RAID技术及应用

www.huawei.com

HUAWEI

目　录

HUAWEI

目　录

HUAWEI

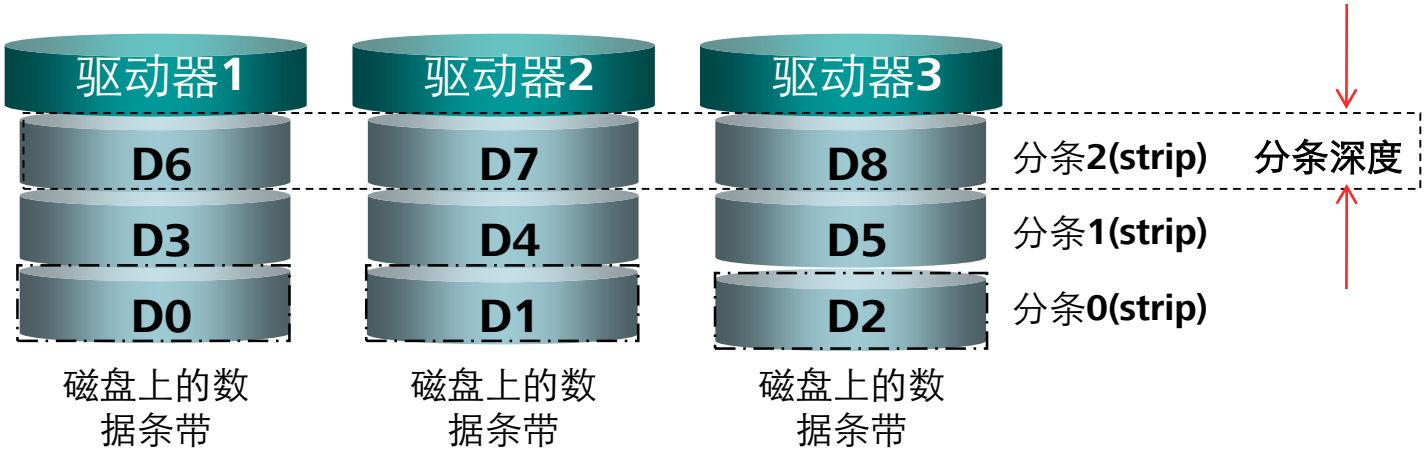
RAID基本概念与技术原理

- **RAID（Redundant Array of Independent Disks）**：独立冗余磁盘阵列，简称磁盘阵列。

- **RAID** 的主要实现方式分为硬件**RAID** 方式和软件**RAID** 方式
  - 硬件**RAID**
  - 软件**RAID**

RAID基本概念与技术原理（续）

- 条带：磁盘中单个或者多个连续的扇区构成一个条带。它是组成分条的元素。
- 分条：同一磁盘阵列中的多个磁盘驱动器上的相同“位置”（或者说是相同编号）的条带。



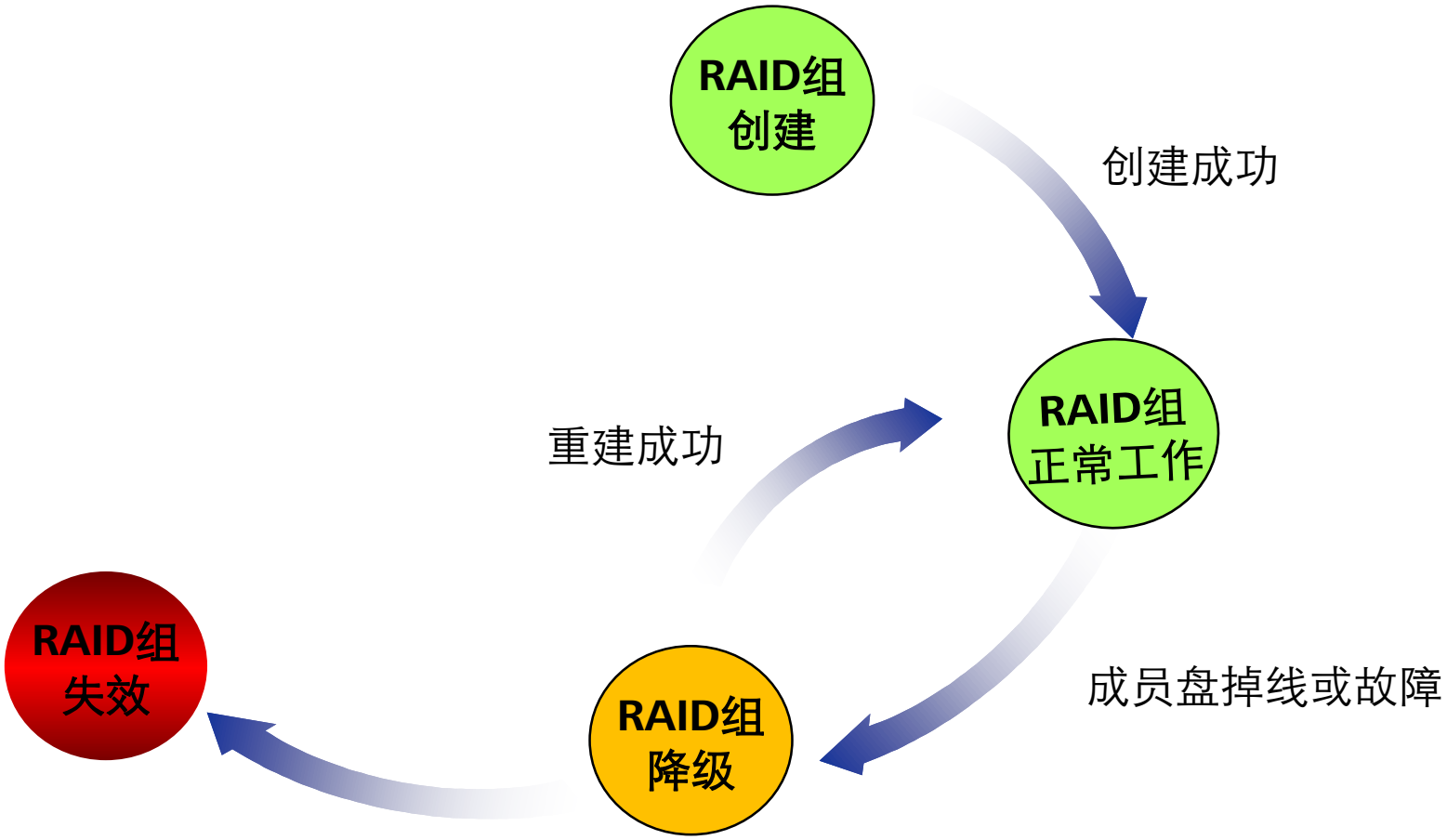
RAID基本概念与技术原理（续）

- **XOR**校验的算法 —— 相同为假，相异为真：

  0⊕0= 0； 0⊕1= 1； 1⊕0= 1； 1⊕1= 0；

异或校验冗余备份

RAID基本概念与技术原理（续）

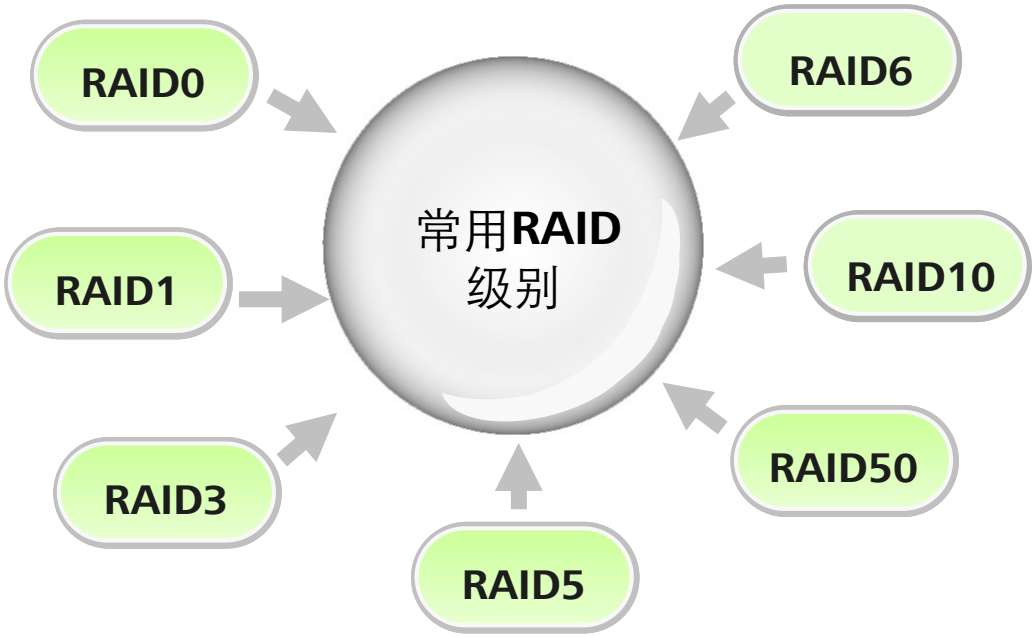


目　录

HUAWEI

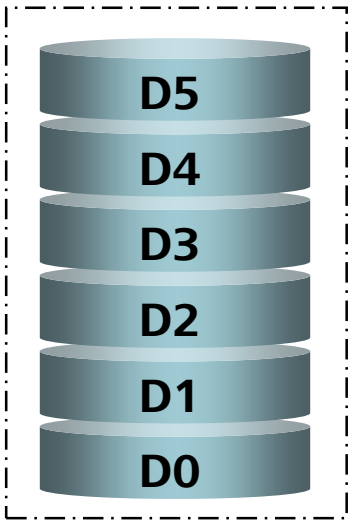
RAID技术与应用 — RAID级别

- **RAID**技术将多个单独的物理硬盘以不同的方式组合成一个逻辑硬盘，提高了硬盘的读写性能和数据安全性，根据不同的组合方式可以分为不同的**RAID**级别。

