计算机组织结构

11 冗余磁盘阵列（RAID）

刘博涵

2023年11月14日

南京大学
NANJING UNIVERSITY

教材对应章节

第8章 互连及输入输出组织

第6章 外部存储器

RAID

- 冗余磁盘阵列 / 独立磁盘冗余阵列：Redundant Arrays of Independent Disks (RAID)
- 基本思想
  - 将多个独立操作的磁盘按某种方式组织成磁盘阵列，以增加容量
  - 将数据存储在多个盘体上，通过这些盘并行工作来提高数据传输率
  - 采用数据冗余来进行错误恢复以提高系统可靠性
- 特性
  - 由一组物理磁盘驱动器组成，被视为单个逻辑驱动器
  - 数据是分布在多个物理磁盘上
  - 冗余磁盘容量用于存储校验信息，保证磁盘万一损坏时能恢复数据

Patterson, David & Gibson, Garth & Katz, Randy. (1988). A case for Redundant Arrays of Inexpensive Disks (RAID). ACM SIGMOD Record. 17. 10.1145/50202.50214.

RAID分类

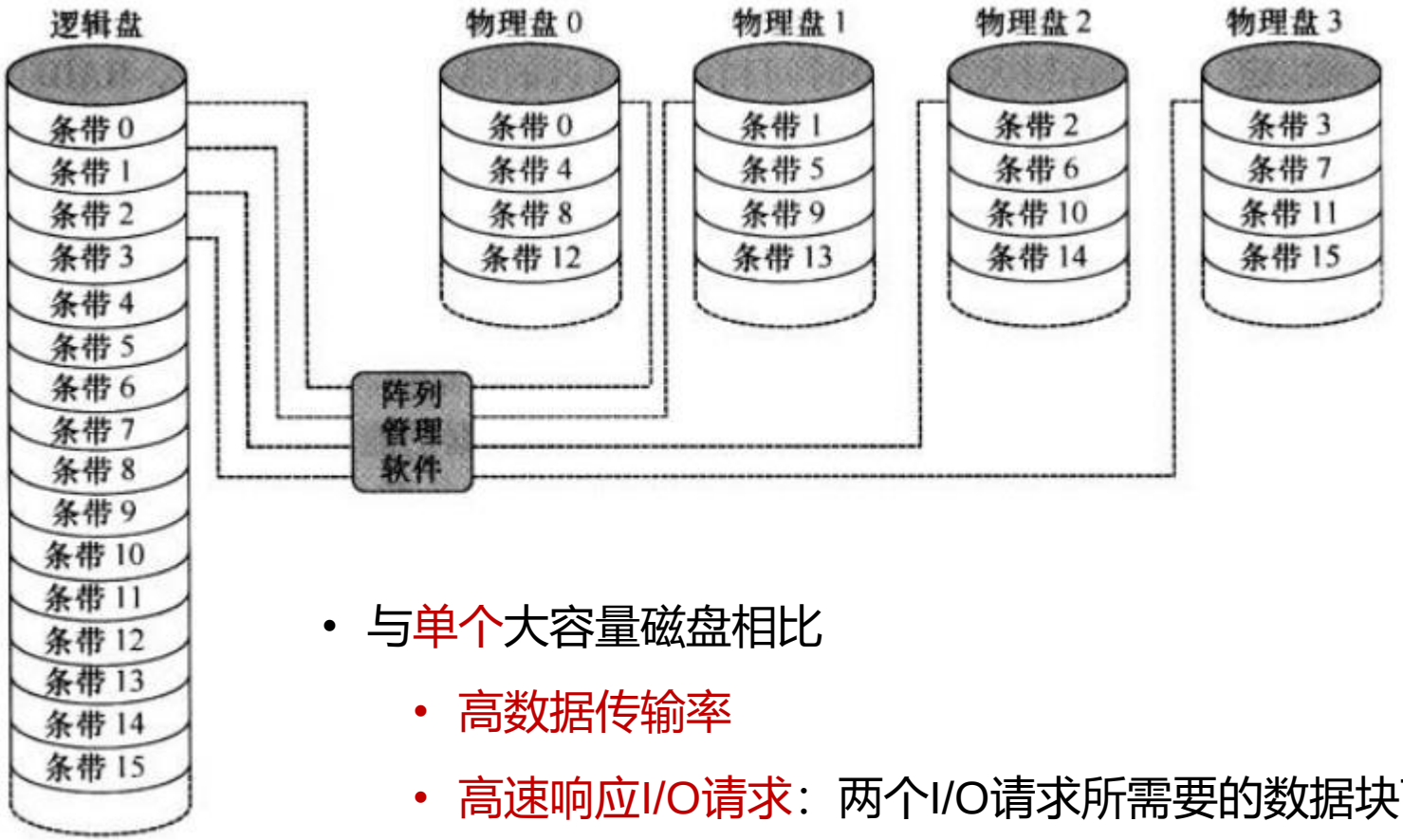
种类	级别	描述	磁盘要求	数据可用性	大I/O数据传输能力	小I/O请求速率
条带化	0	非冗余	N	比单盘低	很高	读和写都很高
镜像	1	镜像	$2N$	比RAID 2、3、4、5高；比RAID 6低	读比单盘高；写与单盘类似	读高达单盘的两倍；写与单盘类似
并行存取	2	汉明码冗余	$N+m$	比单盘高很多，与RAID 3、4、5差不多	列表各级中最高	接近于单盘的两倍
	3	位交错奇偶校验	$N+1$	比单盘高很多；与RAID 2、4、5差不多	列表各级中最高	接近于单盘的两倍
独立存取	4	块交错奇偶校验	$N+1$	比单盘高很多；与RAID 2、3、5差不多	读与RAID 0类似；写低于单盘	读与RAID 0类似；写显著低于单盘
	5	块交错分布式奇偶校验	$N+1$	比单盘高很多；与RAID 2、3、4差不多	读与RAID 0类似；写低于单盘	读与RAID 0类似；写显著低于单盘
	6	块交错分布式奇偶校验	$N+2$	列表各级中最高	读与RAID 0类似；写比RAID 5低	读与RAID 0类似；写显著低于RAID 5

