第47-48 讲 磁盘容错技术

磁盘容错技术

- 保证磁盘数据的*可靠性*
- 磁盘容错技术
  - 通过在系统中设置*冗余部件*来提高系统可靠性。
  - 冗余部件包括增加冗余的磁盘驱动器、磁盘控制器等，使得当磁盘系统某部分出现缺陷或故障时，磁盘仍能正常工作，且不至于造成数据的错误和丢失。

磁盘容错技术级别

也称为系统容错技术 (SFT, System Fault Tolerance)，大体分为三个级别：

1. SFT-I 低级磁盘容错技术，主要防止磁盘表面介质缺陷所引起的数据丢失；
2. SFT-II 中级磁盘容错技术，主要防止磁盘驱动器和磁盘控制器故障所引起的数据丢失；
3. SFT-III 高级系统容错技术，常使用双服务器，以保证在其中一台服务器出现故障，甚至停止工作时，整个系统仍能照常运作。

第一级容错技术

最早出现、最基本的容错技术，包括：

1. 双份目录和双份文件分配表

2. 热修复重定向和写后读校验

双份目录和双份文件分配表

- 可在不同的磁盘上或同一磁盘的不同区域中，分别建立维护两份文件目录和 FAT 。
- 当其中一个目录或 FAT 损坏时，系统便自动启用另一个目录和 FAT ，同时在磁盘的其它区域再建立新的文件目录和 FAT 。
- 每当系统重新启动时，都要对这两份目录和 FAT 进行检查，以保证它们的一致性。

热修复重定向和写后读校验

热修复重定向

- 系统将一定的磁盘容量作为热修复重定向区，用于存放当发现磁盘块有缺陷时的待写数据，并对写入该区的所有数据进行登记，以便于以后对此数据进行访问。

写后读校验

- 每次将数据写到磁盘以后，立即从磁盘上读出该块数据，并进行对比。若写入的数据与读出的数据一致，则表示写入成功；否则，重写数据。若重写后两者仍不一致，则认为该磁盘块有缺陷，便将该块标识为坏块，相应数据写入热修复重定向区中。