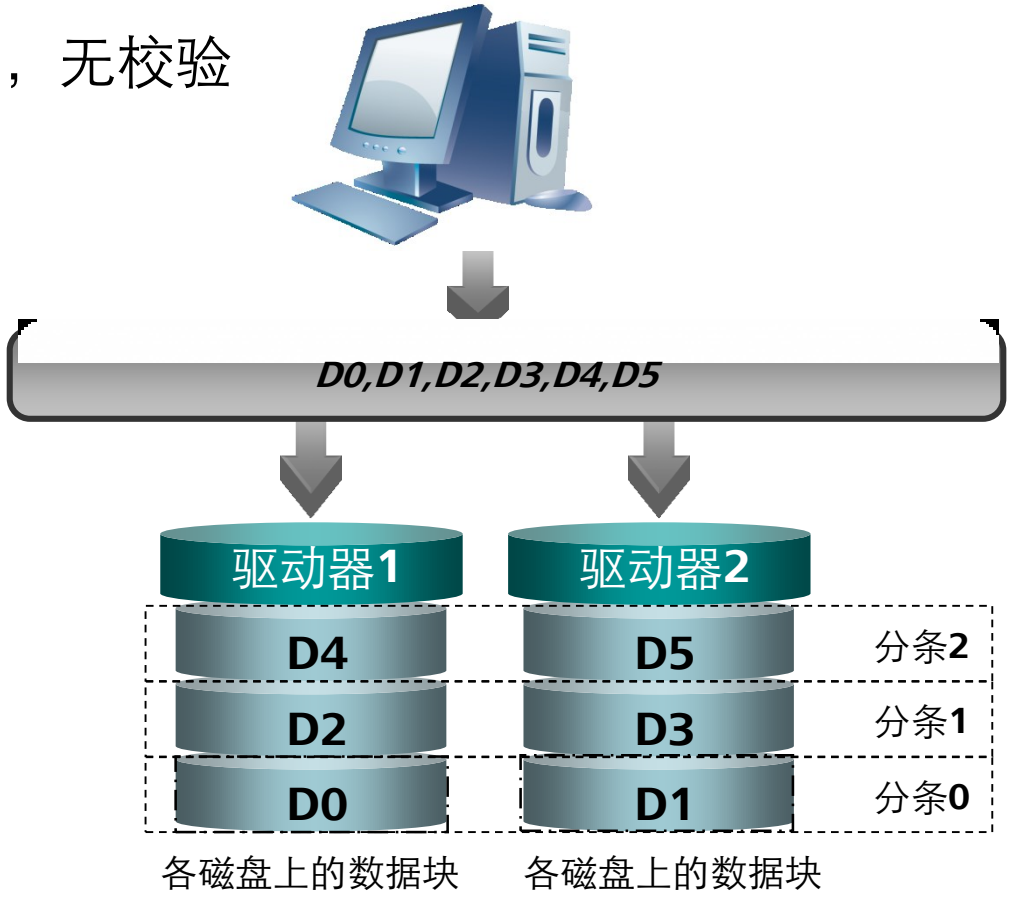


RAID技术与应用 — RAID 0

- **RAID 0**：数据条带化，无校验

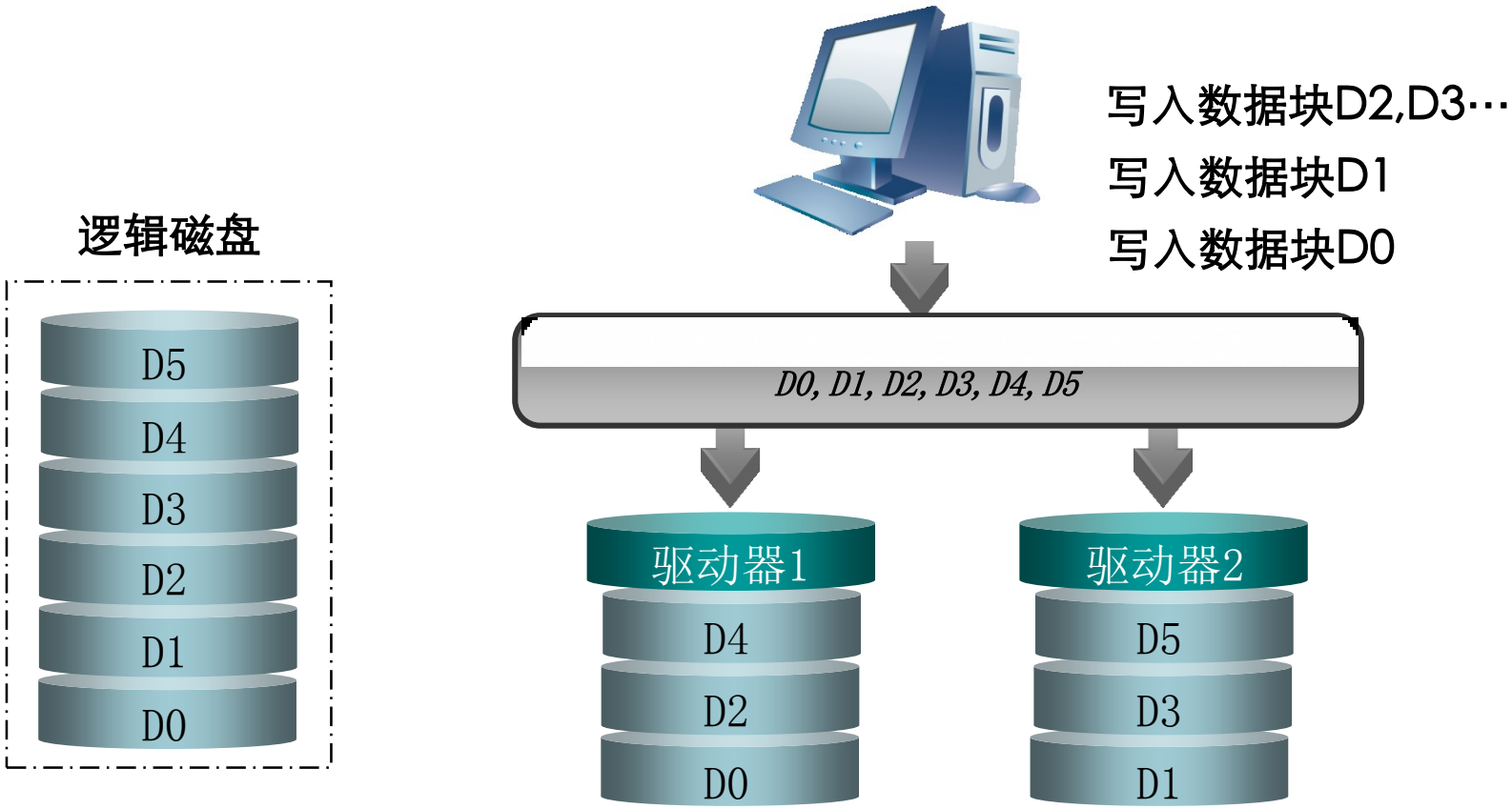


逻辑磁盘

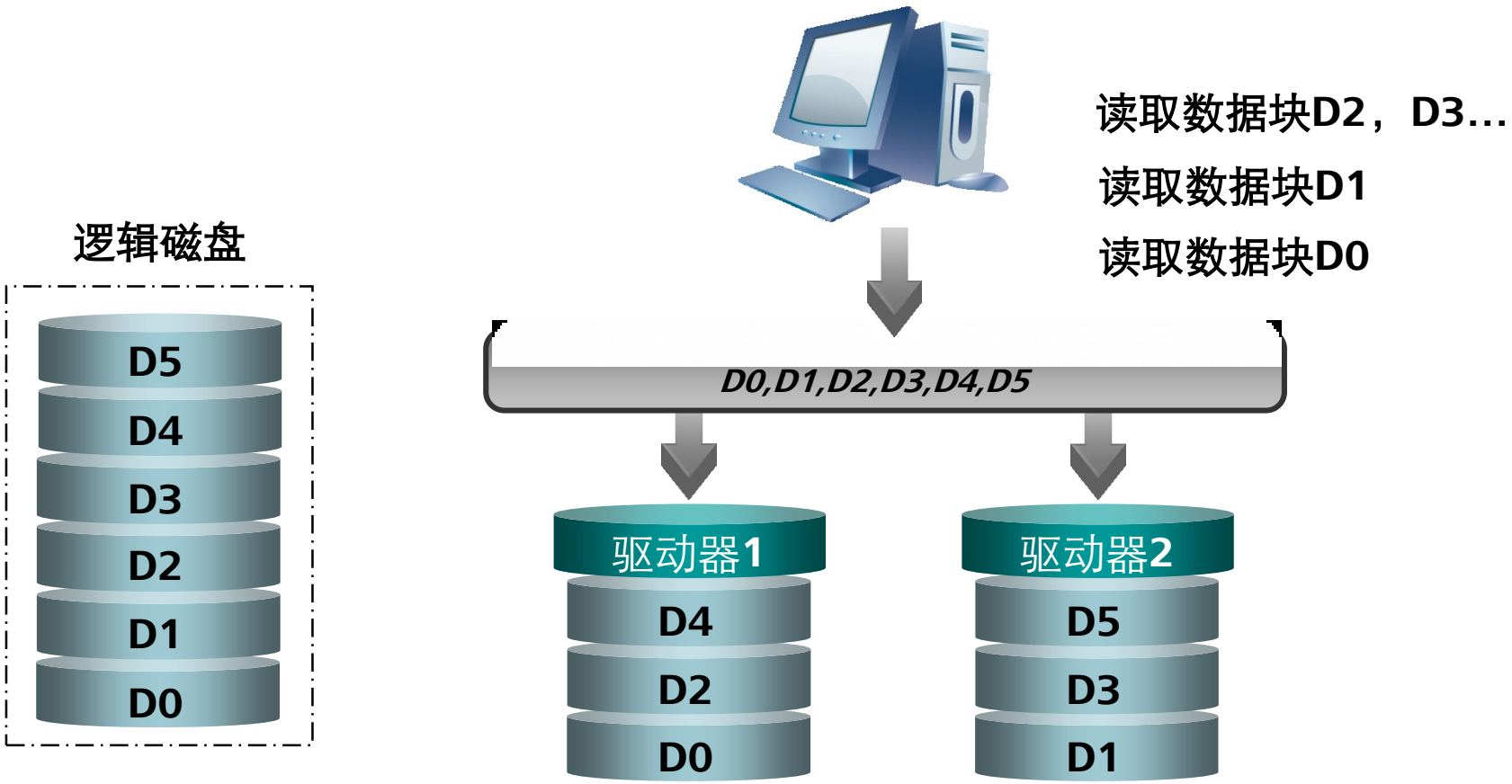


无差错控制的条带化阵列

RAID技术与应用 — RAID 0数据写入

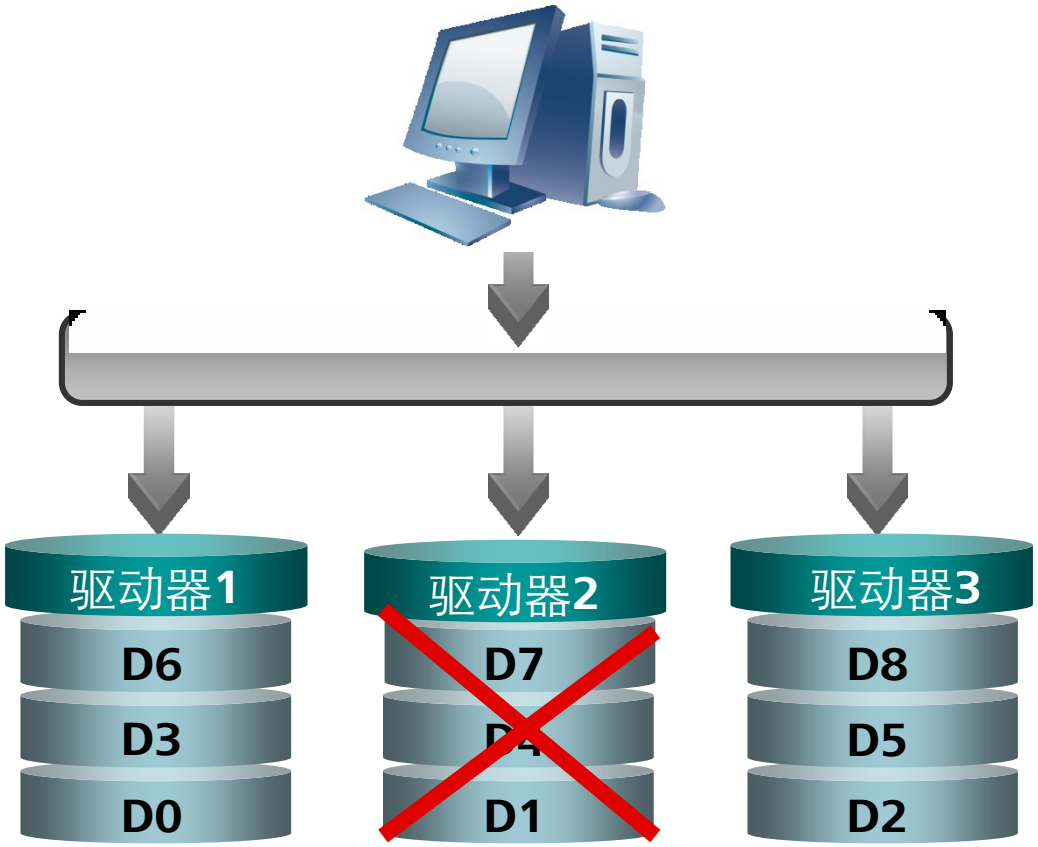


RAID技术与应用 — RAID 0数据读取



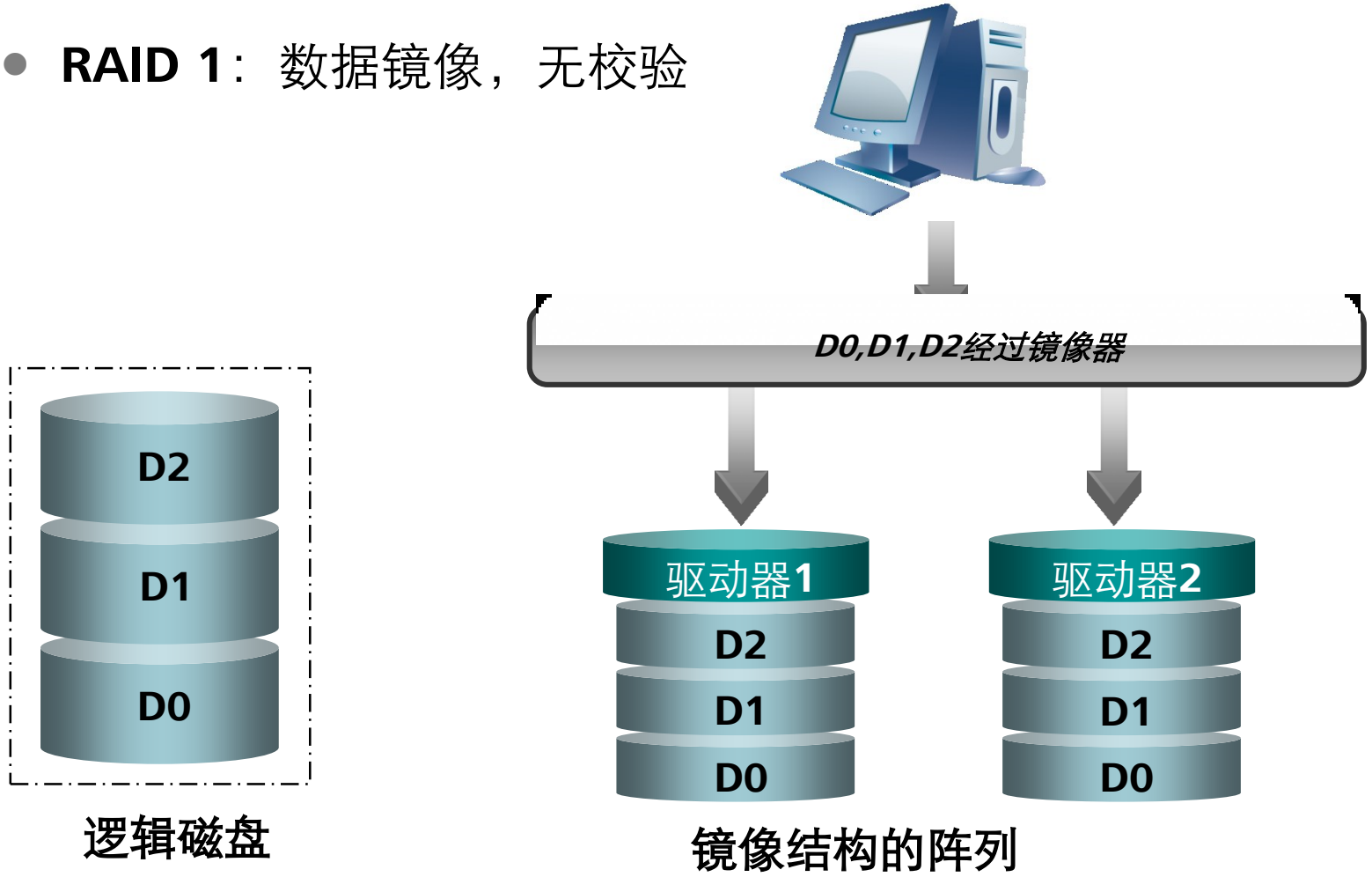
RAID技术与应用 — RAID 0数据恢复

- 阵列中某一个驱动器发生故障，将导致其中的数据丢失。

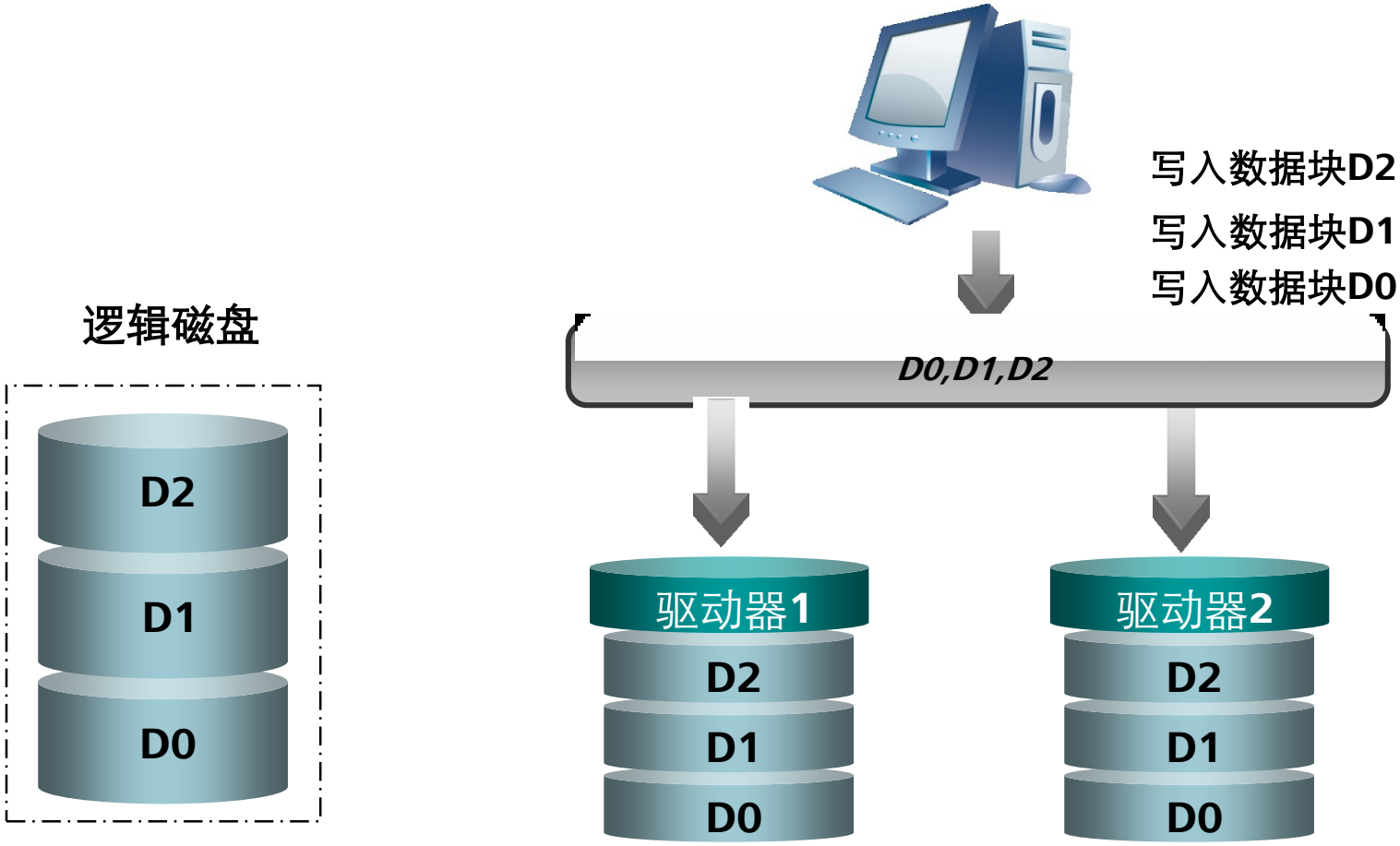


RAID技术与应用 — RAID 1

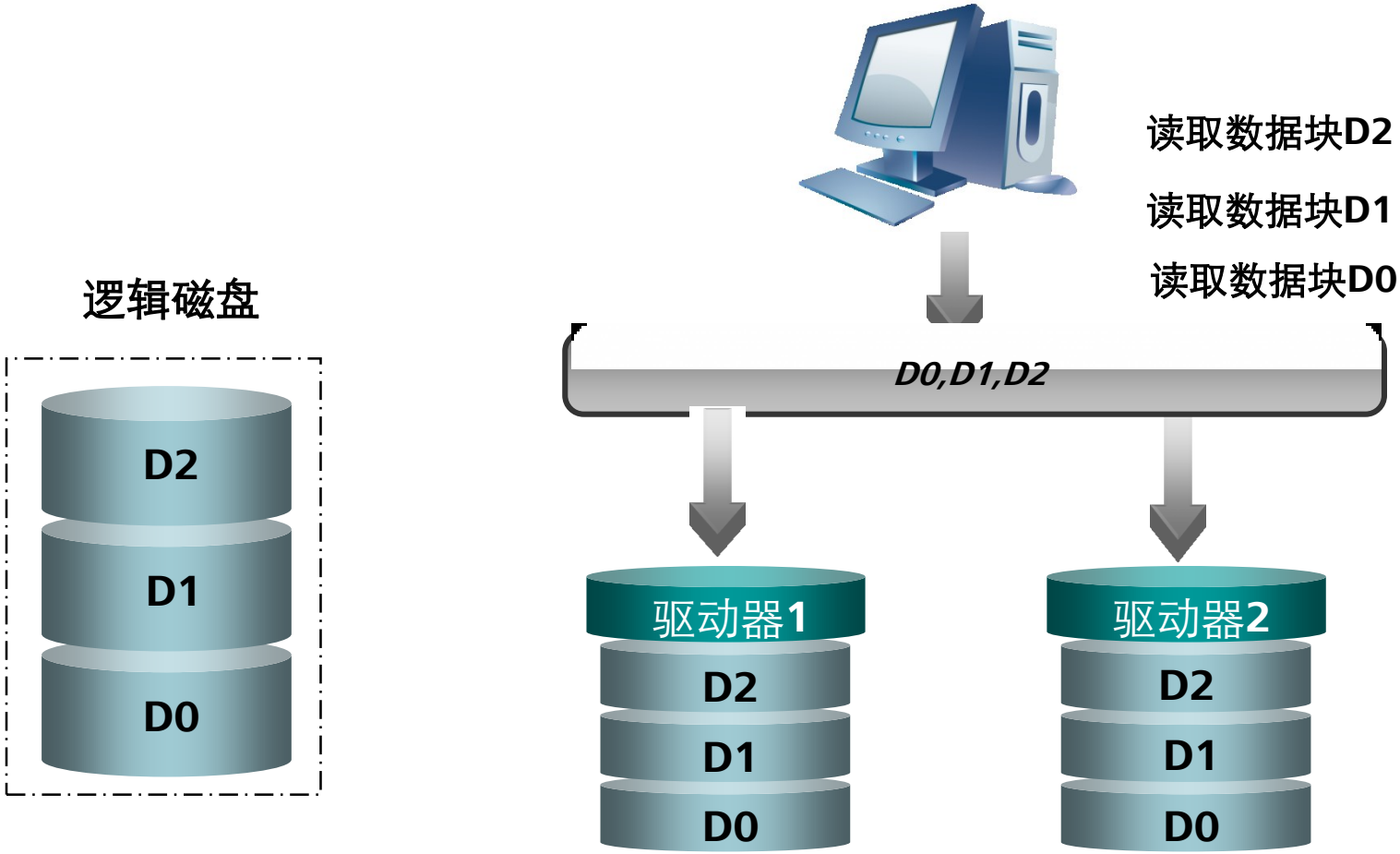
- **RAID 1**：数据镜像，无校验



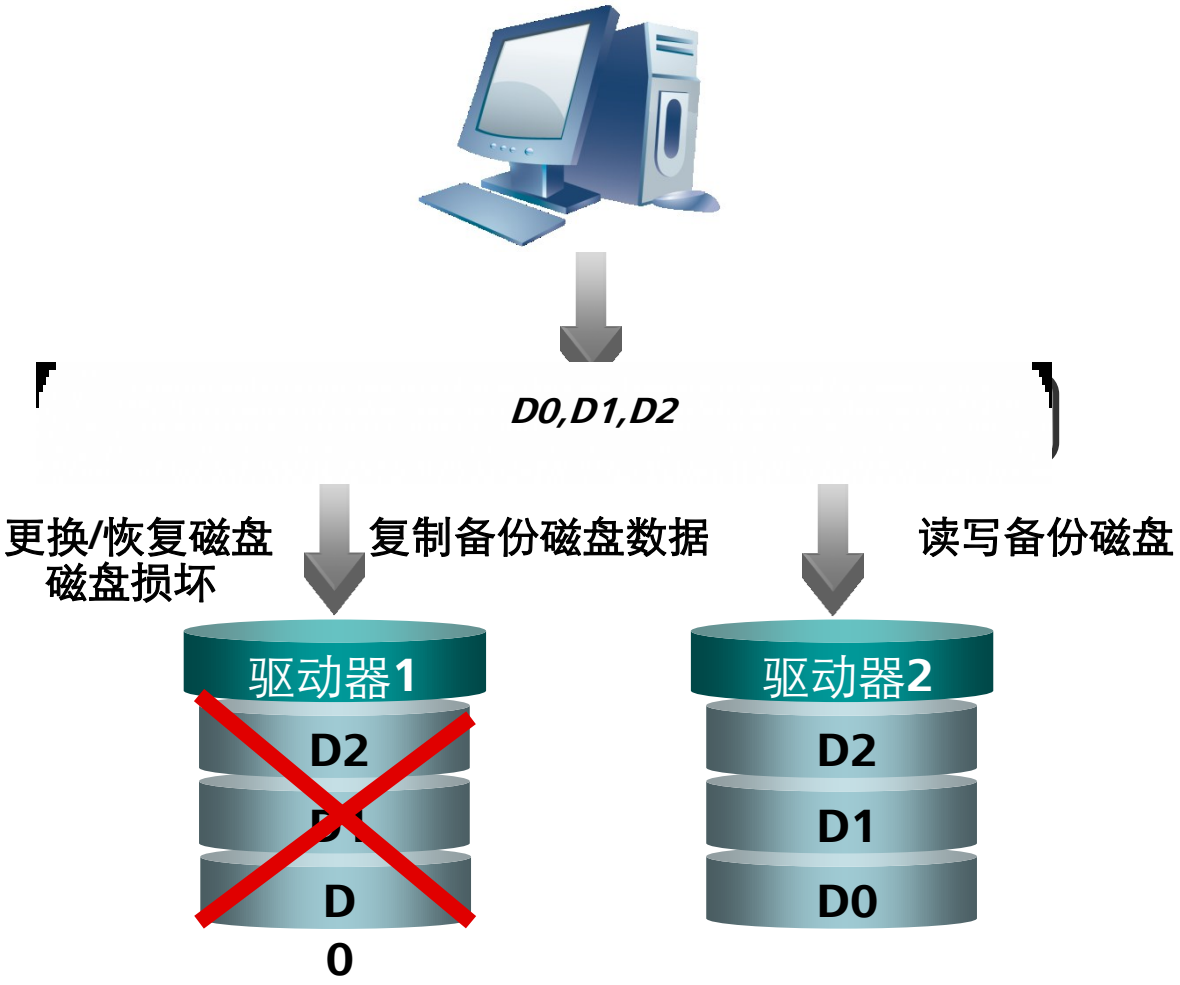
RAID技术与应用 — RAID1 数据写入



RAID技术与应用 — RAID1 数据读取

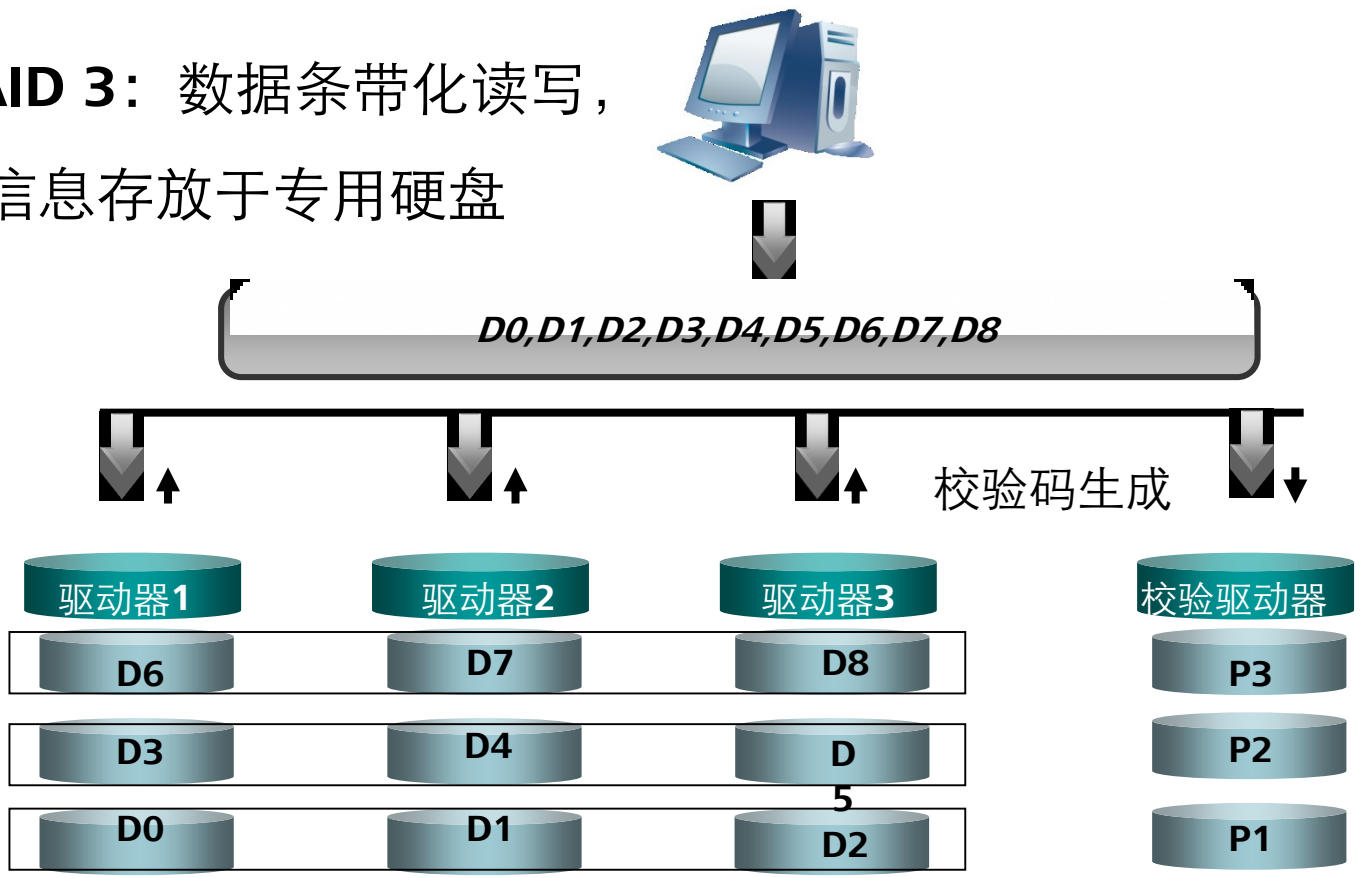