RAID 0

- 数据以条带的形式在可用的磁盘上分布
- 不采用冗余来改善性能（不是RAID家族中的真正成员）
- 用途
  - 高数据传输率
  - 高速响应I/O请求

自定义
16~512KB

RAID 0（续）

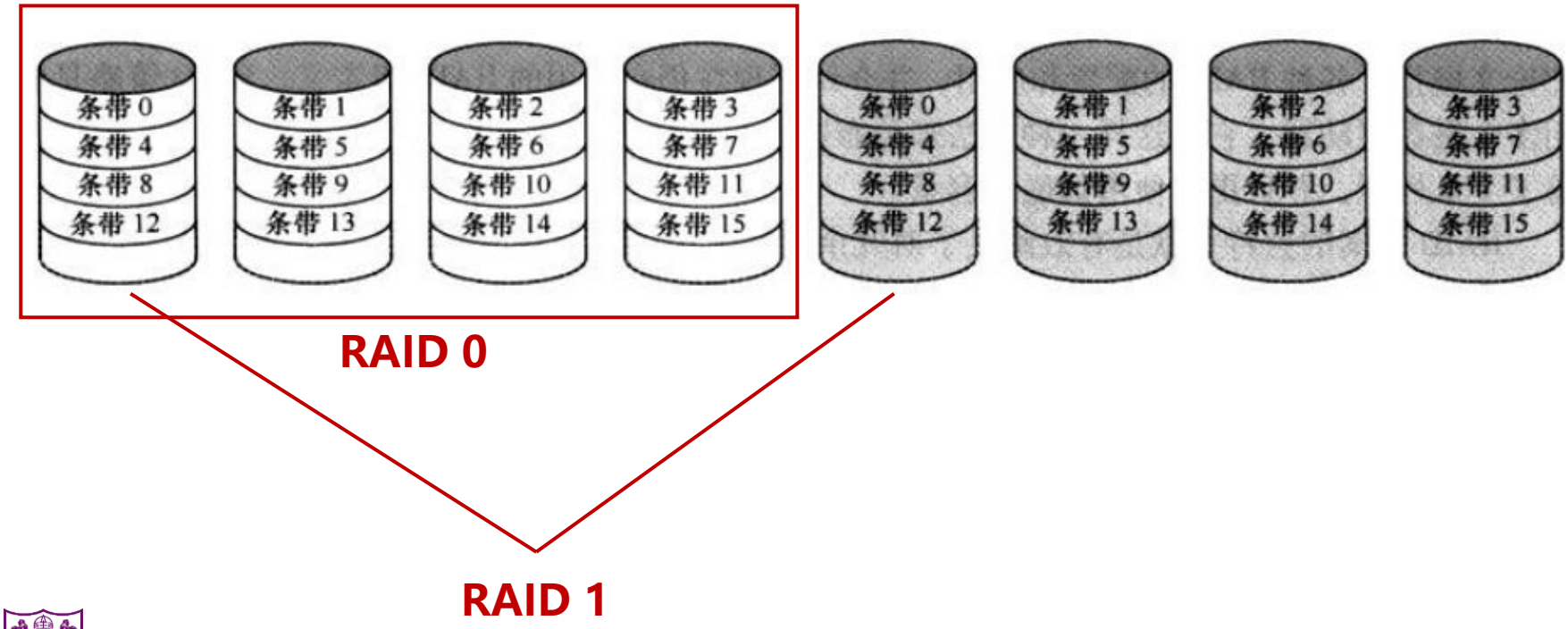


- 与单个大容量磁盘相比
  - 高数据传输率
  - 高速响应I/O请求：两个I/O请求所需要的数据块可能在不同的磁盘上

条带化	0	非冗余	*N*	比单盘低	很高	读和写都很高

RAID 1

- 采用了数据条带
- 采用简单地备份所有数据的方法来实现冗余



RAID 1（续）