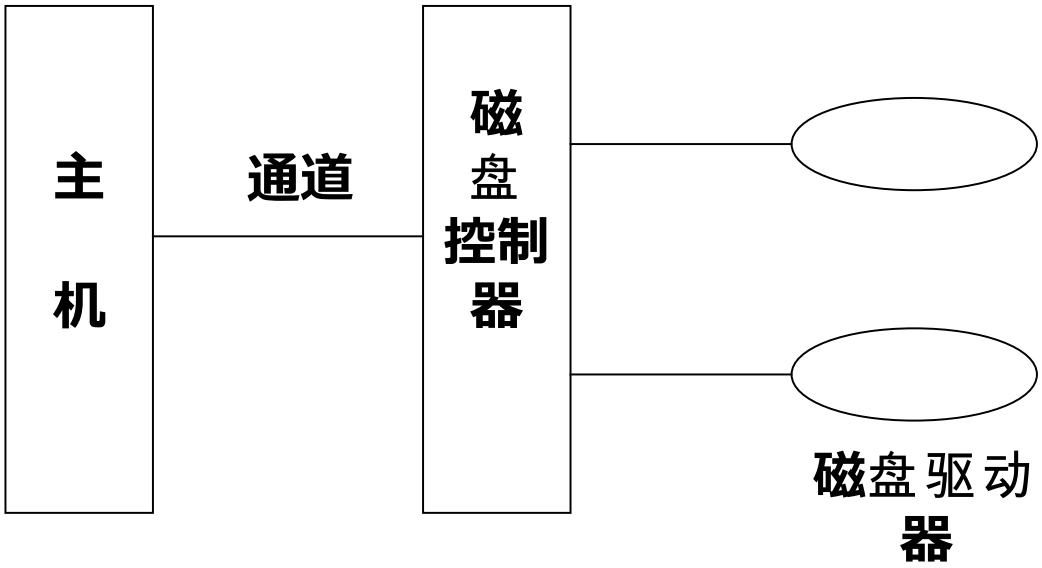
第二级容错技术 SFT-II

防止磁盘驱动器或磁盘控制器发生故障。包括：

- 磁盘镜像
- 磁盘双工

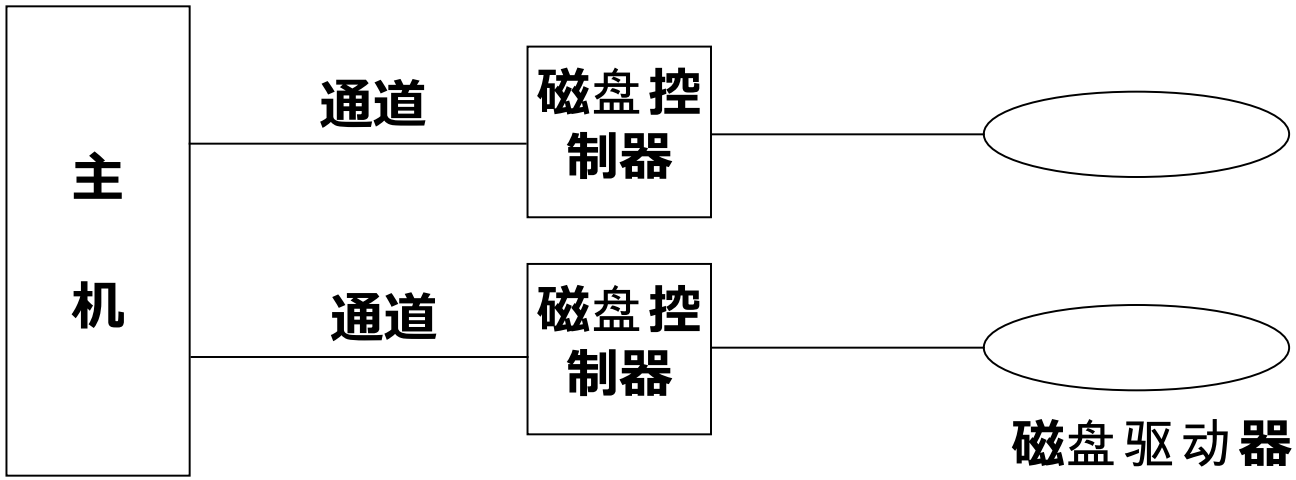
磁盘镜像

在同一磁盘控制器上，连接两个完全相同的磁盘驱动器。同一数据被先后写到两个驱动器上。



磁盘双工

将两台完全相同的磁盘驱动器连接到两个磁盘控制器上。数据被同时写到两个磁盘上。

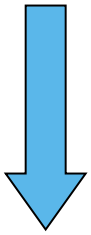


磁盘双工 *vs.* 磁盘镜像

- 磁盘双工的成本较高；
- 可靠性更高；
- 对两个磁盘的写入是*并行*进行，速度较快；
- 在某些实现中读数据时，还可使用分离查找技术，从响应快的通道上取得数据，加快读取速度。

RAID

Redundant Array of Inexpensive Disks



Redundant Array of Independent Disks

RAID 技术

- RAID 磁盘阵列技术能以较低的成本，提供大容量、快速、安全可靠的磁盘存储系统。

- 从容错的角度讲， RAID 技术应属第二级容错技术，但其内涵远远不止容错。

RAID 的基本特征

- RAID 由两部分构成：磁盘阵列（一组可并行工作的磁盘），及磁盘阵列管理软件
- 磁盘阵列管理软件把逻辑上连续的一组数据交叉分布存储在磁盘阵列中的各个磁盘上。好处：磁盘阵列管理软件可以*并行处理*对一组数据中的单个或多个数据的存取请求。

RAID 的基本特征 （续）

- 磁盘阵列管理软件还负责存储相关的校验信息。好处：当磁盘阵列中的某个磁盘发生故障时，磁盘阵列管理软件可以*恢复*存储在该磁盘上的数据。
- 磁盘阵列管理软件屏蔽了磁盘阵列的物理细节，使 OS 的其它成份不知道磁盘阵列的存在；在它们看来，系统中存在一个大容量的逻辑磁盘。