RAID技术与应用 — RAID 1数据恢复



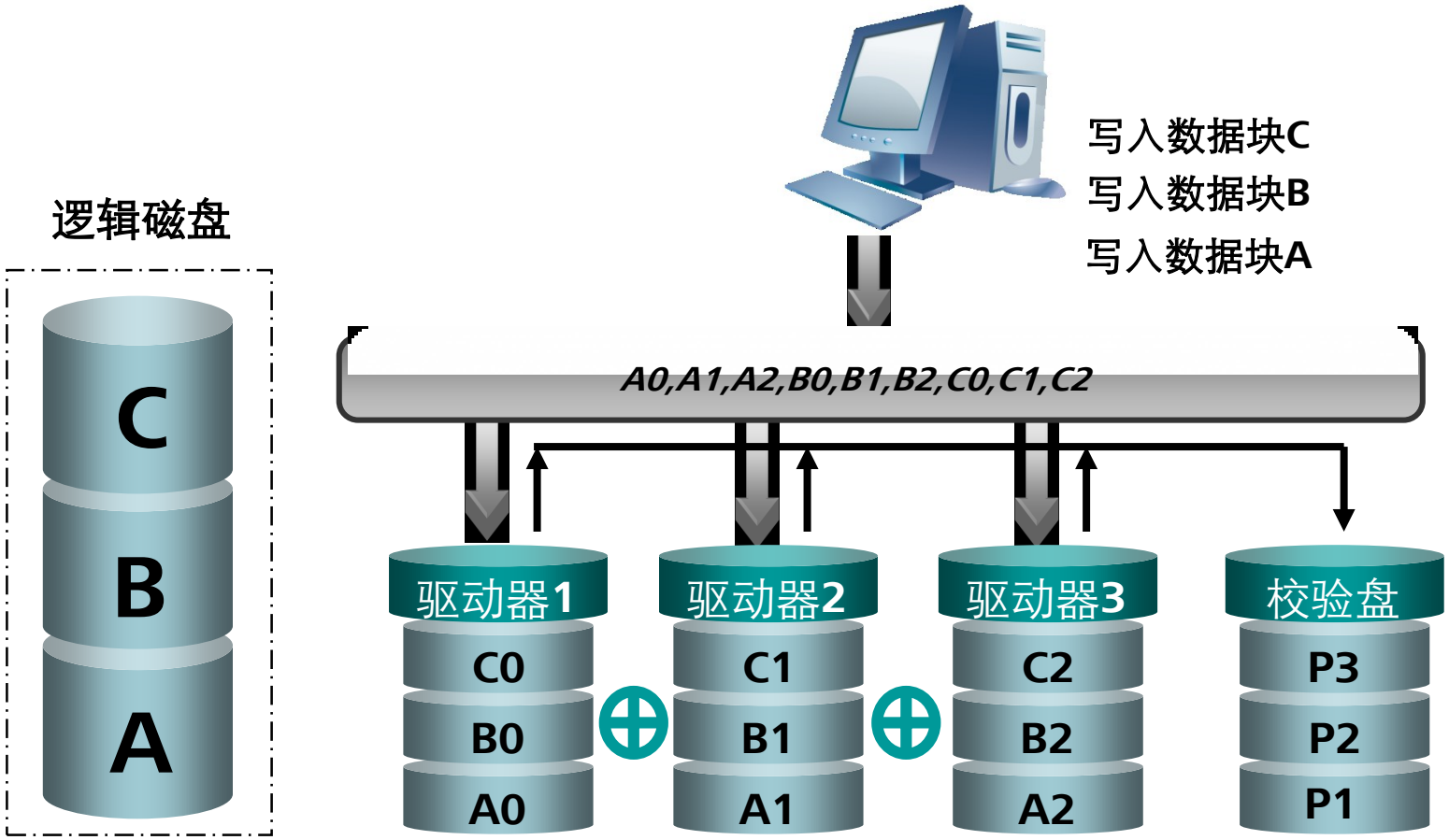
RAID技术与应用 — RAID 3

- **RAID 3**：数据条带化读写，校验信息存放于专用硬盘

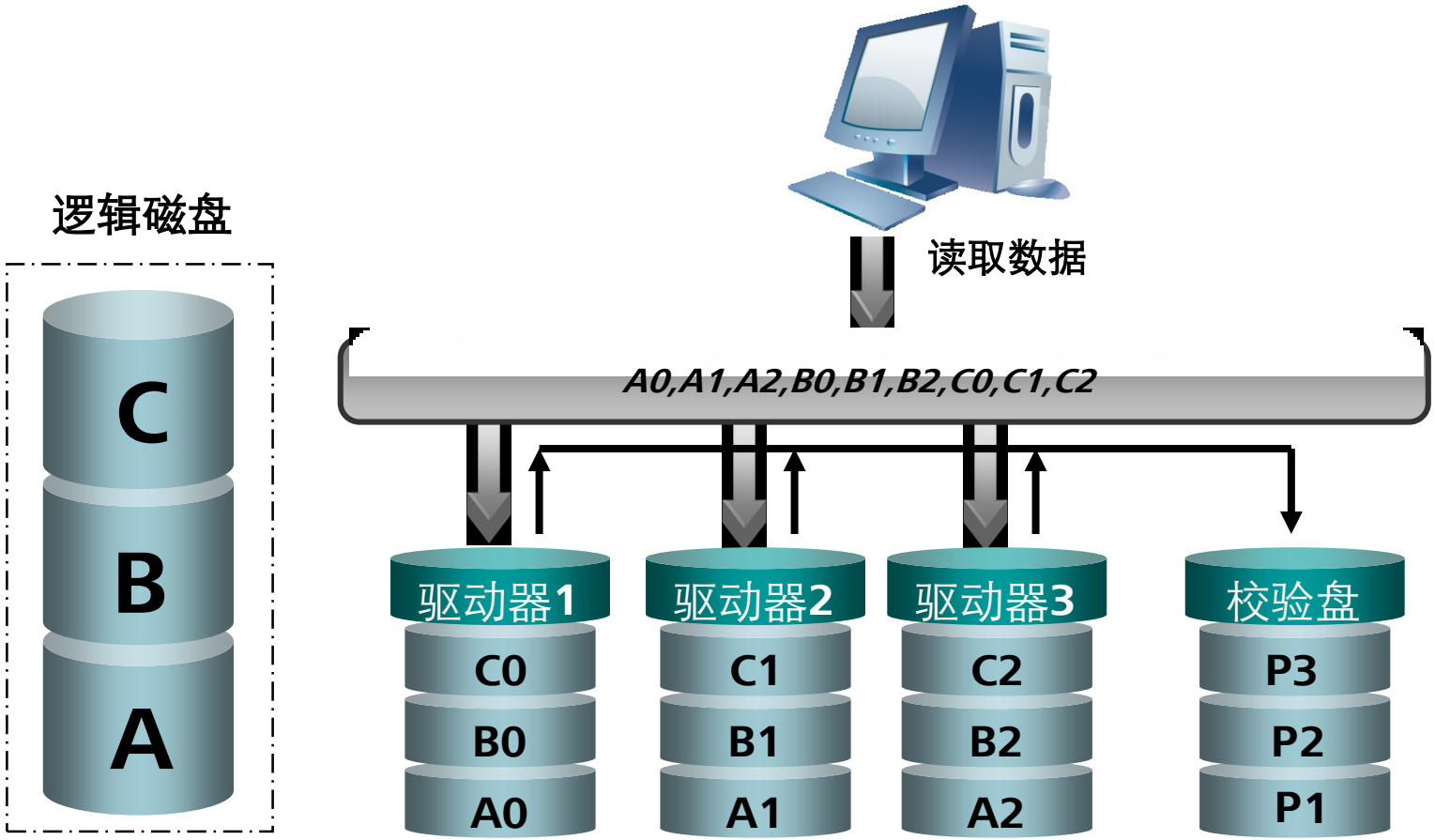


带奇偶校验码的并行阵列

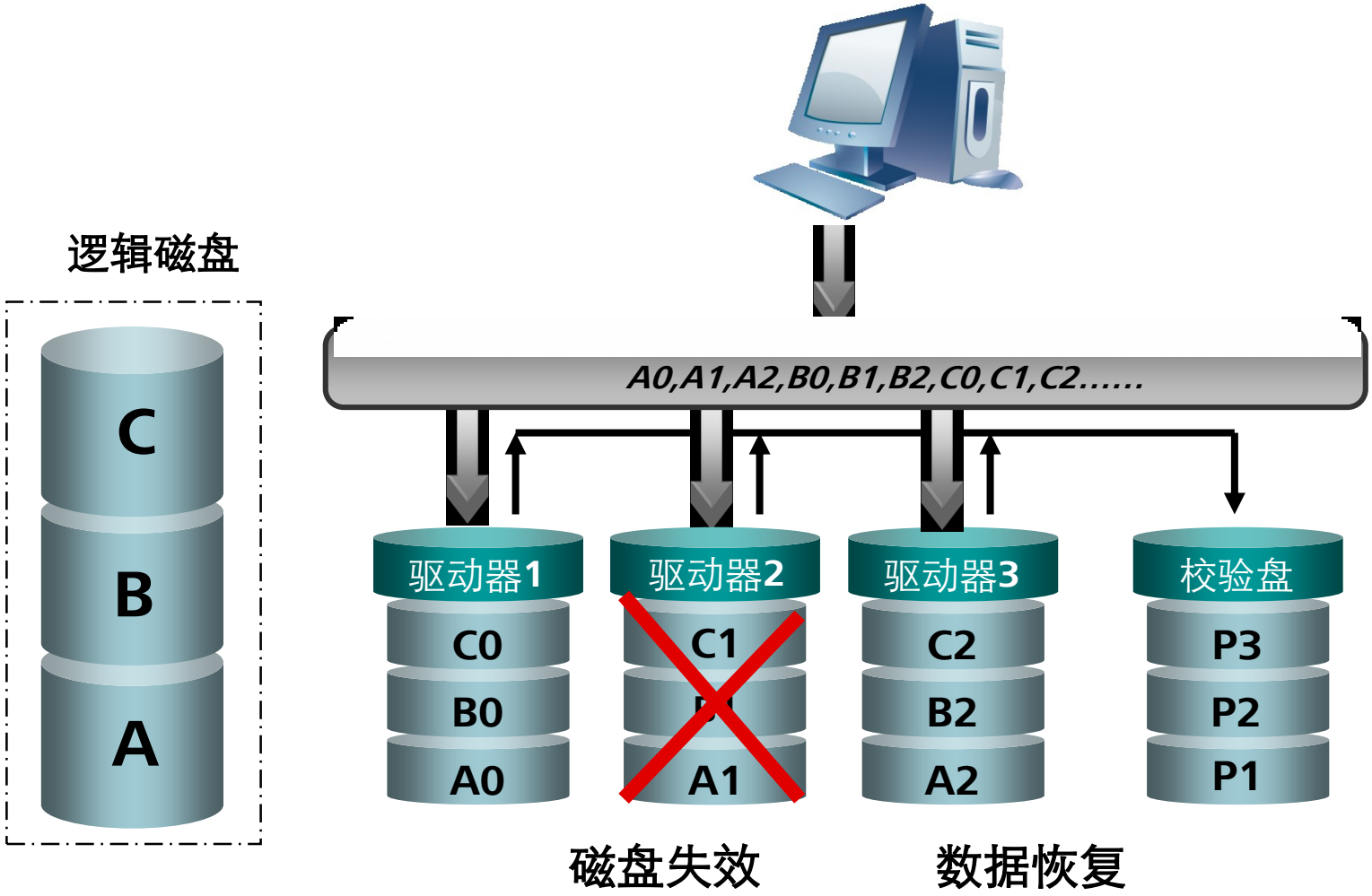
RAID技术与应用 — RAID 3数据写入



RAID技术与应用 — RAID 3数据读取

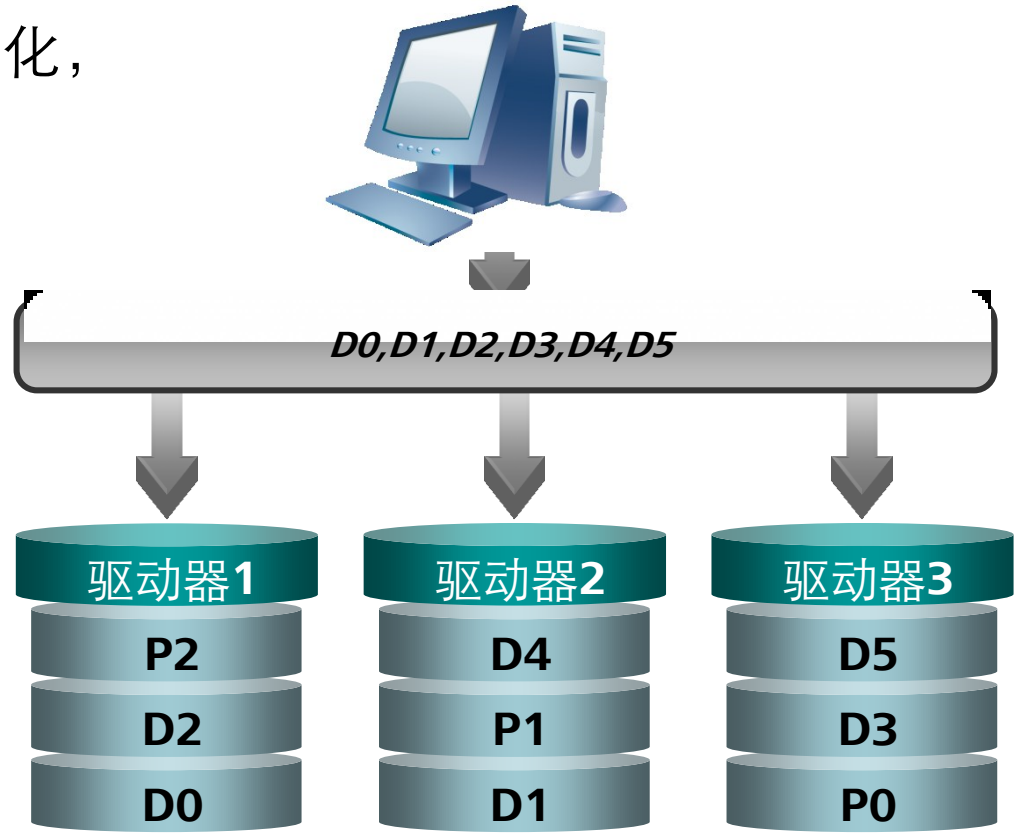


RAID技术与应用 — RAID 3数据恢复



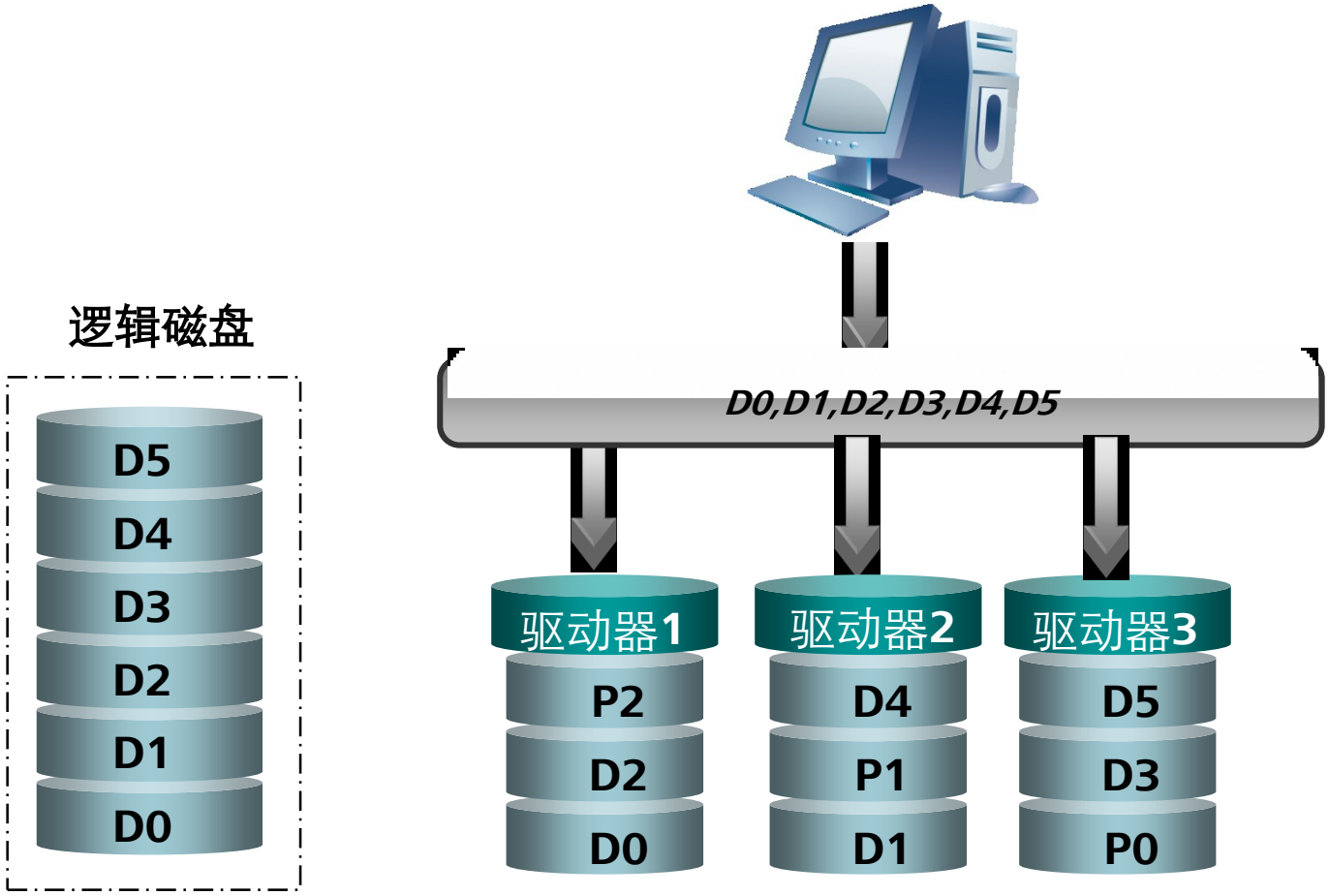
RAID技术与应用 — RAID 5

- **RAID 5**：数据条带化，校验信息分布式存放

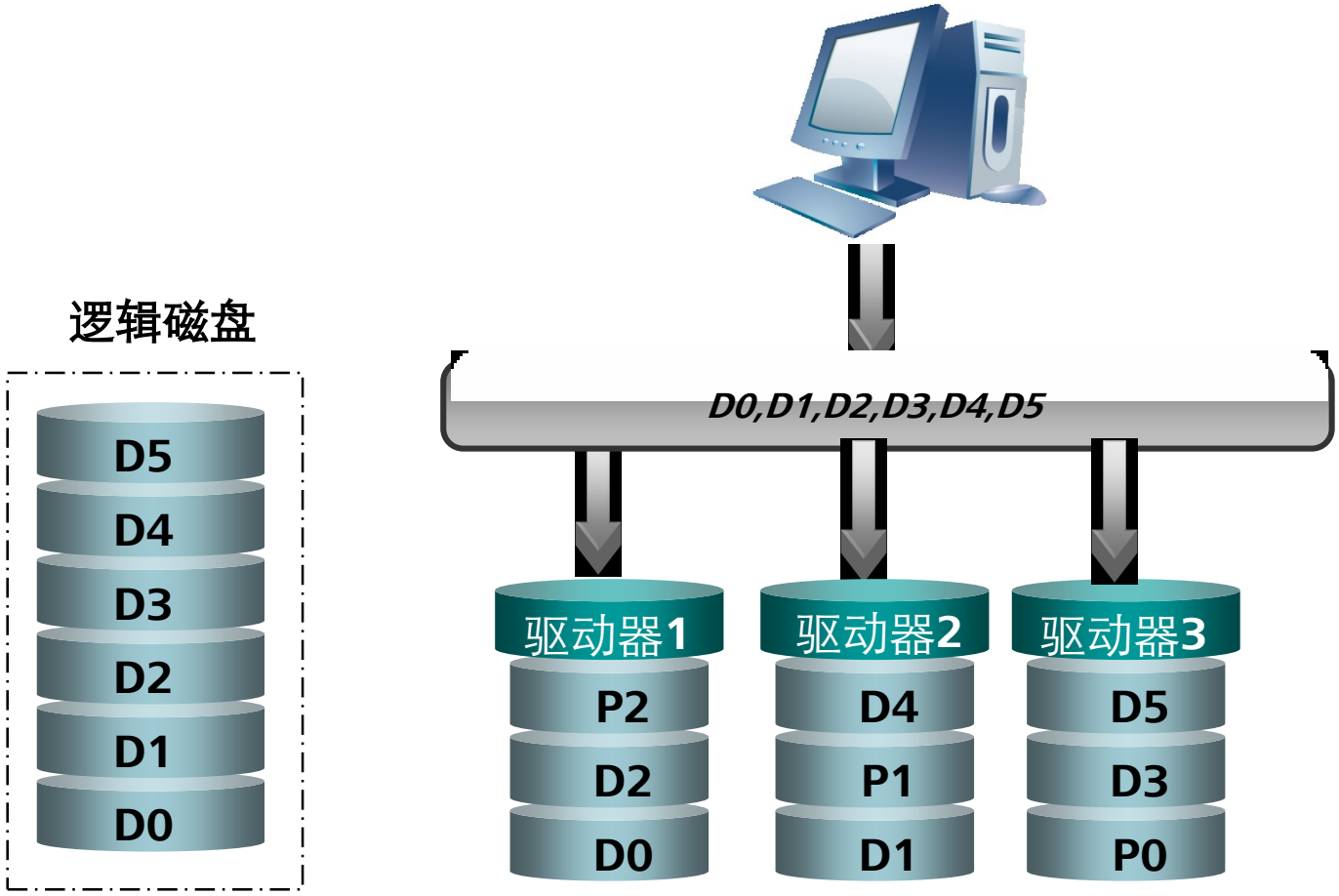


分布式奇偶校验码的独立磁盘结构

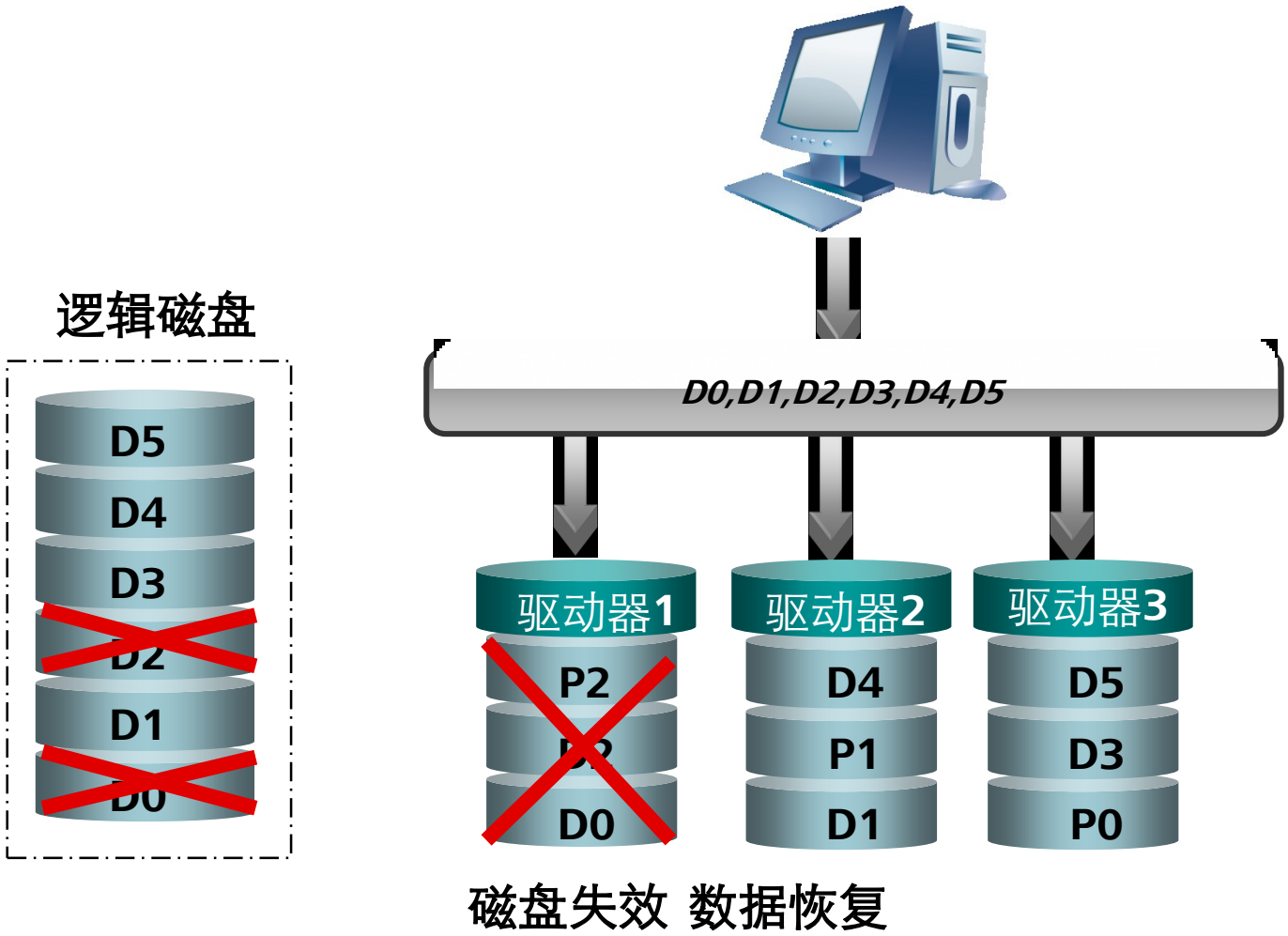
RAID技术与应用 — RAID 5数据写入



RAID技术与应用 — RAID 5数据读取



RAID技术与应用 — RAID 5数据恢复



磁盘失效 数据恢复

RAID技术与应用 — RAID 6