- **优点**
  - 高速响应I/O请求：即便是同一个磁盘上的数据块，也可以由两组硬盘分别响应
  - 读请求可以由包含请求数据的两个对应磁盘中的某一个提供服务，可以选择寻道时间较小的那个
  - 写请求需要更新两个对应的条带：可以并行完成，但受限于写入较慢的磁盘
  - 单个磁盘损坏时不会影响数据访问，恢复受损磁盘简单
- **缺点**
  - 价格昂贵（一半的容量）

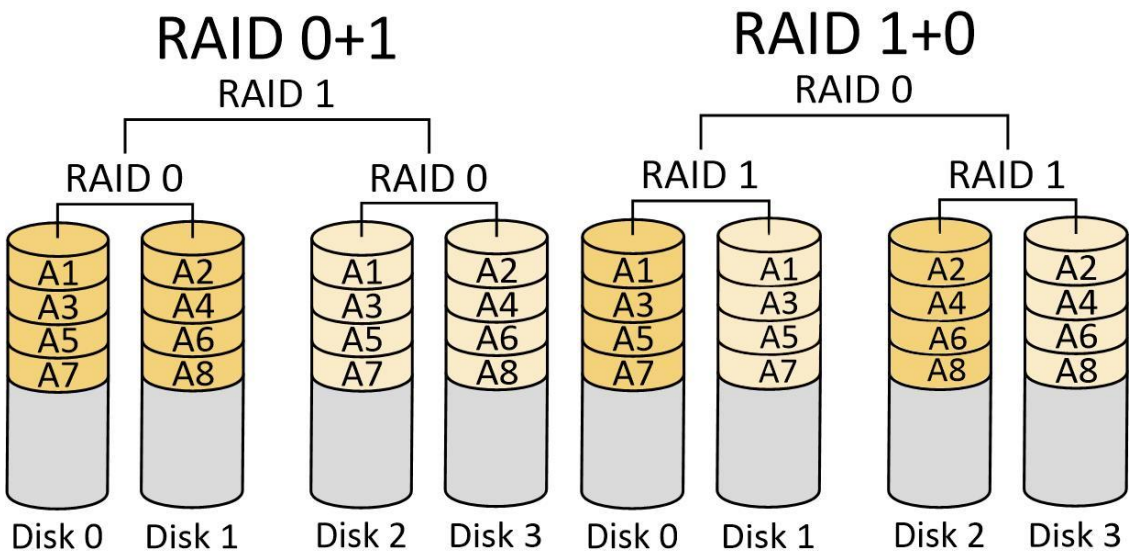
镜像	1	镜像	2*N*	比 RAID 2、3、4、5 高；比 RAID 6 低	读比单盘高；写与单盘类似	读高达单盘的两倍；写与单盘类似