RAID 中的数据存储布局

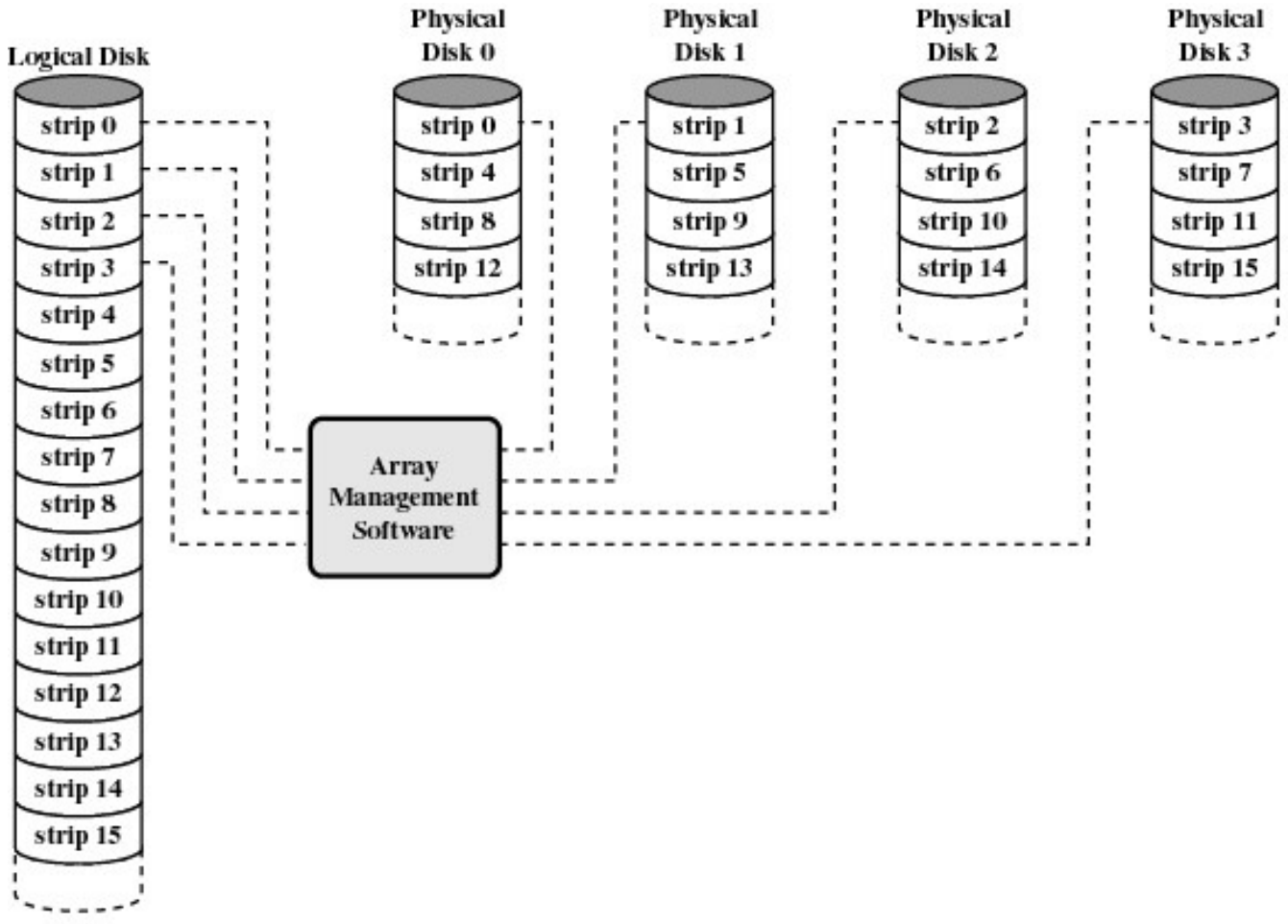


Figure 11.10 Data Mapping for a RAID Level 0 Array [MASS97]

