，达到可以容纳整个一磁道的主存储器缓冲区。通过从源磁道读取数据到主存储器缓冲区，然后将数据从缓冲区写入目标磁道的方式，在磁盘位置之间传输数据。

a. 将轨道8上的扇区1转移到轨道9上的扇区1需要多长时间？

b. 将轨道8的所有扇区转移到轨道9的相应扇区需要多长时间？

6.8 应该清楚的是，与I/O请求大小相比，当条带大小较小时，磁盘条带可以提高数据传输速率。还应该清楚的是，RAID 0相对于单个大磁盘提供了改进的性能，因为多个I/O请求可以并行处理。然而，在后一种情况下，是否需要磁盘条带？也就是说，与没有分条的相当的磁盘阵列相比，磁盘分条是否提高了I/O请求率性能？

6.9 考虑一个6驱动器，每驱动器磁盘阵列150 GB大小的RAID阵列。每个RAID级别0、1、3、4、5和6的可用数据存储容量分别是多少？

6.10 对于CD，音频被转换成按照16位采样的数字信号，并且被处理为8位字节流以便存储。存储这种数据的一个简单方案，称为直接记录，是将1表示为平地，将0表示为凹坑。相反，每个字节被扩展成一个14位二进制数。结果发现，在16134（214）个14位数中，有确切的256（28）个数在每对1之间至少有两个0，而这些是选择用于从8位扩展到14位的数字。光学系统通过检测从凹坑到平地或从平地到凹坑的过渡来检测1的存在。它通过测量两个变化强度之间的距离来检测0的存在。这个方案要求没有连续的1，因此使用第8到第14位的代码。

该方案的优点如下。对于给定的激光束直径，存在一个最小的凹坑尺寸，而不管比特如何表示。用这个方案中，这个最小坑大小存储3比特，因为每1比特至少跟随两个0。通过直接记录，同一个凹坑将只能存储一位。同时考虑每个坑中存储的比特数和第8到14位的扩展，哪一种方案存储最多的比特数？通过什么因素存储？

6.11 设计一个计算机系统的备份策略。一种选择是使用插入式外部磁盘，每个500GB大小的磁盘需要150美元。另一种选择是花2500美元买一台磁带机，400GB的磁带每台50美元。（这些价格在2008年是比较现实的。）典型的备份策略是在线设置两组备份媒体，交替地在它们上写备份，因而如果系统在进行备份时失败，则前一个版本仍然完整。还有第三套设备保持离线，离线设备定期与在线设备交换。

a. 假设您有1个TB（1000 GB）的数据要备份。磁盘备份系统要花多少钱？

b. 备份1个TB，使用磁带备份系统需要花费多少钱？

c. 为了使磁带策略更便宜，每个备份必须有多大？

d. 什么样的备份策略更青睐磁带？