RAID 1（续）

- **用途**
  - 只限于用在存储系统软件、数据和其他关键文件的驱动器中
- **与 RAID 0 相比**
  - 如果有大批的读请求，则RAID 1能实现高速的I/O速率，性能可以达到RAID 0的两倍
  - 如果I/O请求有相当大的部分是写请求，则它不比RAID 0的性能好多少

RAID 01 vs. RAID 10

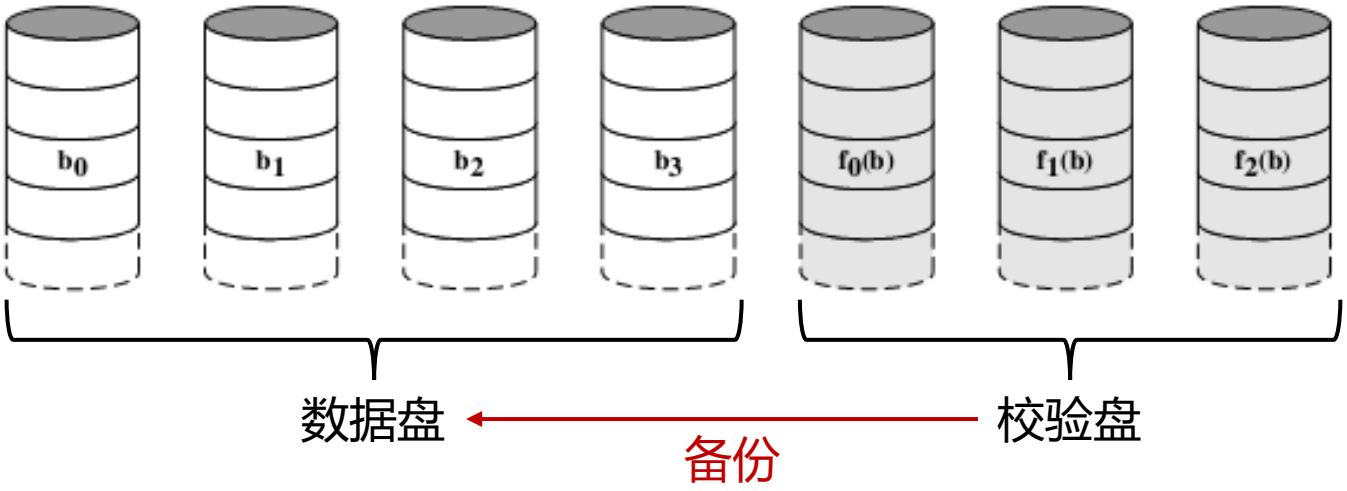
- RAID 01 = RAID 0+1：先做RAID 0，再做RAID 1
- RAID 10 = RAID 1+0：先做RAID 1，再做RAID 0
- 两者在数据传输率和磁盘利用率上没有明显区别，主要区别是对磁盘损坏的**容错能力**



思考：如果Disk 0和Disk 3坏了会怎么样?