条带 Strip 的粒度

- 条带可以是细粒度的（如一个字节或字）， 也可以是粗粒度的（如一个扇区或多个扇区）。

- 若采用细粒度条带，几乎每个存取请求都会导致同时存取 RAID 中的所有磁盘，使得无法同时响应多个存取请求。 因此，细粒度的条带只利于对单个存取请求进行并行处理

- 若采用粗粒度条带，不会使每个存取请求都同时存取 RAID 中的所有磁盘；但，多个独立的存取请求通常会导致同时存取 RAID 中的所有磁盘。 显然，粗粒度的条带只利于对多个独立的存取请求进行并行处理

磁盘阵列管理软件

- 磁盘阵列管理软件在存储数据的同时还将存储相关的校验信息；

- 使得当磁盘阵列中的某个磁盘发生故障时，磁盘阵列管理软件可以恢复存储在该磁盘上的数据。

RAID 中的数据校验方式

- 对数据进行镜像存储
- 对数据进行 Hamming 编码
- 存储数据的奇偶校验信息

RAID 中奇偶校验信息的存储布局

- 用一组专用磁盘存储奇偶校验信息
- 把奇偶校验信息分布存储在各个磁盘上

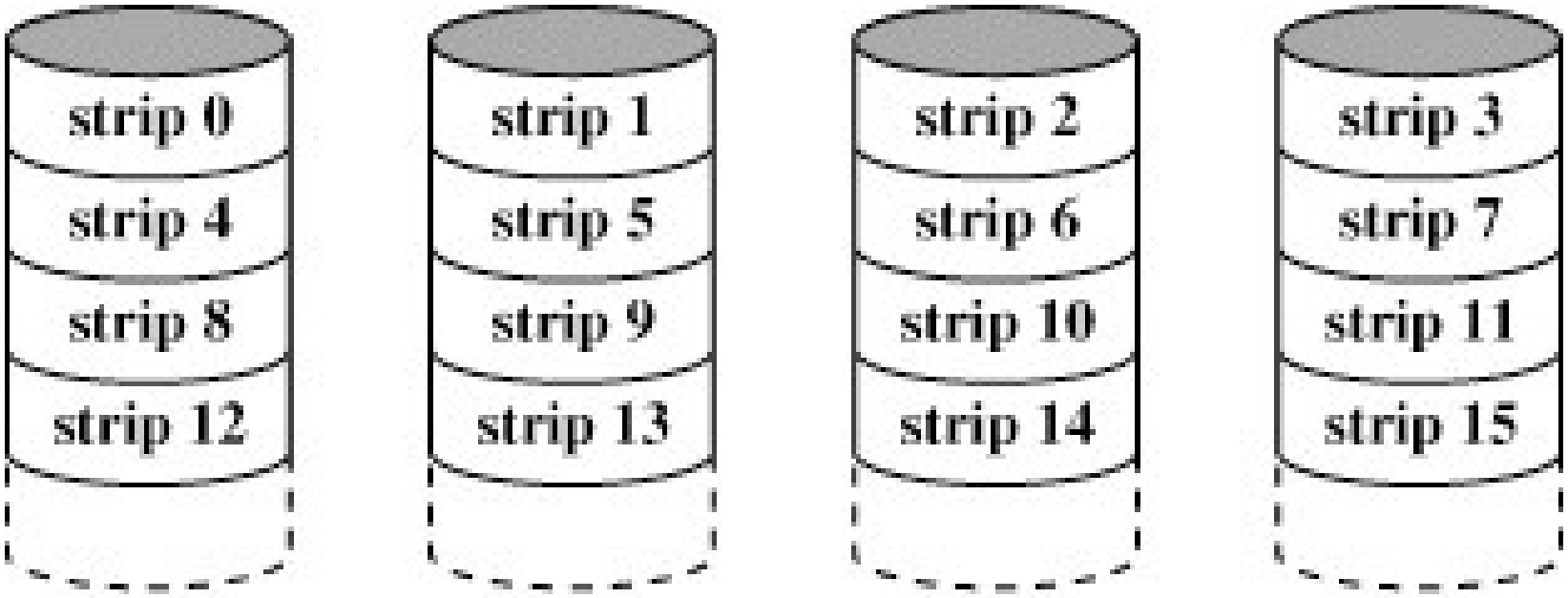
- 如果磁盘阵列管理软件在主机系统中执行，则属*软件*实现的 RAID；如果磁盘阵列管理软件在磁盘子系统中执行，则属*硬件*实现的 RAID，相应的硬件称为 RAID 控制器或 RAID 卡。