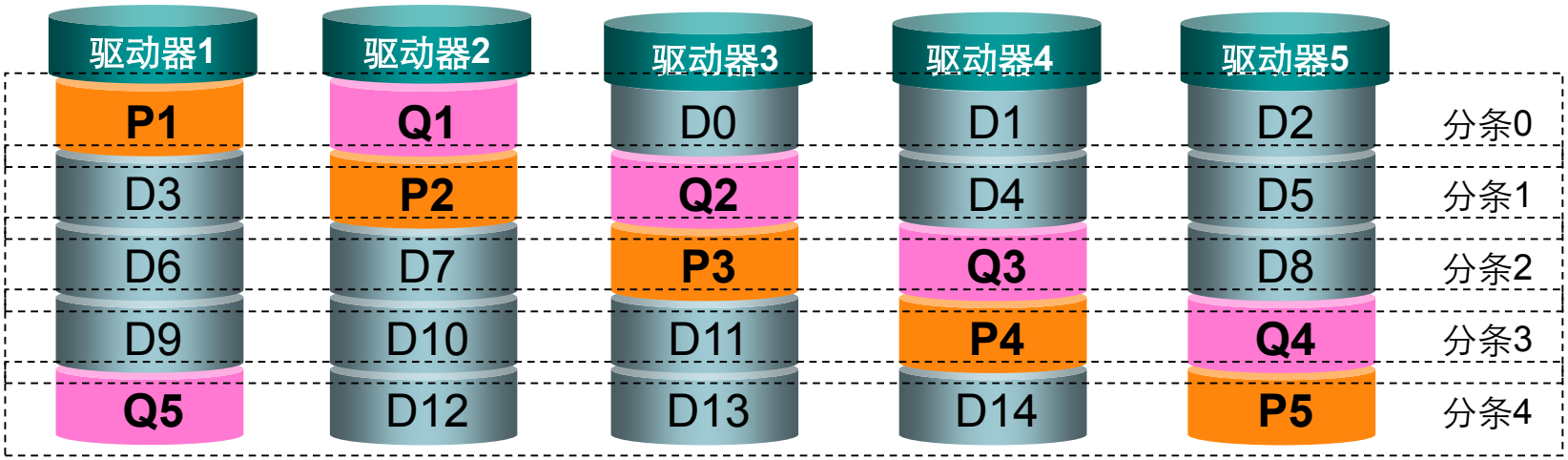
- **RAID 6**是带有两种校验的独立磁盘结构，采用两种奇偶校验方法，需要至少**N+2(N>2)**个磁盘来构成阵列，一般用在数据可靠性、可用性要求极高的应用场合 。
- 常用的**RAID 6**技术
  - **RAID6 P + Q**
  - **RAID6 DP**

RAID技术与应用 — RAID6 P+Q

- **RAID6 P+Q**需要计算出两个校验数据**P**和**Q**，当有两个数据丢失时，根据**P**和**Q**恢复出丢失的数据。校验数据**P**和**Q**是由以下公式计算得来的：

  P=D0⊕ D1 ⊕ D2 ……

  Q=(α⊗D0)⊕(β⊗D1)⊕(γ⊗D2)……

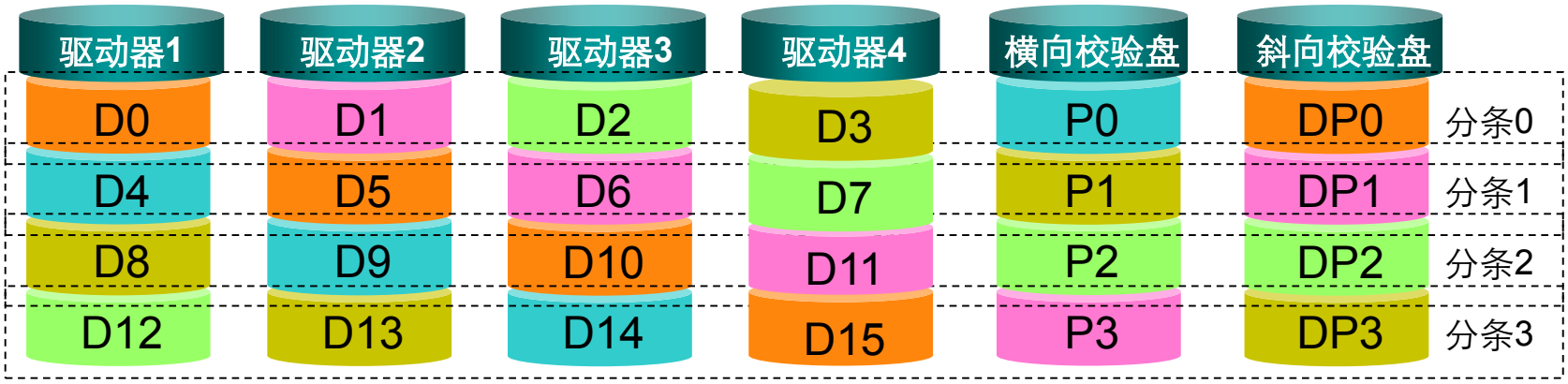


RAID技术与应用 — RAID6 DP

- **DP – Double Parity**，就是在**RAID4**所使用的一个行**XOR**校验磁盘的基础上又增加了一个磁盘用于存放斜向的**XOR**校验信息
  - 横向校验盘中**P0—P3**为各个数据盘中横向数据的校验信息

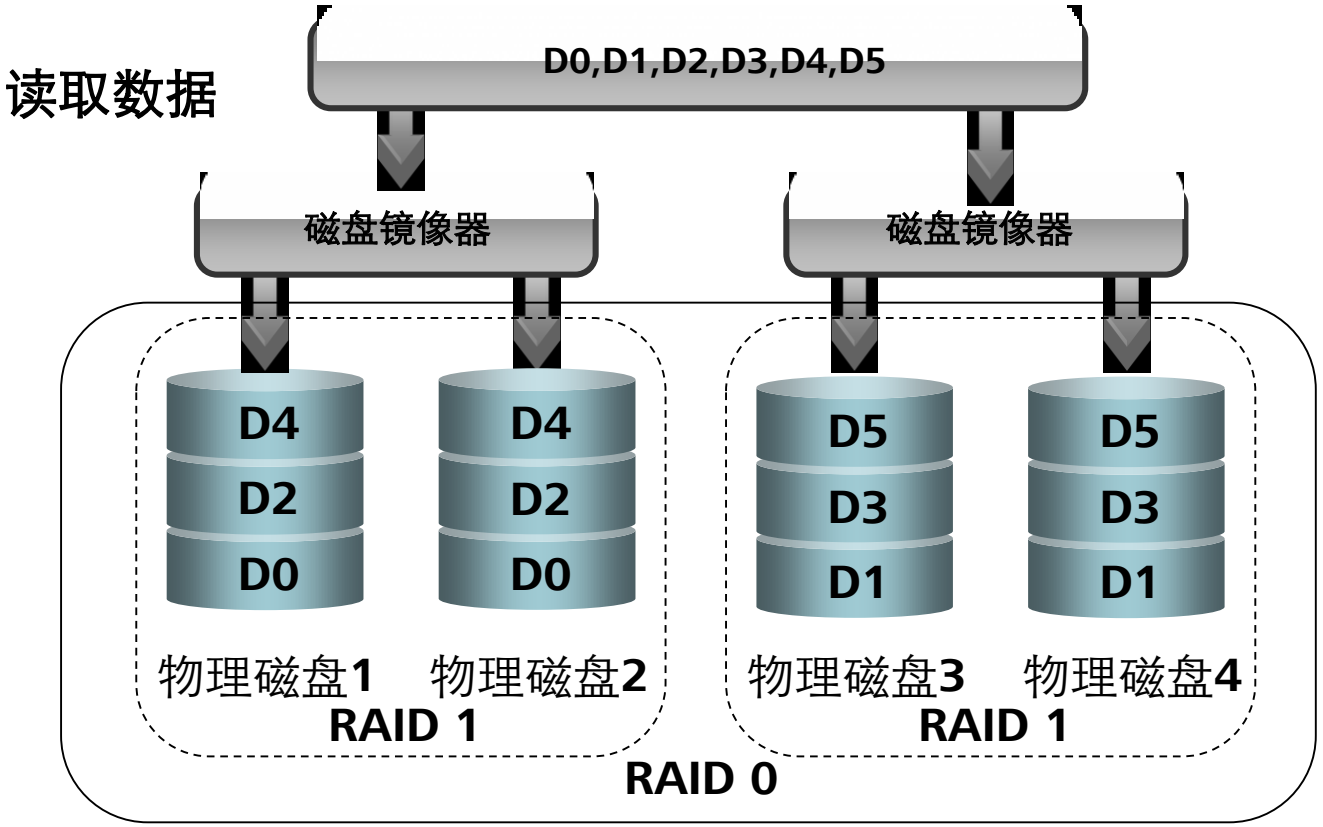
    例：**P0=D0 XOR D1 XOR D2 XOR D3**
  - 斜向校验盘中**DP0—DP3**为各个数据盘及横向校验盘的斜向数据校验信息