RAID 2

- **采用并行存取技术**
- **目标**
  - 所有磁盘都参与每个I/O请求的执行
- **特点**
  - 各个驱动器的轴是同步旋转的，因此每个磁盘上的每个磁头在任何时刻都位于同一位置
  - 采用数据条带：条带非常小，经常只有一个字节或一个字



RAID 2（续）

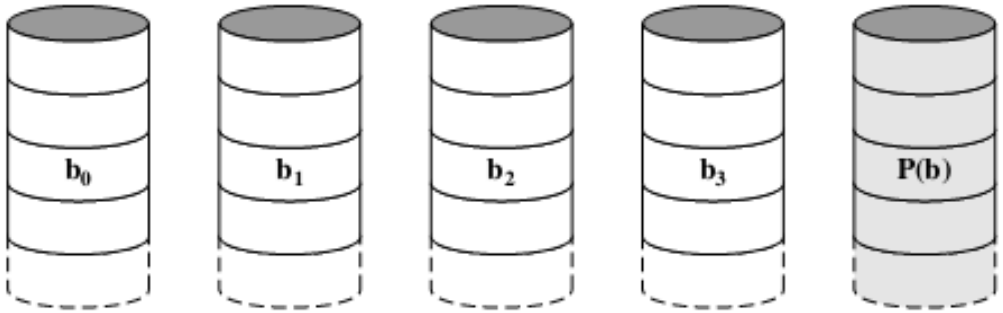
- **纠错**
  - 对位于同一条带的各个数据盘上的数据位计算校验码（通常采用海明码），校验码存储在该条带中多个校验盘的对应位置
- **访问**
  - 读取：获取请求的数据和对应的校验码
  - 写入：所有数据盘和校验盘都被访问
- **缺点**
  - 冗余盘依然比较多，价格较贵
  - 适用于多磁盘易出错环境，对于单个磁盘和磁盘驱动器已经具备高可靠性的情况没有意义（实际基本弃用）

2	汉明码冗余	$N+m$	比单盘高很多，与RAID 3、4、5差不多	列表各级中最高	接近于单盘的两倍

RAID 3

- **采用并行存取技术**
  - 各个驱动器的轴同步旋转
  - 采用非常小的数据条带
- **校验**：对所有数据盘上同一位置的数据计算奇偶校验码
  - 当某一磁盘损坏时，可以用于重构数据　　回顾：奇偶校验码距是多少？

$$b_0 = P(b) \oplus b_1 \oplus b_2 \oplus b_3$$



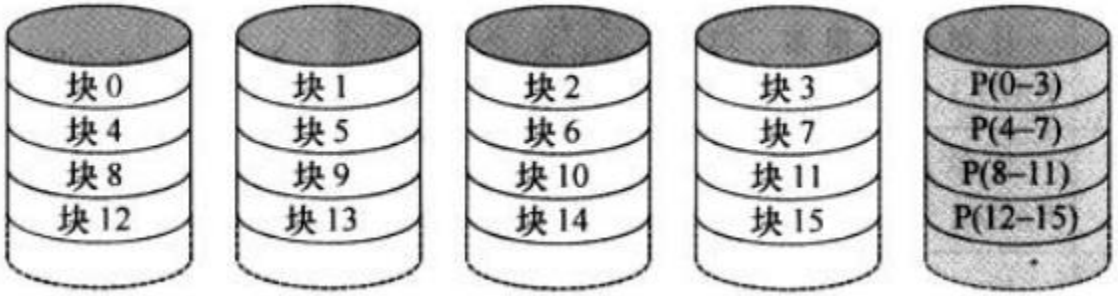
RAID 3 （续）

- **优点**
  - 能够获得非常高的数据传输率，对于大量读请求，性能改善特别明显
- **缺点**
  - 一次只能执行一个I/O请求，在面向多个IO请求时，性能将受损

3	位交错奇偶校验	$N+1$	比单盘高很多；与RAID 2、4、5 差不多	列表各级中最高	接近于单盘的两倍

RAID 4

- 采用独立存取技术
  - 每个磁盘成员的操作是独立的，各个I/O请求能够并行处理
- 采用相对较大的数据条带（常见的是4KB）
- 根据各个数据盘上的数据来逐位计算奇偶校验条带，奇偶校验位存储在奇偶校验盘的对应条带上