- 采用 RAID 控制器方案的成本较高，但其效率高，不增加主机负担。此类 RAID 技术对主机完全透明，在主机看来， RAID 就是一个容量大、速度快、可靠性高的磁盘。

RAID 的常见组织形式（6 种）

- RAID Level 0
- RAID Level 1
- RAID Level 2
- RAID Level 3
- RAID Level 4
- RAID Level 5
- 还可对基本 RAID 级别进行组合

RAID 0 (non-redundant)



(a) RAID 0 (non-redundant)

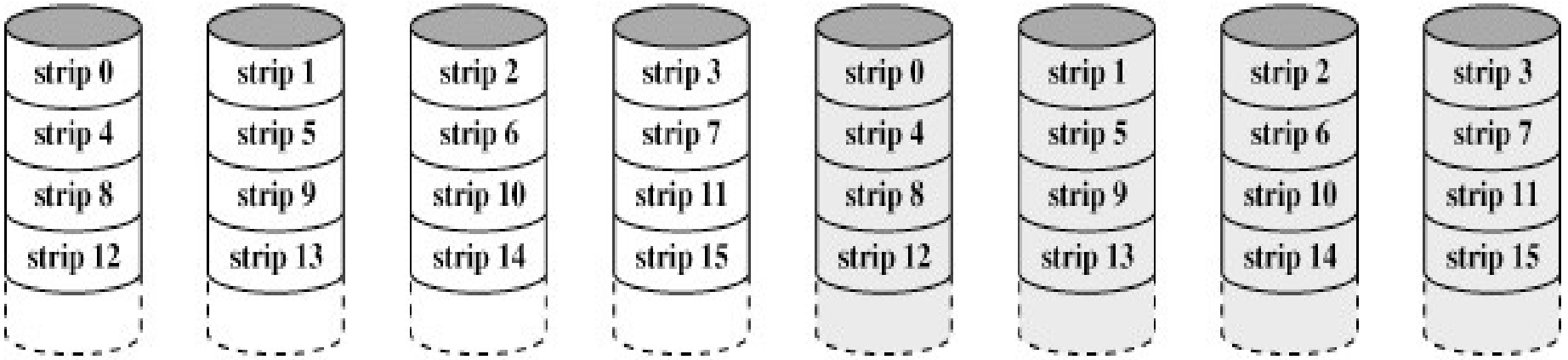
Figure 11.9 RAID Levels (page 1 of 2)

RAID0

- 仅使用了条带化技术
- 不存储数据的校验信息
- 能提供大容量、快速的磁盘存储能力，
- 具备最好的读／写性能和最低的成本
- 磁盘容量的利用率为 100%

- 但其安全性最低，其中任何一个磁盘损坏便会导致整个系统不可使用。

RAID 1 (mirrored)