    例：**DP0=D0 XOR D5 XOR D10 XOR D15**



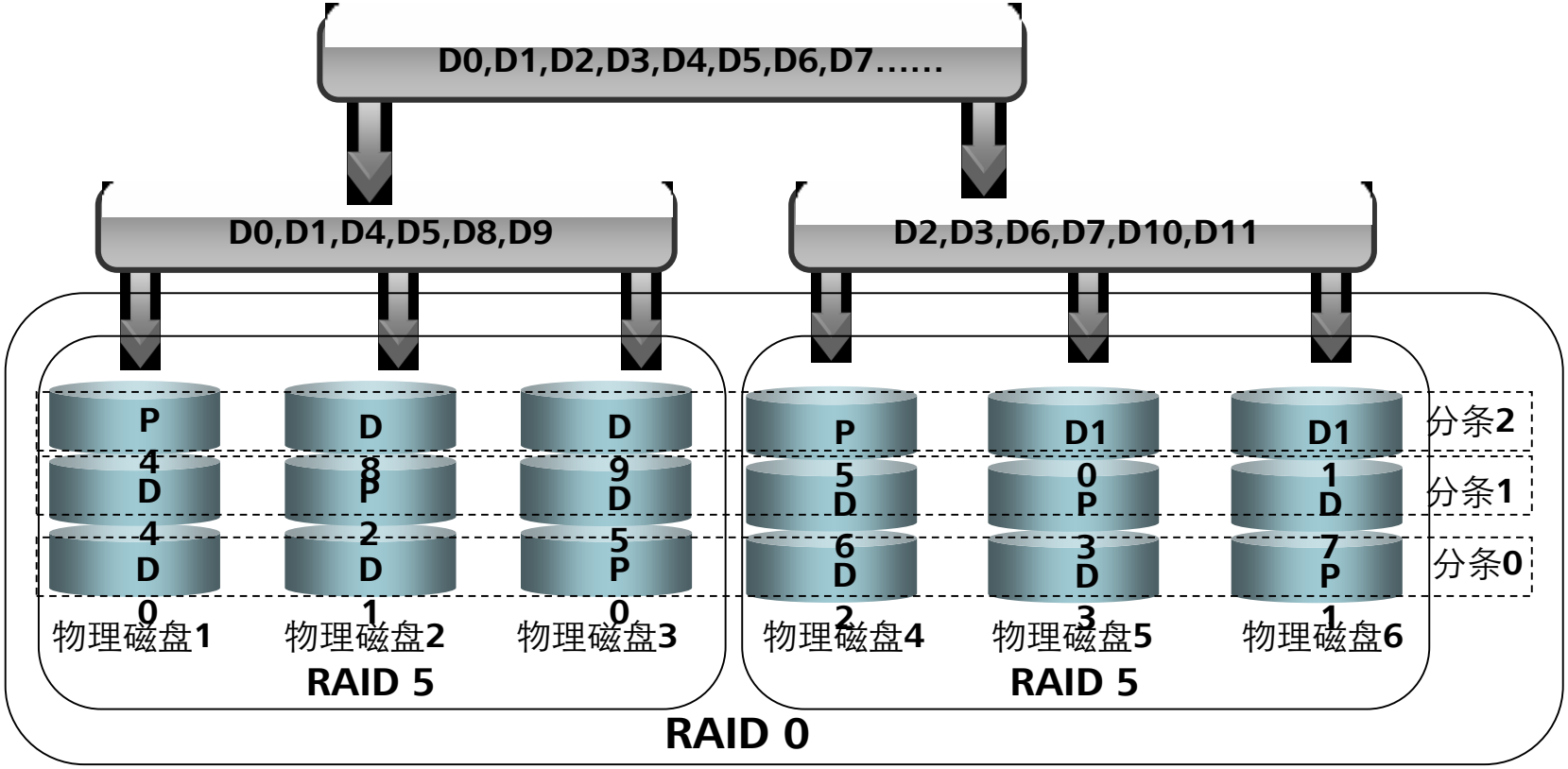
RAID技术与应用 — RAID 10

- **RAID 10**：是将镜像和条带进行组合的**RAID**级别，先进行**RAID1**镜像然后再做**RAID 0**。**RAID 10**也是一种应用比较广泛的**RAID**级别。



RAID技术与应用 — RAID50

- **RAID 50**：是将**RAID 5**和**RAID 0**进行两级组合的**RAID**级别，第一级是 **RAID 5**，第二级为**RAID 0**。



RAID技术与应用 — RAID比较

RAID级别	RAID 0	RAID 1	RAID 3	★RAID 5	★RAID6	★RAID 10	RAID 50
容错性	无	有	有	有	有	有	有
冗余类型	无	复制	奇偶校验	奇偶校验	奇偶校验	复制	奇偶检验
热备盘选项	无	有	有	有	有	有	有
读性能	高	低	高	高	高	一般	高
随机写性能	高	低	最低	低	低	一般	低
连续写性能	高	低	低	低	低	一般	低
最小硬盘数	2块	2块	3块	3块	4 块	4块	6块
可用容量	N * 单块硬盘容量	(1/N) * 单块硬盘容量	(N -1) * 单块硬盘容量	(N -1) * 单块硬盘容量	(N -2) * 单块硬盘容量	(N /2) * 单块硬盘容量	(N -2) * 单块硬盘容量

RAID技术与应用 — 应用场景

RAID级别	RAID 0	RAID 1	RAID 3	RAID 5 /6	RAID 10	RAID50
典型应用环境	迅速读写，安全性要求不高，如图形工作站等	随机数据写入，安全性要求高，如服务器、数据库存储领域	连续数据传输，安全性要求高，如视频编辑、大型数据库等	随机数据传输，安全性要求高，如邮件服务器，文件服务器等	数据量大，安全性要求高，如银行、金融等领域	随机数据传输，安全性要求高，并发能力要求高，如邮件服务器，www服务器等。

RAID技术与应用 — RAID级别选择

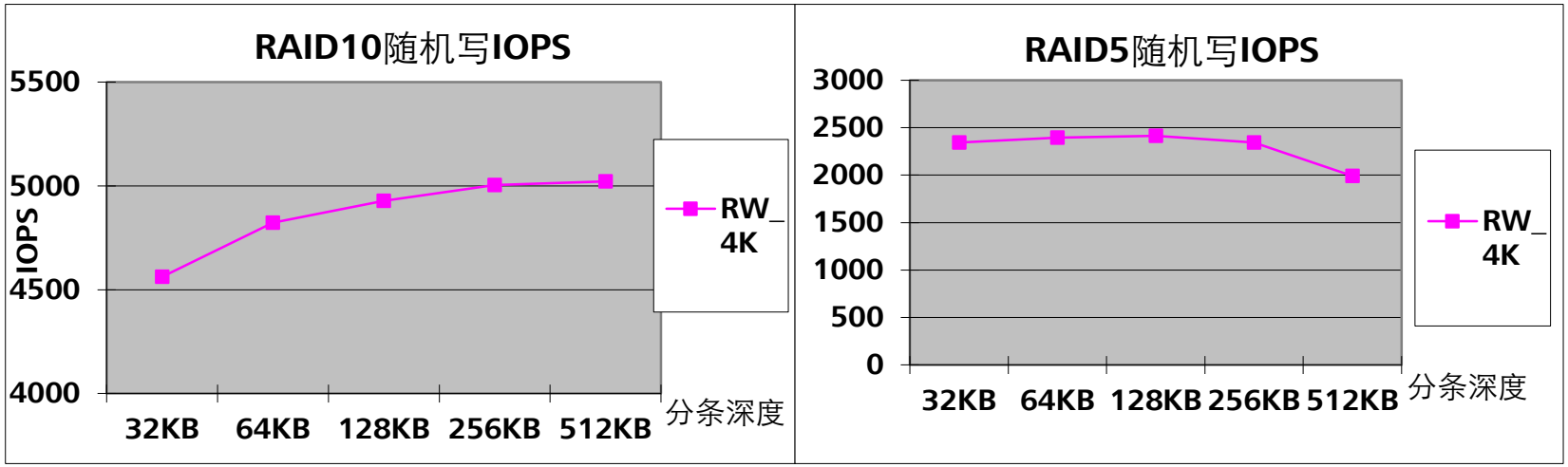
- 从可靠性、性能和成本简单比较各**RAID**级别的优劣（相对而言），供在实际项目中选择时参考。

参考	RAID 0	RAID 1	RAID 3	RAID 5	RAID 10	RAID6
可靠性	★	★★★★	★★	★★★	★★★★	★★★★
性能	★★★★	★★★★	★★★	★★★	★★★★	★★
成本	★★★★	★★	★★★	★★★	★★	★★

- **空间利用率:**
  RAID5明显优于**RAID10**
- **可靠性:**
  RAID5低于**RAID10**
- **性能:**
  业务是一些大文件的读写操作时，**RAID5**的性能会明显好于**RAID10**
  业务以随机的小数据块读写为主的时候，**RAID10**是最优的选择

RAID技术与应用 — 随机写性能比较

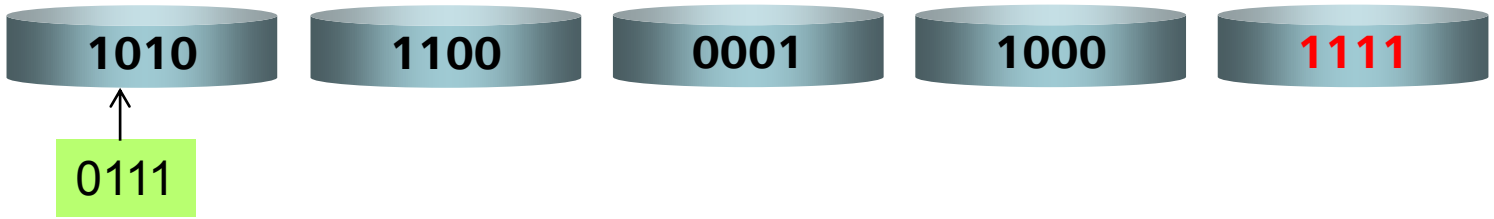
- **RAID5和RAID10下分条深度变化随机写性能规律**



- **RAID5规律**：随着分条深度的增加，随机写**IOPS**先会不断的增加，到达一定程度之后，随机写**IOPS**会不断的递减;
- **RAID10规律**：随着分条深度的增加，随机写**IOPS**不断的增长，当分条深度增大到一定程度后，随机写**IOPS**保持一个较为稳定的状态;

RAID技术与应用 — 写惩罚

假设由5块硬盘组成的RAID5，每块盘同一条带数据如下：



如果有一个数据要写入，假设在第**1**个磁盘上写入的数据为**0111**。那么整个**RAID5**需要完成写入的过程分为：

1. 读取原数据，然后与新数据**0111**做**XOR**操作： **1010 XOR 0111 = 1101**
2. 读取原有的校验位**1111**
3. 用第一步算出的数值与原校验再做一次**XOR**操作： **1101 XOR 1111 = 0010**
4. 然后将**0111**新数据写入到数据磁盘，将第三步计算出来的新的校验位写入校验盘。

由上述几个步骤可见，对于任何一次写入，在存储端，需要分别进行两次读+两次写，所以说**RAID5**的**Write Penalty**的值是**4**

RAID技术与应用 — 写惩罚（续）

常见RAID级别的Write Penalty值：

RAID	Write Penalty
0	1
1	2
5	4
6	6
10	2

在实际存储方案设计的过程中，计算实际可用**IOPS**的过程中必须纳入**RAID**的写惩罚计算。
计算的公式如下：
物理磁盘总的**IOPS** = 物理磁盘的**IOPS** × 磁盘数目

可用的**IOPS** = (物理磁盘总的**IOPS** × 写百分比 ÷ **RAID**写惩罚) + (物理磁盘总的**IOPS** × 读百分比)

假设组成**RAID5**的物理磁盘总共可以提供500 I**OPS**，使用该存储的应用程序读写比例是50%/50%，那么对于前端主机而言，实际可用的I**OPS**是：

(500 ×50% ÷ 4)+ (500 * 50%) = 312.5 IOPS

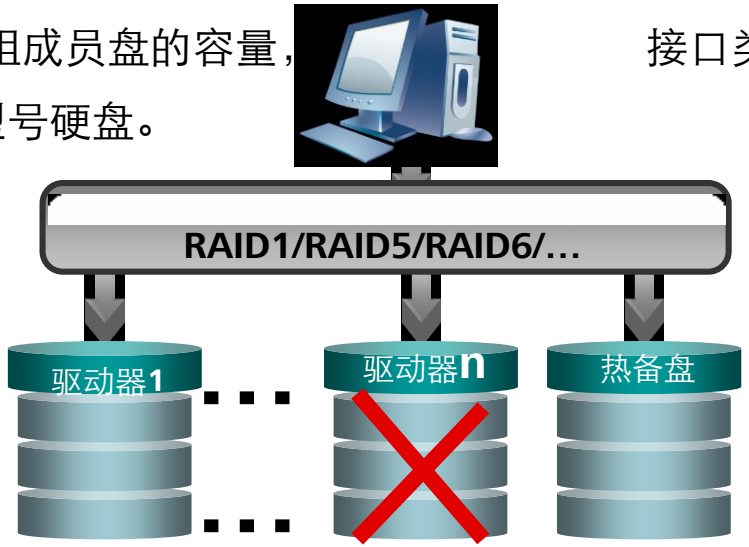
目　录

HUAWEI

RAID数据保护 — 热备盘

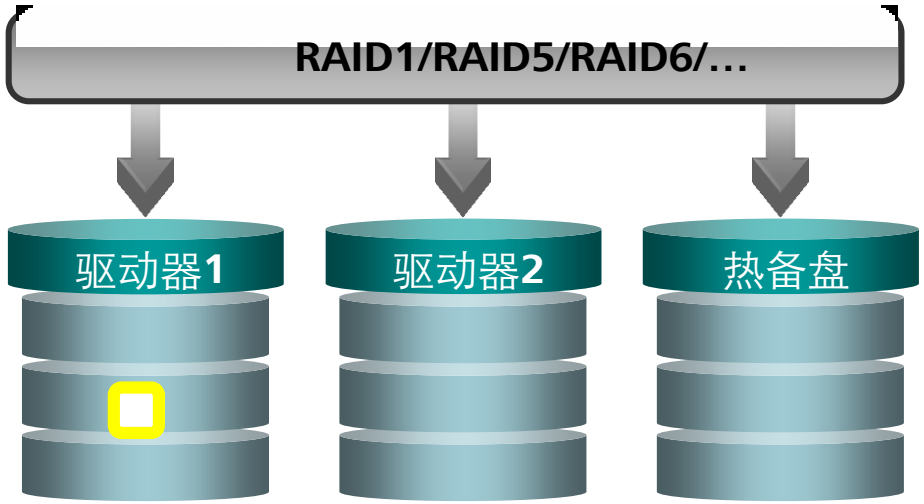
- **热备（Hot Spare）**：当冗余的**RAID**阵列中某个磁盘失效时，在不干扰当前**RAID**系统正常使用的情况下，用**RAID**系统中另外一个正常的备用磁盘顶替失效磁盘。
  - 全局热备盘
  - 局部热备盘。

热备盘要求和**RAID**组成员盘的容量，接口类型，速率一致，最好是采用同一厂家的同型号硬盘。



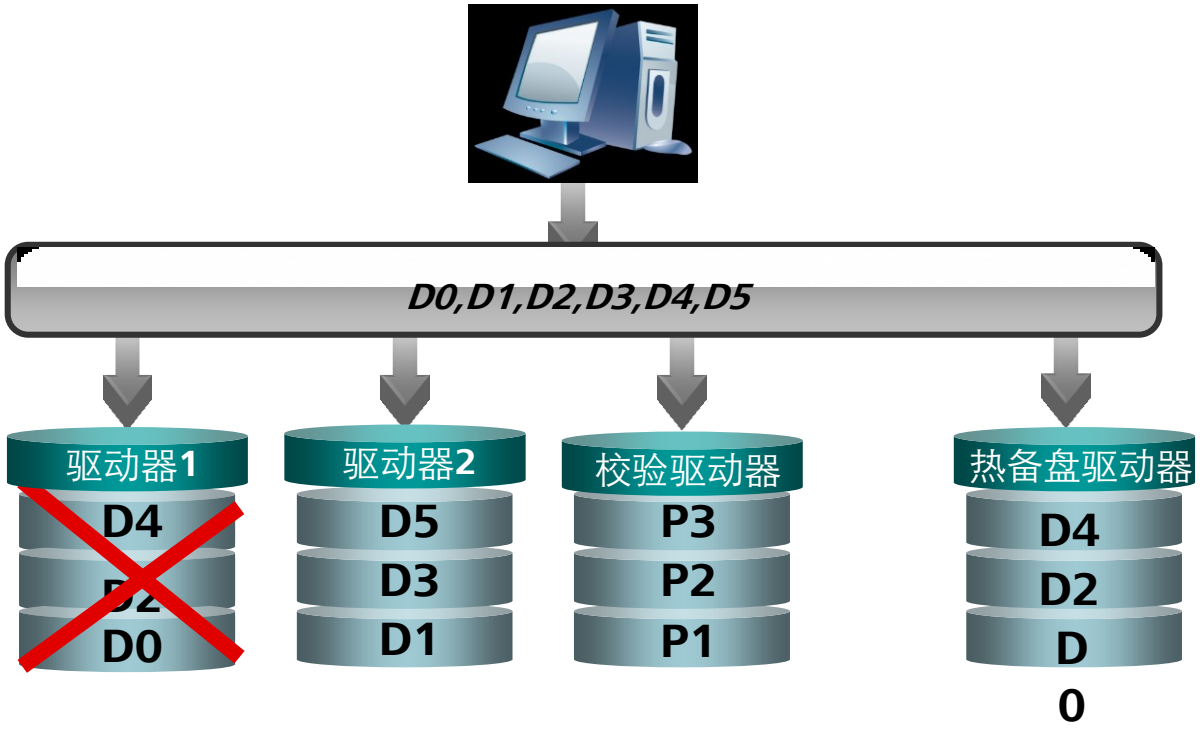
RAID数据保护 — 预拷贝

- **预拷贝：** 系统通过监控发现**RAID**组中某成员盘即将故障时，将即将故障成员盘中的数据提前拷贝到热备盘中，有效降低数据丢失风险。



RAID数据保护 — 重构

- **重构：RAID**阵列中发生故障的磁盘上的所有用户数据和校验数据的重新生成，并将这些数据写到热备盘上的过程。

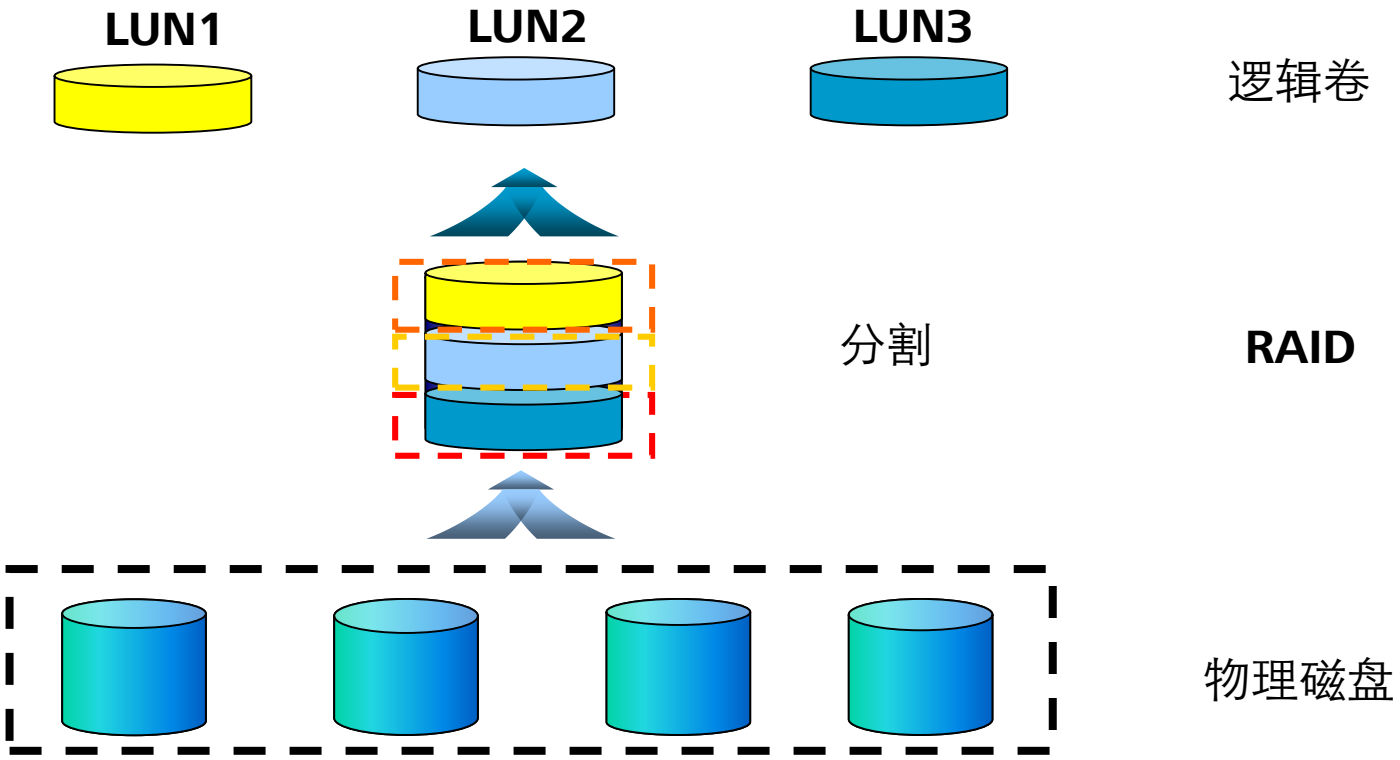


目　录

HUAWEI

RAID与LUN

- **RAID**由几个硬盘组成，从整体上看相当于由多个硬盘组成的一个大的物理卷。在物理卷的基础上可以按照指定容量创建一个或多个逻辑单元，这些逻辑单元称作**LUN**，可以做为映射给主机的基本块设备。