RAID 4（续）

- **性能**
  - 当执行较小规模的I/O写请求时，RAID 4会遭遇写损失
    - 对于每一次写操作，阵列管理软件不仅要修改用户数据，而且要修改相应的校验位

$$P'(B) = P(B) \oplus B_0 \oplus B_0'$$

  - 当涉及所有磁盘的数据条带的较大I/O写操作时，只要用新的数据位来进行简单的计算即可得到奇偶校验位
  - 每一次写操作必须涉及到唯一的校验盘，校验盘会成为瓶颈（实际基本弃用）

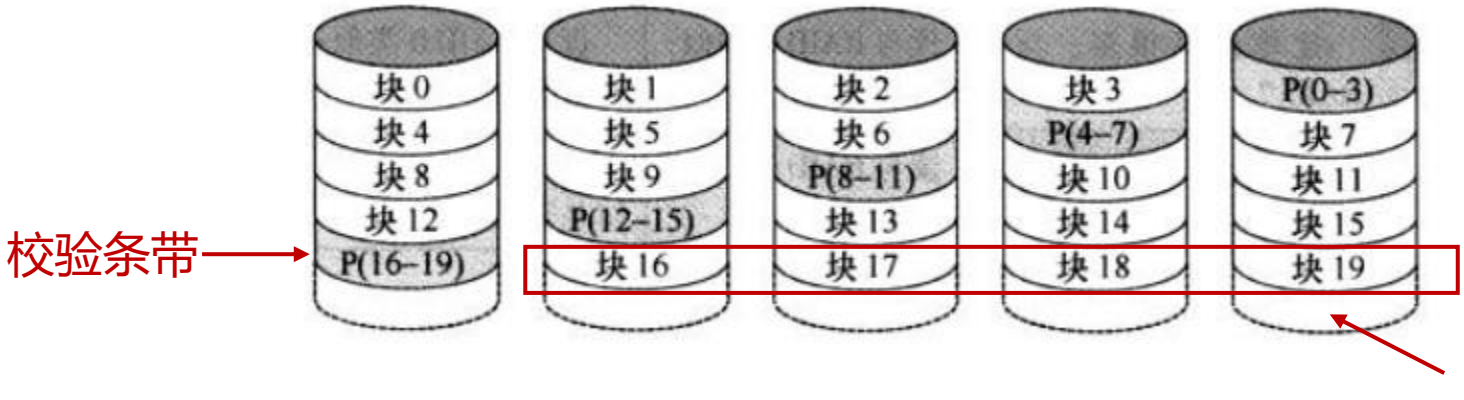
4	块交错奇偶校验	$N+1$	比单盘高很多；与RAID 2、3、5差不多	读与RAID 0类似；写低于单盘	读与RAID 0类似；写显著低于单盘

RAID 5

- 与RAID 4 组织方式相似（常用）
- 在所有磁盘上都分布了奇偶校验条带
  - 避免潜在的I/O瓶颈问题
- 访问时的“两读两写”：读在写前，读/写不需要并行