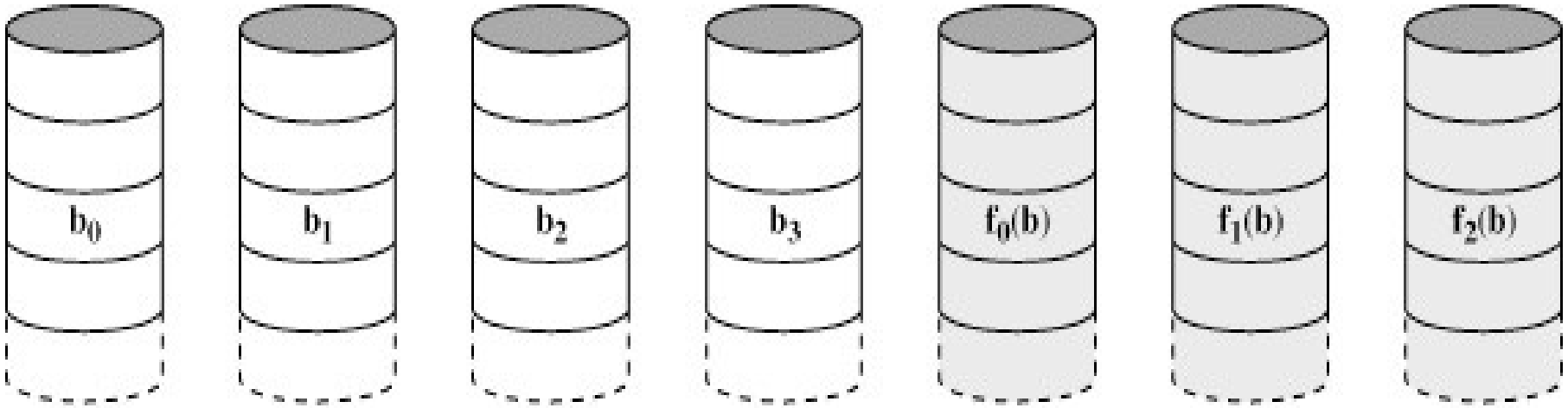
(b) RAID 1 (mirrored)

Figure 11.9 RAID Levels (page 1 of 2)

RAID1

- 仅使用了磁盘镜像或磁盘双工技术；
- 能提供最好的安全性，其中任何一个磁盘损坏都不会导致数据丢失；
- 但磁盘容量的利用率只有 50%，相对于其它 RAID 组织形式，成本较高；
- 相对于 RAID0， RAID1 读性能较好 .

RAID 2 (redundancy through Hamming code)



(c) RAID 2 (redundancy through Hamming code)

Figure 11.9 RAID Levels (page 1 of 2)