目　录

HUAWEI

目　录

RAID 2.0原理

RAID 2.0：为增强型**RAID**技术，有效解决了机械硬盘容量越来越大，重构一块机械硬盘所需时间越来越长，传统**RAID**组重构窗口越来越大而导致重构期间又故障一块硬盘而彻底丢失数据风险的问题。

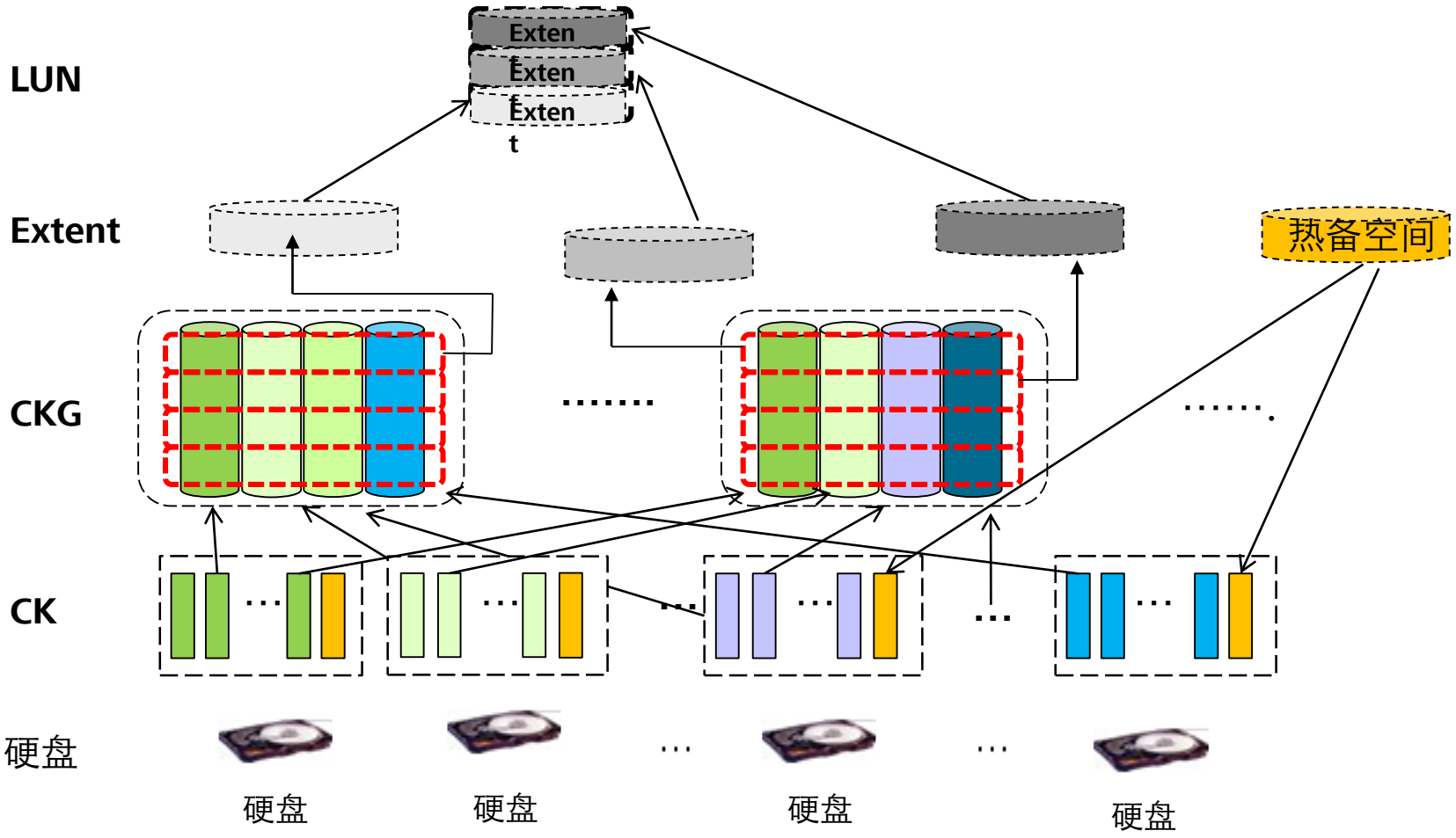
RAID2.0+：在**RAID 2.0**的基础上提供了更细粒度（可以达几十**KB**粒度）的资源颗粒，形成存储资源的标准分配及回收单位，类似计算虚拟化中的虚拟机，我们称之为虚拟块技术。

华为RAID2.0+：是华为针对传统 **RAID** 的缺点，设计的一种满足存储技术虚拟化架构发展趋势的全新的 **RAID** 技术，其变传统固定管理模式为两层虚拟化管理模式，在底层块级虚拟化（**Virtual for Disk**）硬盘管理的基础之上，通过一系列 **Smart** 效率提升软件，实现了上层虚拟化（**Virtual for Pool**）的高效资源管理。

RAID 2.0原理（续）

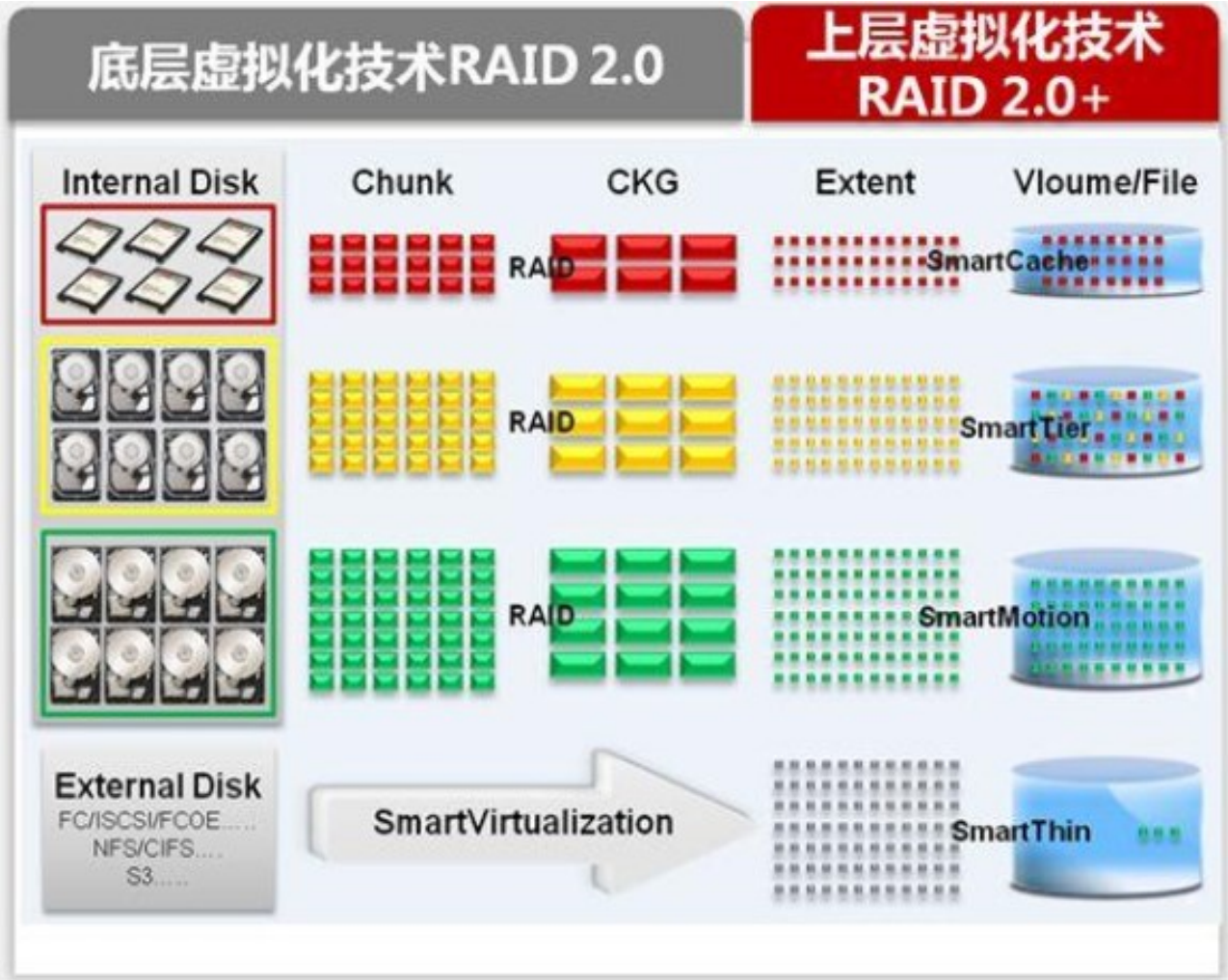
- 将硬盘划分成若干个连续的固定大小的存储空间，称为存储块（**chunk**）简称**CK**。
- **Ck**按**RAID**策略组合成**RAID**组，称为存储块组（**chunk group**）简称**CKG**。
- 在**CKG**中划分若干小数据块（**extent**），**LUN**就是由来自不同**CKG**的**extent**组成。
- 用作热备空间的**CK**也是分散在各个盘上的。