$$P'(B) = P(B) \oplus B_0 \oplus B_0'$$

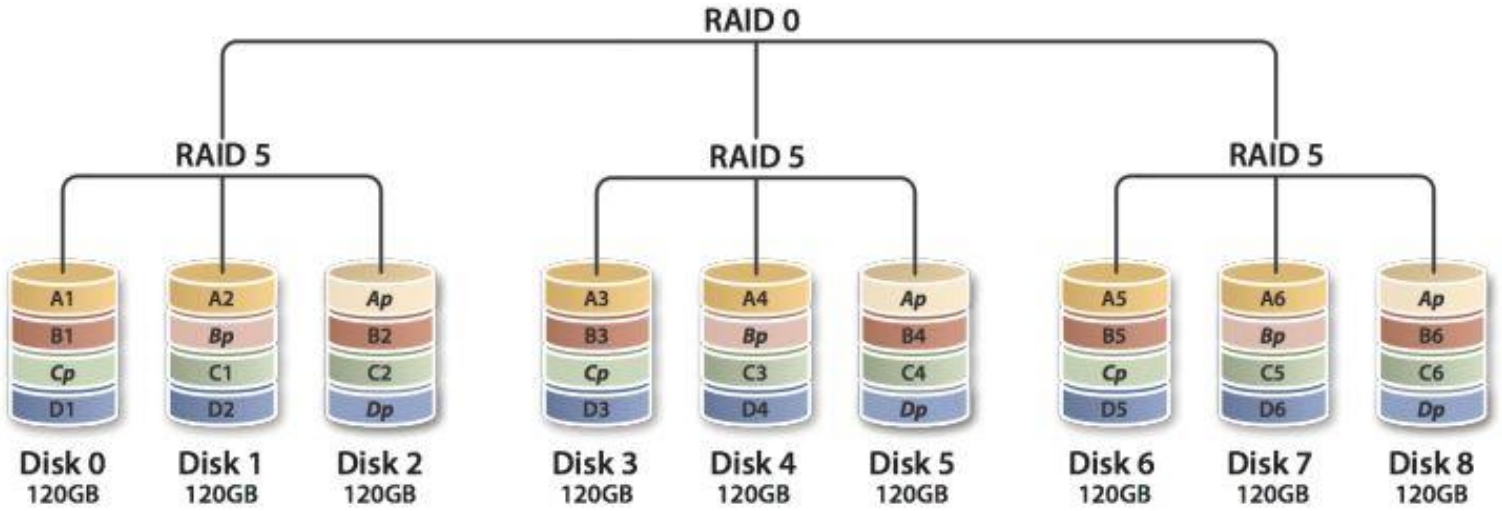
读　写



5	块交错分布式奇偶校验	$N+1$	比单盘高很多；与RAID 2、3、4 差不多	读与 RAID 0 类似；写低于单盘	读与 RAID 0 类似；写显著低于单盘

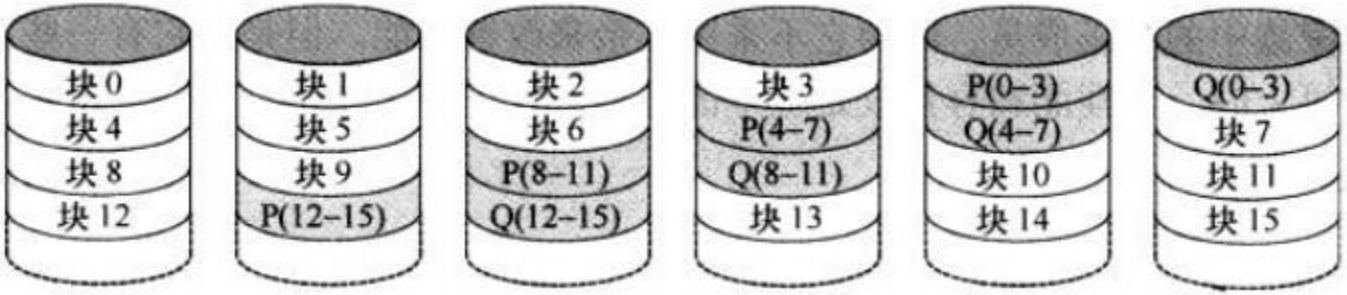
RAID 50

- RAID 5与RAID 0的组合，先作RAID 5，再作RAID 0，也就是对多组RAID 5彼此构成条带访问
- RAID 50在底层的任一组或多组RAID 5中出现1颗硬盘损坏时，仍能维持运作；如果任一组RAID 5中出现2颗或2颗以上硬盘损毁，整组RAID 50就会失效
- RAID 50由于在上层把多组RAID 5进行条带化，性能比起单纯的RAID 5高，但容量利用率比RAID5要低



RAID 6

- 采用两种不同的校验码，并将校验码以分开的块存于不同的磁盘中
- **优点**
  - 提升数据可用性：只有在平均修复时间间隔内3个磁盘都出了故障，才会造成数据丢失
- **缺点**
  - 写损失：每次写都要影响两个校验块（读3个写3个磁盘）