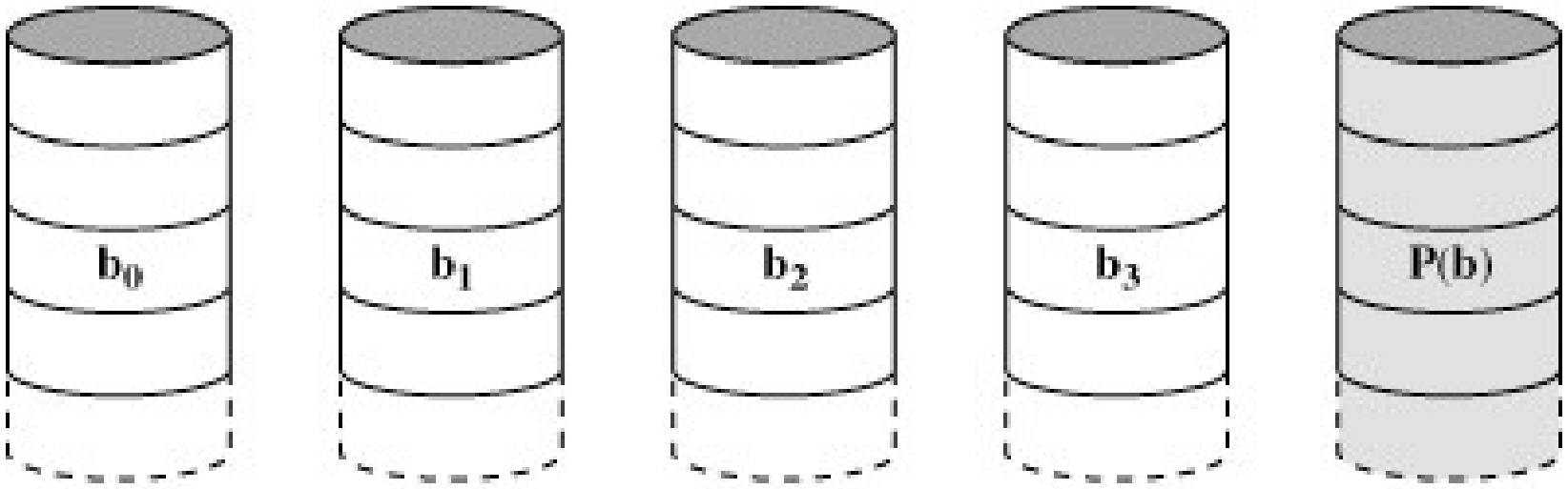
RAID Level 2 的特征

- 使用细粒度 Strip
- 对数据进行 Hamming 编码：能纠正 1 位错误，检测 2 位错误
- Hamming 编码信息存放在专用的磁盘上

对 RAID Level 2 的评价

- 相对于 RAID1， RAID2 成本较低； 但相对于 RAID Level 3 、 4 、 5， RAID2 成本较高；
- 相对于 RAID Level 4 、 5， RAID2 具有较好的数据传输性能，但具有*较差的 I/O 请求响应能力*

RAID 3 (bit-interleaved parity)



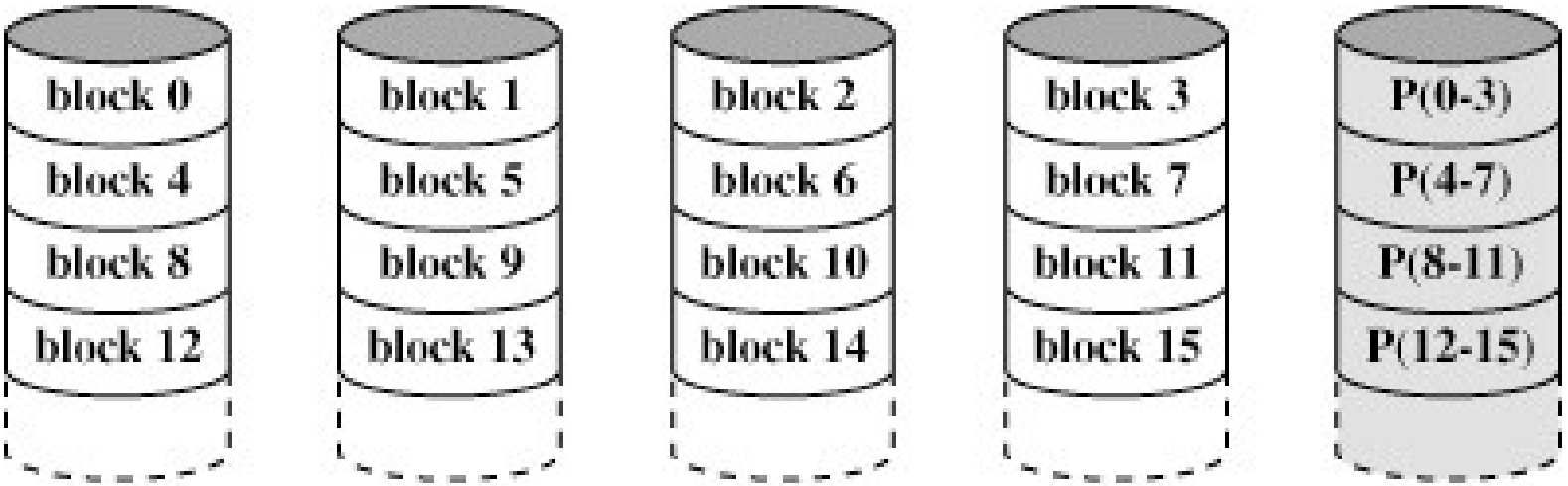
(d) RAID 3 (bit-interleaved parity)

Figure 11.9 RAID Levels (page 2 of 2)

RAID Level 3 的特征

- 使用细粒度 Strip
- 存储数据的奇偶校验信息
- 奇偶校验信息存放在一个专用的磁盘上
- 相对于 RAID Level 2， RAID3 *成本较低*

RAID 4 (block-level parity)



(e) RAID 4 (block-level parity)

Figure 11.9 RAID Levels (page 2 of 2)