RAID 2.0原理（续）

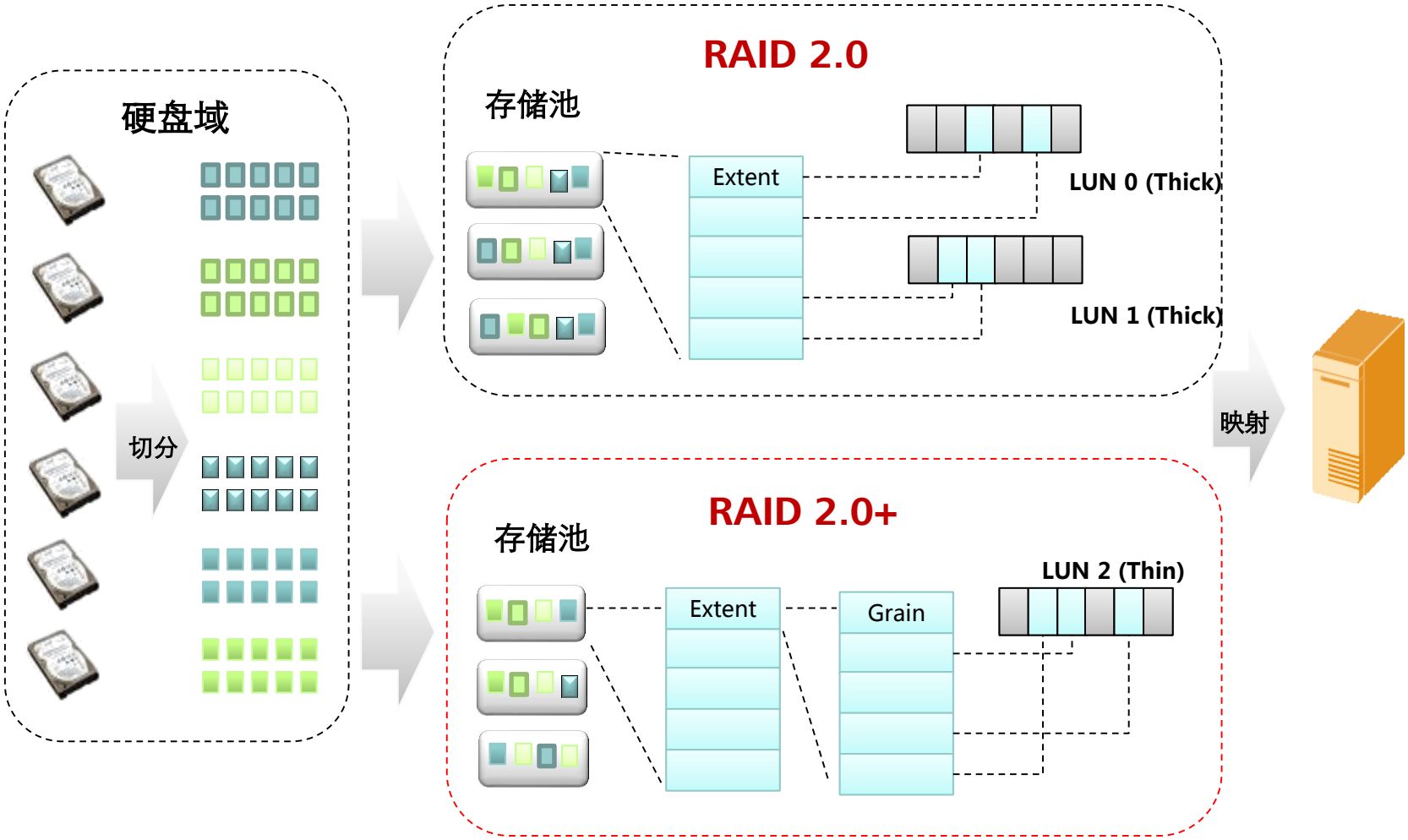


目　录

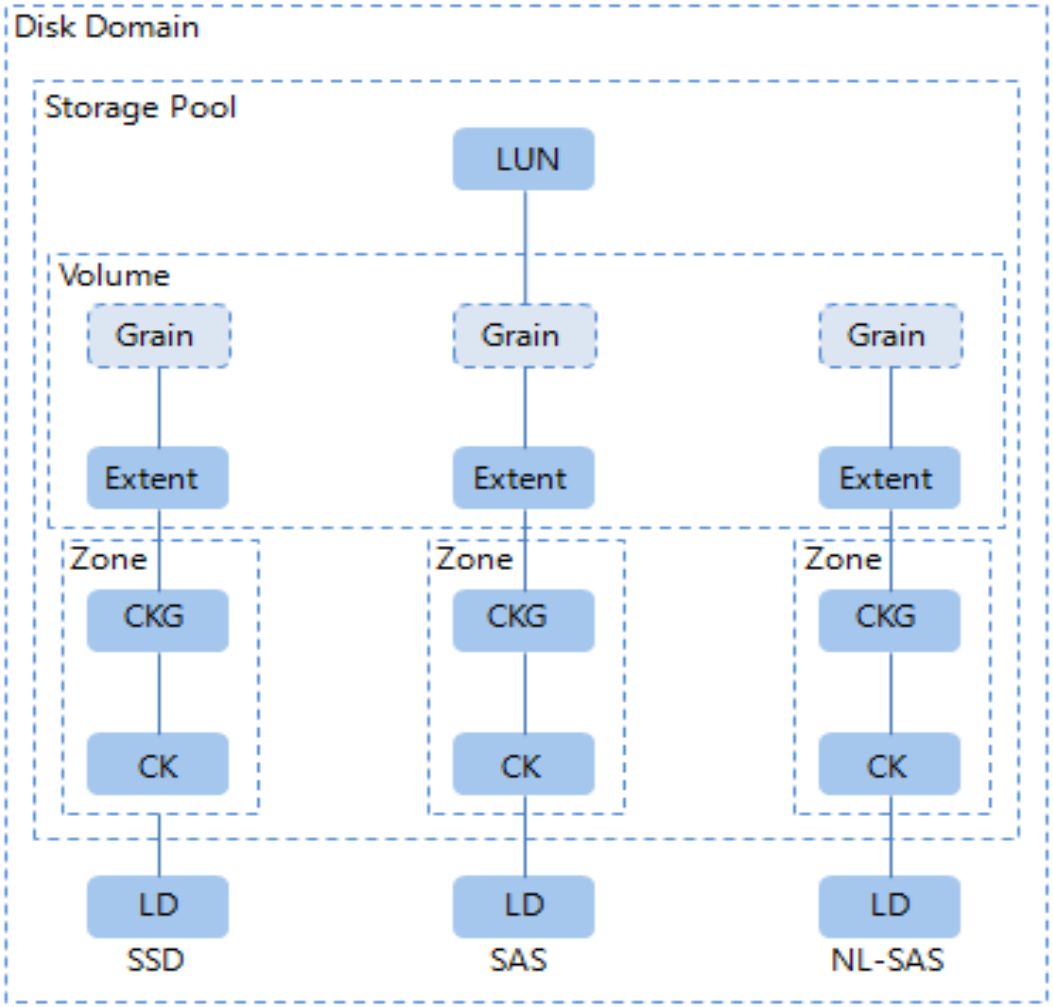
RAID 2.0+原理



RAID2.0+原理（续）

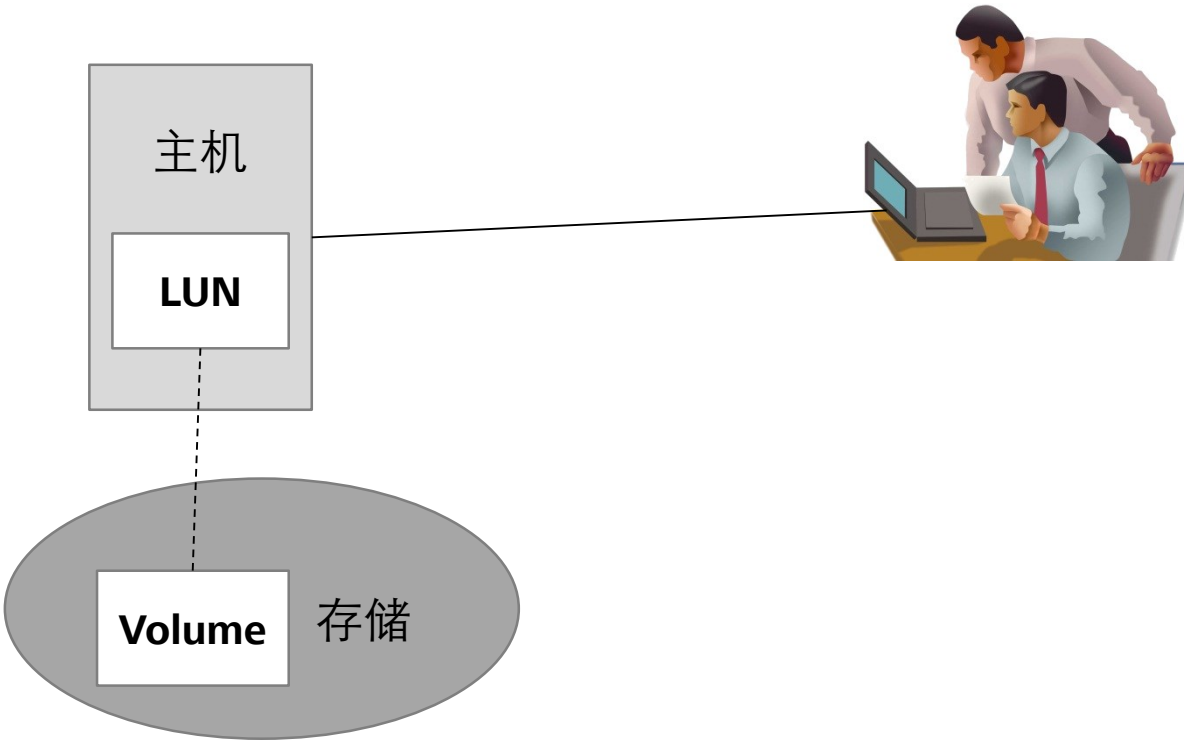


RAID2.0+原理 — 软件逻辑对象



RAID2.0+原理 — Volume & LUN

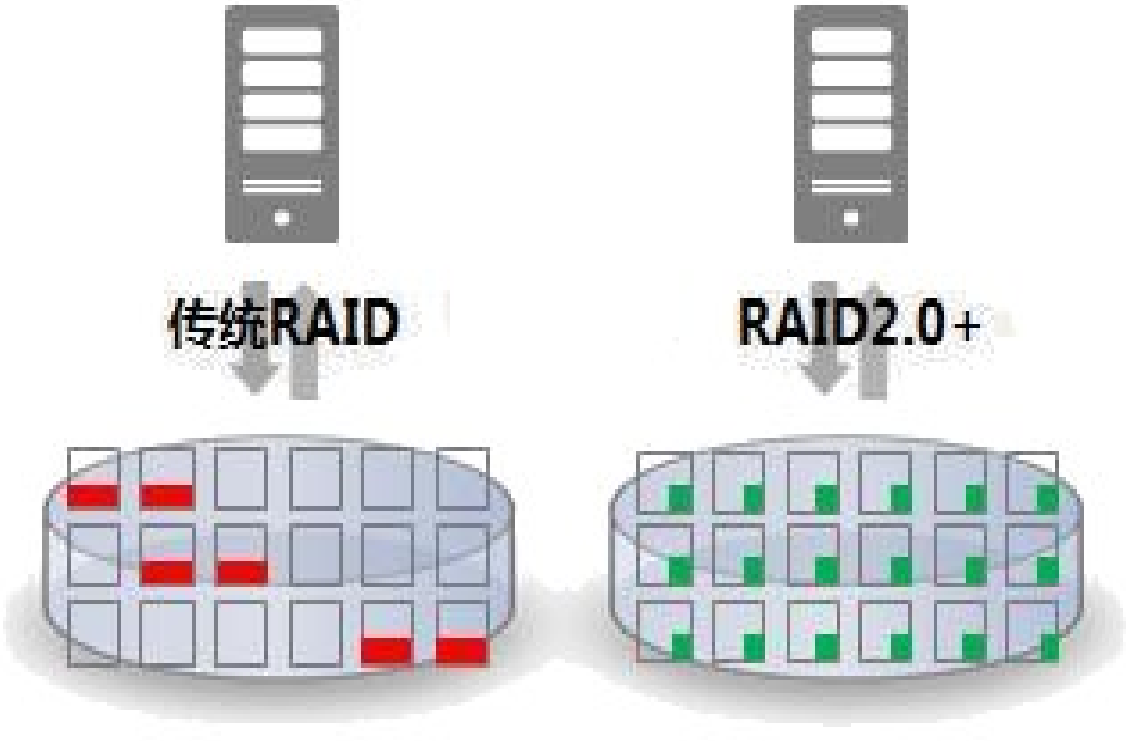
- **Volume**即卷，是存储系统内部管理对象。
- **LUN**是可以直接映射给主机读写的存储单元，是**Volume**对象的对外体现。



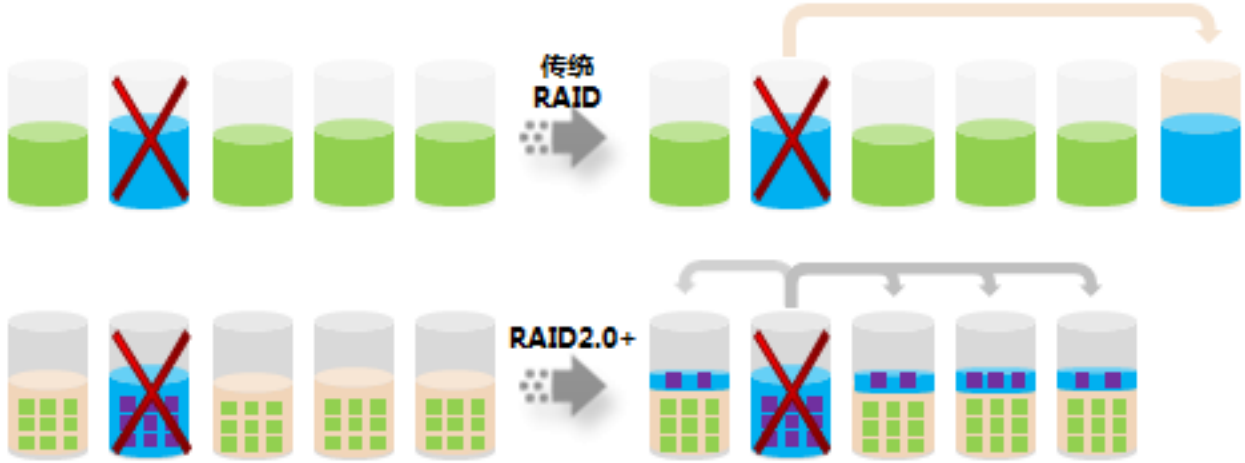
目　录

RAID 2.0+优势 — 自动负载均衡

- RAID 2.0+使得各硬盘均衡分担负载，不再有热点硬盘，提升了系统的性能和硬盘可靠性

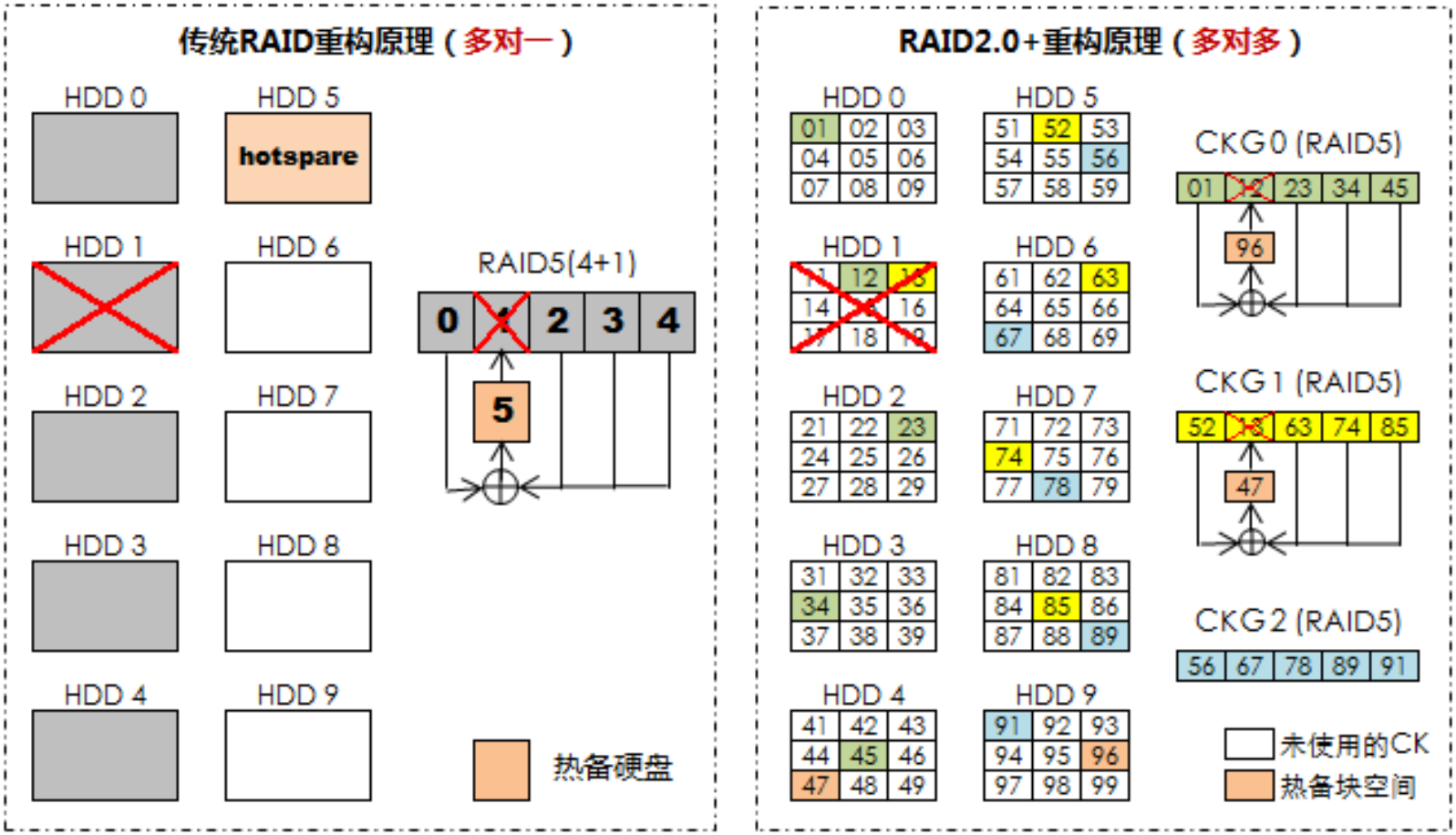


RAID 2.0+优势 — 系统可靠性高

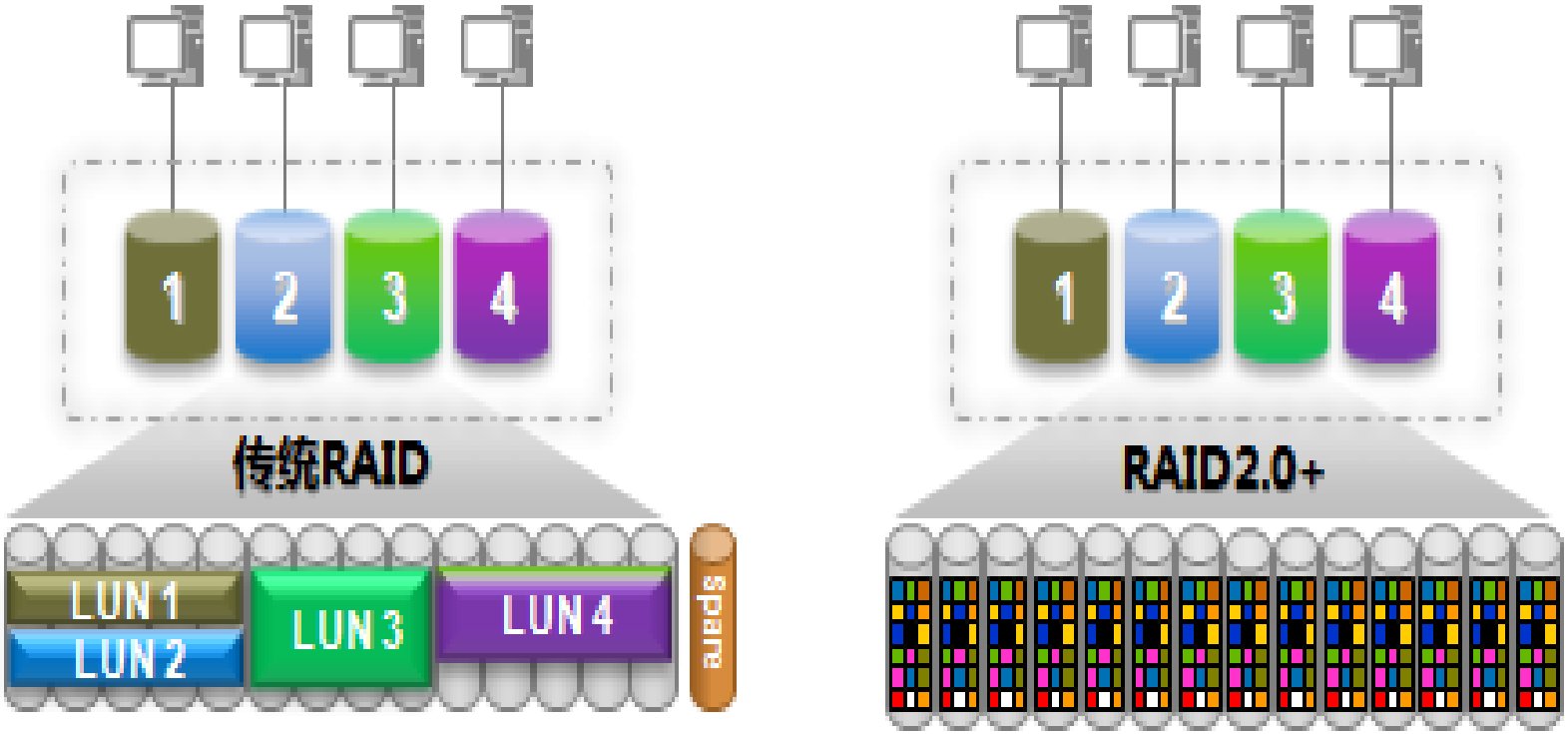


传统RAID	RAID2.0+
需要手动配置单独的全局或局部热备磁盘	分布式的热备空间，无需单独配置
多对一的重构，重构数据流串行写入单一的热备磁盘	多对多的重构，重构数据流并行写入多块磁盘
存在热点，重构时间长	负载均衡，重构时间短