6	块交错分布式奇偶校验	$N+2$	列表各级中最高	读与 RAID 0 类似；写比 RAID 5 低	读与 RAID 0 类似；写显著低于 RAID 5

RAID 比较

级	优　点	缺　点	应　用
0	通过将 I/O 负载分散到多个通道和驱动器，极大地改善了 I/O 性能 无奇偶计算开销 很简单的设计 易实现	只要有某一个驱动器失效就导致阵列全部数据丢失	视频制作和编辑 图像编辑 预压缩应用 任何要求高带宽的应用
1	数据 100% 的冗余，意味着磁盘失效时无需重构，只需对替代盘拷贝即可 某些环境下，RAID 1 能承受多个驱动器同时失效 最简单的 RAID 存储子系统设计	在所有 RAID 类型中，磁盘数开销最大（100%）——低效	统计、工资单、财务和任何要求很高可用性的应用

RAID 0：提升I/O响应能力，但数据可用性低
RAID 1：提升数据可用性，但容量利用率低

RAID 比较（续）

2	可能有极高的数据传输率 数据传输率要求得越高，数据盘对 ECC 盘的比值越好 与 RAID3、4 和 5 级相比，控制器设计相对简单	短字长时，ECC 盘对数据盘的比值非常高——低效 入门级成本很高——要求证实很高数据传输率的需求是恰当的	无商品实现的存在/无商业化应用
3	很高的读数据传输率 很高的写数据传输率 磁盘失效时对吞吐率无显著影响 ECC（奇偶）盘对数据盘的低比率意味着高效率	最好情况（如果主轴同步旋转）下的事务率等同于单盘的事务率 控制器设计相当复杂	视频制作和直播 图像编辑 视频编辑 预压缩应用 任何要求高吞吐率的应用

RAID 2 和 RAID3：提升数据可用性和数据传输率，但一次只能处理一个I/O请求

RAID 比较（续）

级	优　点	缺　点	应　用
4	很高的读数据事务率 ECC（奇偶）盘对数据盘的低比率意味着高效率	十分复杂的控制器设计 最差的写事务率和写聚集传输率 磁盘失效事件中，数据重构困难并低效	无商品实现的存在/无商业化应用
5	最高的读数据事务率 ECC（奇偶）盘对数据盘的低比率意味着高效率 好的聚集传输速率	最复杂的控制器设计 磁盘失效事件中，数据重构困难（与 RAID 1 级相比）	数据和应用服务器 数据库服务器 Web、E-mail 和新闻组服务器 Intranet 服务器 用途最多的 RAID 级
6	提供极高的数据故障容忍能力并能承受多个驱动器同时失效	较复杂的控制器设计 计算奇偶校验地址的控制器开销非常高	对丢失数据严重的应用是理想的解决方案

RAID 4 和 RAID 5 和 RAID 6：提升数据可用性和读速率，但写速率受限

总结

种类	级别	描述	磁盘要求	数据可用性	大 I/O 数据传输能力	小 I/O 请求速率
条带化	0	非冗余	N	比单盘低	很高	读和写都很高
镜像	1	镜像	$2N$	比 RAID 2、3、4、5 高；比 RAID 6 低	读比单盘高；写与单盘类似	读高达单盘的两倍；写与单盘类似
并行存取	2	汉明码冗余	$N+m$	比单盘高很多，与 RAID 3、4、5 差不多	列表各级中最高	接近于单盘的两倍
	3	位交错奇偶校验	$N+1$	比单盘高很多；与 RAID 2、4、5 差不多	列表各级中最高	接近于单盘的两倍
独立存取	4	块交错奇偶校验	$N+1$	比单盘高很多；与 RAID 2、3、5 差不多	读与 RAID 0 类似；写低于单盘	读与 RAID 0 类似；写显著低于单盘
	5	块交错分布式奇偶校验	$N+1$	比单盘高很多；与 RAID 2、3、4 差不多	读与 RAID 0 类似；写低于单盘	读与 RAID 0 类似；写显著低于单盘
	6	块交错分布式奇偶校验	$N+2$	列表各级中最高	读与 RAID 0 类似；写比 RAID 5 低	读与 RAID 0 类似；写显著低于 RAID 5

谢谢

bohanliu@nju.edu.cn

南京大學
NANJING UNIVERSITY