RAID Level 4 的特征

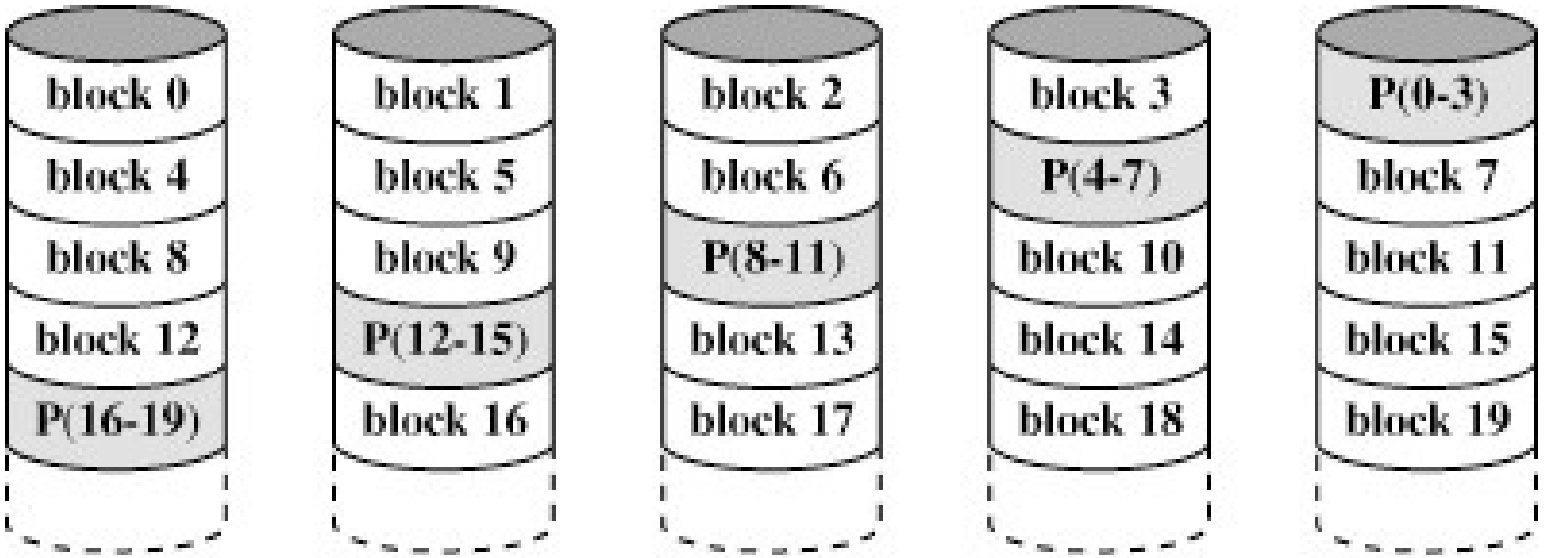
- 使用粗粒度 Strip
- 存储数据的奇偶校验信息
- 奇偶校验信息存放在专用的磁盘上

对 RAID Level 4 的评价

- 当用户进程请求把一个尺寸较小的数据写入 RAID 中时，需要读两次磁盘、写两次磁盘 *(Write Penalty)*

- 任何情况下，一个写请求都会导致磁盘阵列管理软件访问奇偶校验盘；因此，多个独立的写请求很难真正并行处理。 *(Parity disk become a bottleneck)*

RAID 5 (block-level distributed parity)



(f) RAID 5 (block-level distributed parity)

Figure 11.9 RAID Levels (page 2 of 2)

RAID Level 5 的特征

- 使用粗粒度 Strip
- 存储数据的奇偶校验信息
- 奇偶校验信息分布存储在各个磁盘上
- 多个独立的写请求可以真正地被并行处理。
  - *(Avoid bottleneck found in RAID Level 4)*

RAID 的优点

- 效率高
  - 系统可以并行存取存储在 RAID 中的数据。
- 可靠性高
  - 系统可以恢复存储在故障磁盘中的数据。
- 性价比高
  - 磁盘阵列可以用廉价磁盘组成。

Thank You !

UESTC

电子科技大学
University of Electronic Science and Technology of China