RAID 2.0+优势 — 快速精简重构

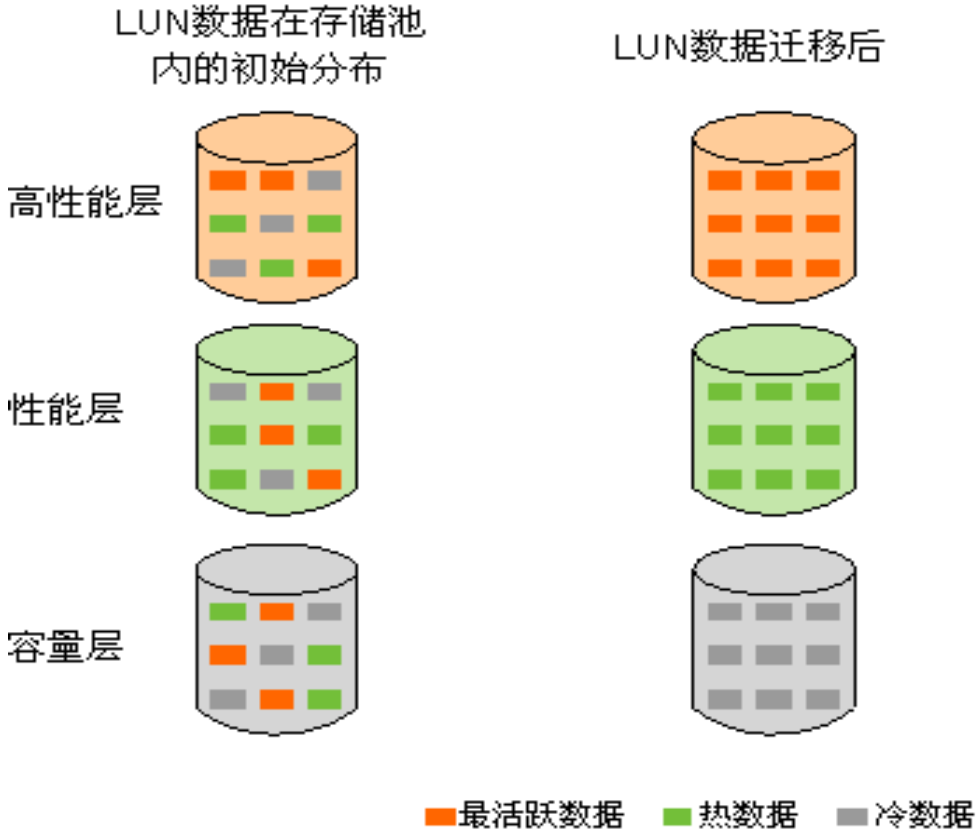


RAID 2.0+优势 — 提升单LUN性能

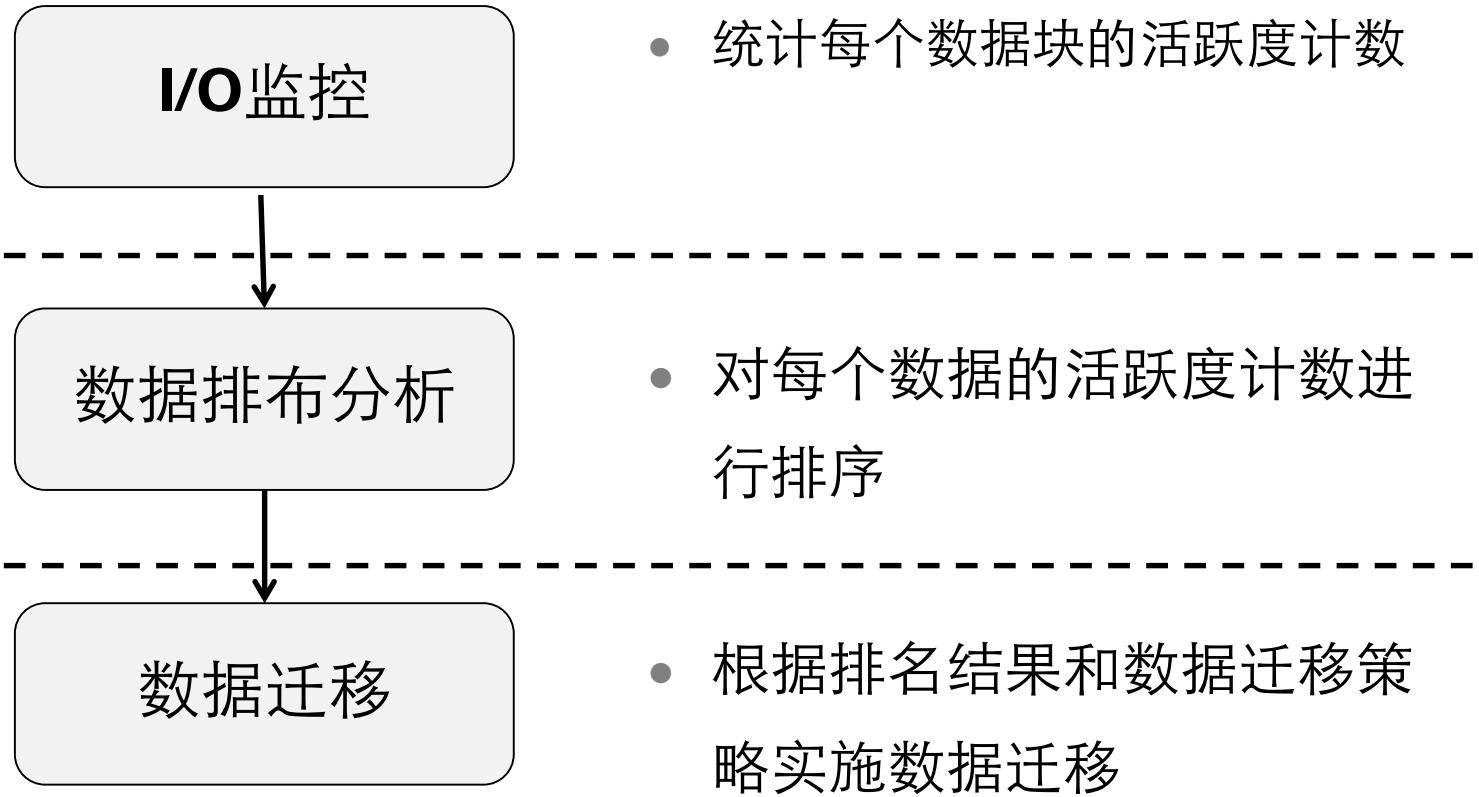


RAID 2.0+优势 — SmartTier

- **LUN**上的数据可以根据数据的活跃度，自动调整，迁移到存储池中的不同存储层。



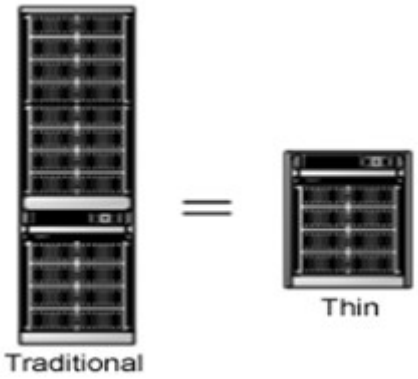
RAID 2.0+优势 — SmartTier（续）

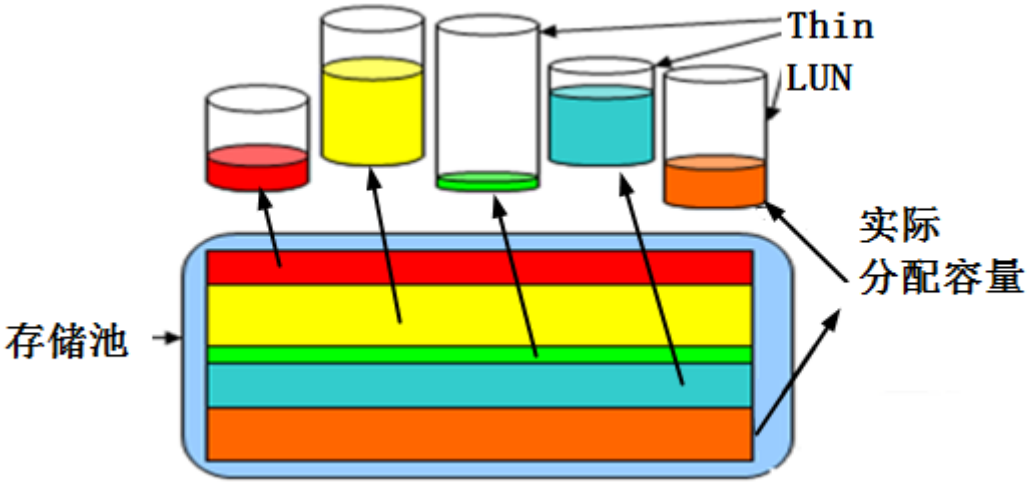


- 统计每个数据块的活跃度计数
- 对每个数据的活跃度计数进行排序
- 根据排名结果和数据迁移策略实施数据迁移

RAID 2.0+优势 — SmartThin

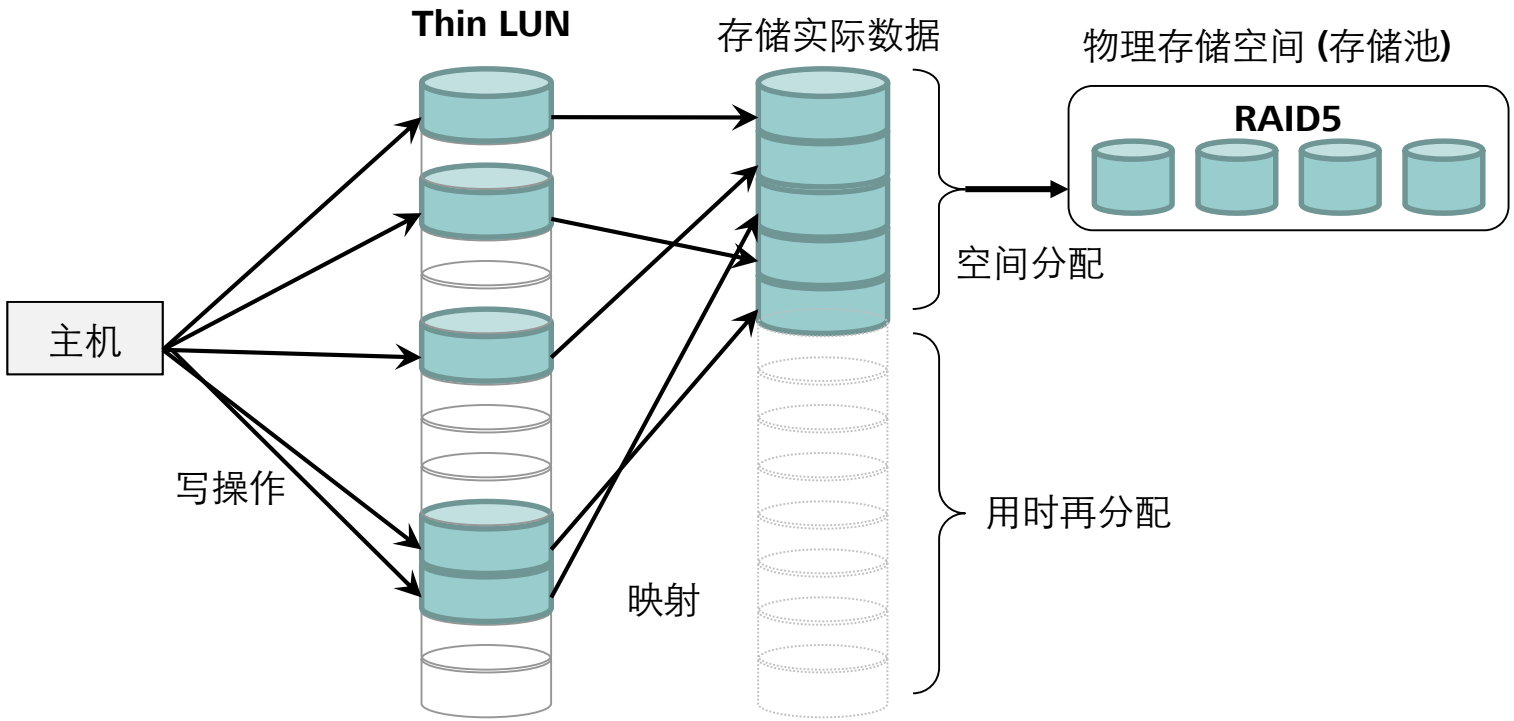
- **Smart Thin特点：**
  - 存储容量虚拟化
  - 按需分配
  - 可在线扩容
  - 容量管理自动化
  - 告警阈值





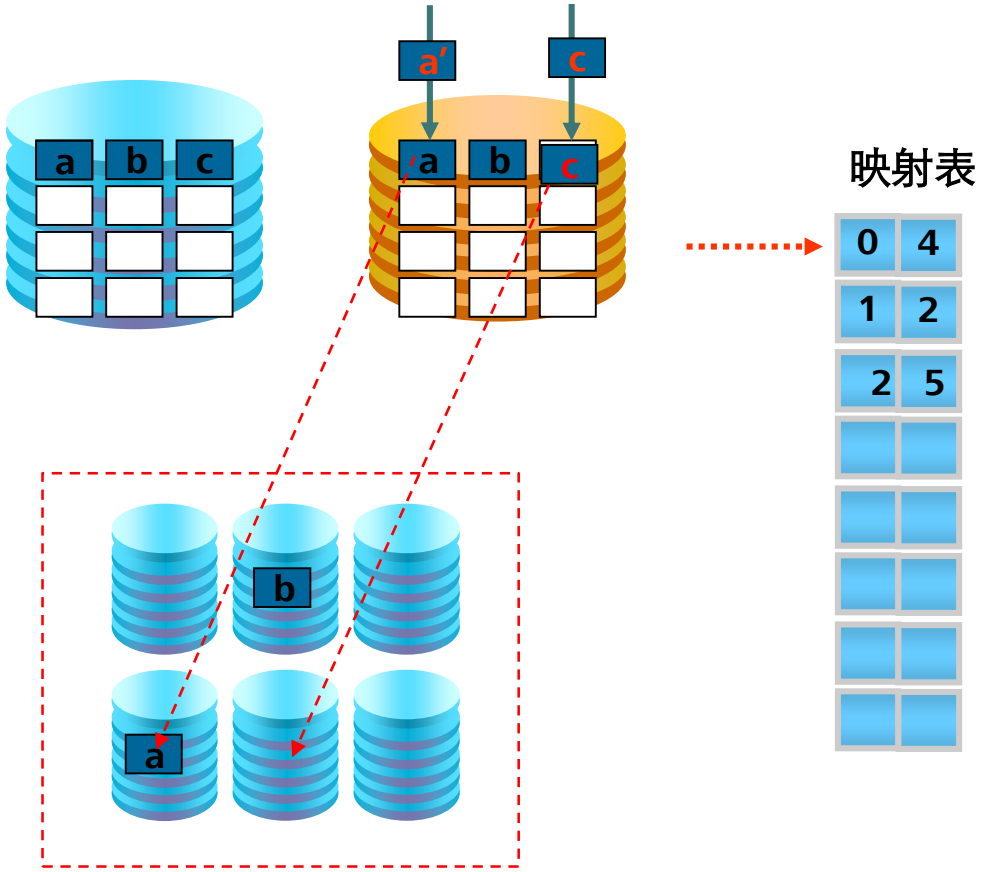
RAID 2.0+优势 — SmartThin（续）

- 写时空间分配： **Capacity-on-write**
- 读写重定向： **Redirect-on-time**



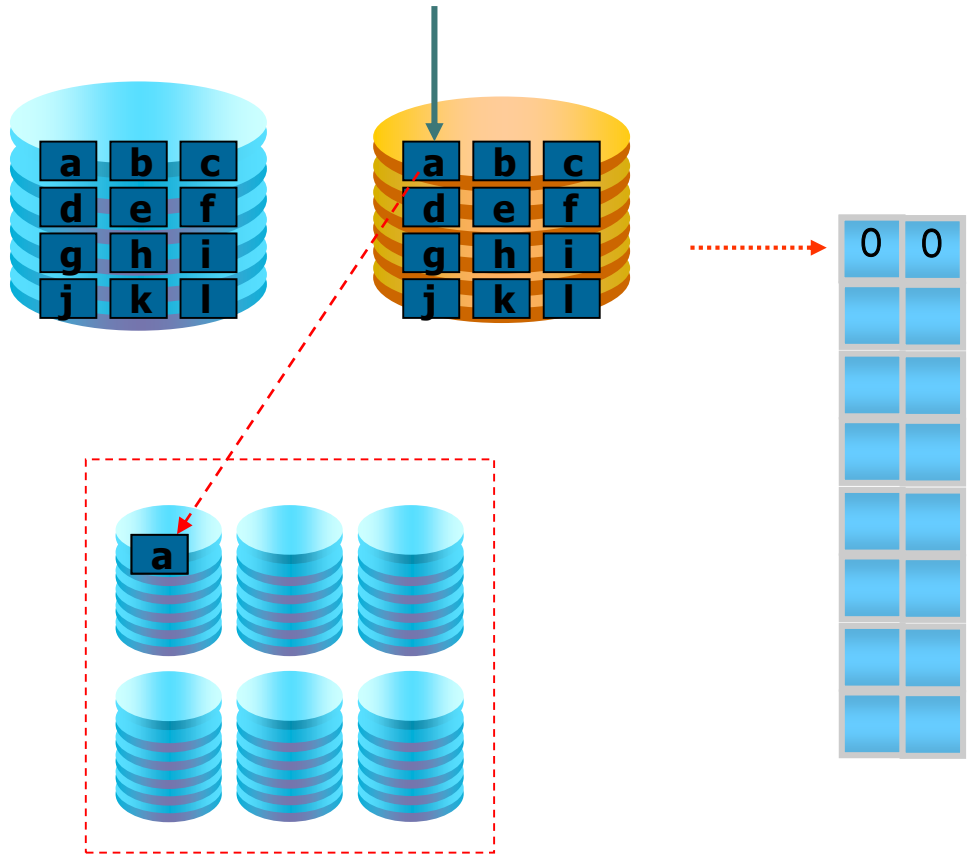
RAID 2.0+优势 — SmartThin（续）

1. 收到写请求
2. 查映射表
3. 写重定向
4. 写数据



RAID 2.0+优势 — SmartThin（续）

1. 收到读请求
2. 查映射表
3. 重定向请求
4. 读数据数据



RAID 2.0+优势 — SmartThin（续）

- 适用的场合

业务类型	分析	举例
对业务连续性要求较高的系统核心业务	在线对系统进行扩容，不会中断业务	银行票据交易系统
应用系统数据增长速度无法准确评估的业务	按需分配物理存储空间，避免浪费	E-mail邮箱服务、网盘服务等
多种业务系统混杂并且对容量需求不一的业务	让不同业务去竞争物理存储空间，实现物理存储空间的优化配置。	运营商服务等

- 不适用的场合

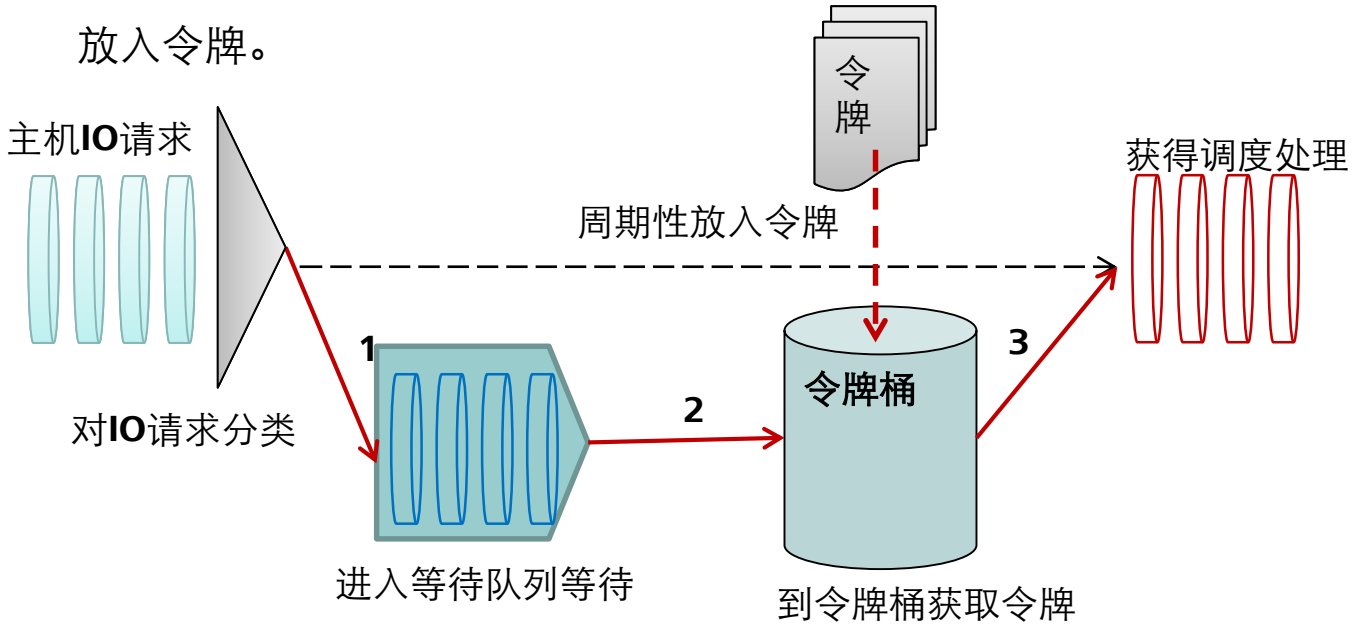
业务类型	分析	举例
对I/O性能要求很高的场合	读写重定向，对性能有一定影响	在线交易

RAID 2.0+优势 — Smart QoS

- 背景：
  - 不同应用程序之间由于业务模型和**I/O**特征不同相互影响，导致存储系统整体性能受到影响；
  - 不同应用程序相互争抢系统带宽和**IOPS**资源，关键业务性能无法得到保证。
- 需求：
  - 保证关键型应用程序的性能；
  - 保证高级别用户的性能。
- **Smart QoS**原理：
  - 该特性允许用户根据应用程序的一系列特征（**IOPS**或带宽），对每一种应用程序设置特定的性能目标；
  - 存储系统根据设定的性能目标，动态分配存储系统的资源来满足特定应用程序的服务级别要求，从而优先保证关键性应用程序服务级别的需求。

RAID 2.0+优势 — SmartQoS

- **SmartQoS基于令牌桶原理**
  - 用户每配置一个**SmartQoS**策略，系统会根据用户设置的性能目标生成一个令牌桶，按照用户配置的性能目标周期性向令牌桶中放入一定数量的令牌。
  - 每一个受这个**SmartQoS** 策略控制的**I/O**请求都必须从令牌桶中获得一个令牌才能得到处理；如果令牌桶中的令牌取空，则只能在等待队列中等待系统下一次放入令牌。



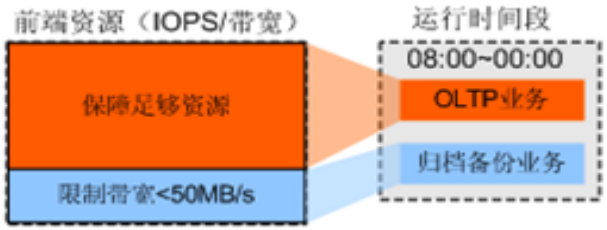
RAID 2.0+优势 — SmartQoS（续）

业务类型	I/O特征	主要运行时间段
在线交易业务	随机小I/O，通常以IOPS来衡量	08:00至00:00
归档备份业务	顺序大I/O，通常以带宽来衡量	00:00至08:00

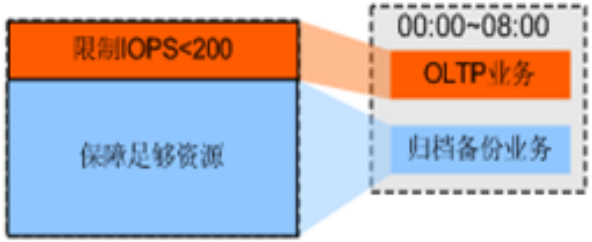
这两种业务在各自对应的时间段内都需要保障足够的系统资源。

- 创建两个Smart QoS策略：
  - 策略A：在08:00至00:00时间段内限制备份归档业务的带宽（例如<50MB/s）。
  - 策略B：在00:00至08:00时间段内限制在线交易业务的IOPS（例如<200）。

策略A



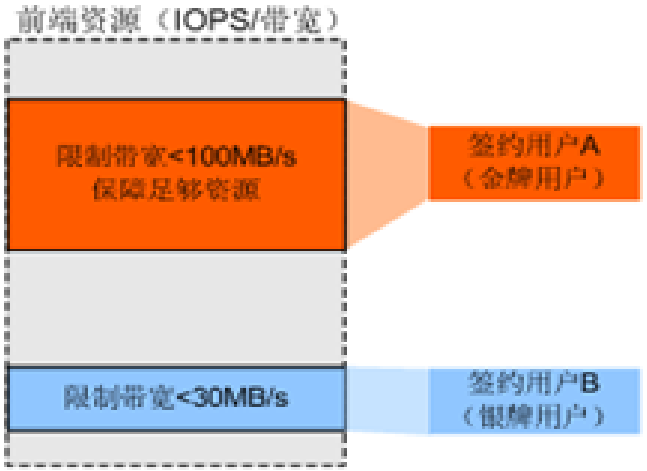
策略B



RAID 2.0+优势 — SmartQoS（续）

用户分类	业务质量要求
签约用户A（金牌用户）	高
签约用户B（银牌用户）	中

- 创建两个**Smart QoS**策略：
  - 策略**A**：限制金牌用户**A**的业务带宽（例如**<100MB/s**）。
  - 策略**B**：限制银牌用户**B**的业务带宽（例如**<30MB/s**）。



谢谢

www.huawei.com