MODULE – 3

DATA PROTECTION – RAID

School of software ,BUAA

前情回顾

- 磁盘总体服务时间由哪几部分构成，分别解释影响时间的因素有哪些？

Disk service time = seek time + rotational latency + data transfer time

前情回顾

- 什么是磁盘驱动器的合并操作？
  A. 将多个物理驱动器分组到逻辑驱动器
  B. 将物理驱动器分为多个逻辑驱动器
  C. 在逻辑驱动器上写入磁盘元数据的过程
  D. 通过碎片整理向物理驱动器添加更多容量

前情回顾

- 对于敏感型应用程序，磁盘空间限制利用率是多少？
  - A. 70%
  - B. 75%
  - C. 80%
  - D. 85%

Module 3: Data Protection – RAID

Upon completion of this module, you should be able to:

- Describe RAID implementation methods
- Describe the three RAID techniques
- Describe commonly used RAID levels
- Describe the impact of RAID on performance
- Compare RAID levels based on their cost, performance, and protection

Module 3: Data Protection – RAID

Lesson 1: RAID Overview

During this lesson the following topics are covered:

- RAID Implementation methods
- RAID array components
- RAID techniques

Why RAID? Redundant Array of Inexpensive/Independent Disks

RAID

It is a technique that combines multiple disk drives into a logical unit (RAID set) and provides protection, performance, or both.

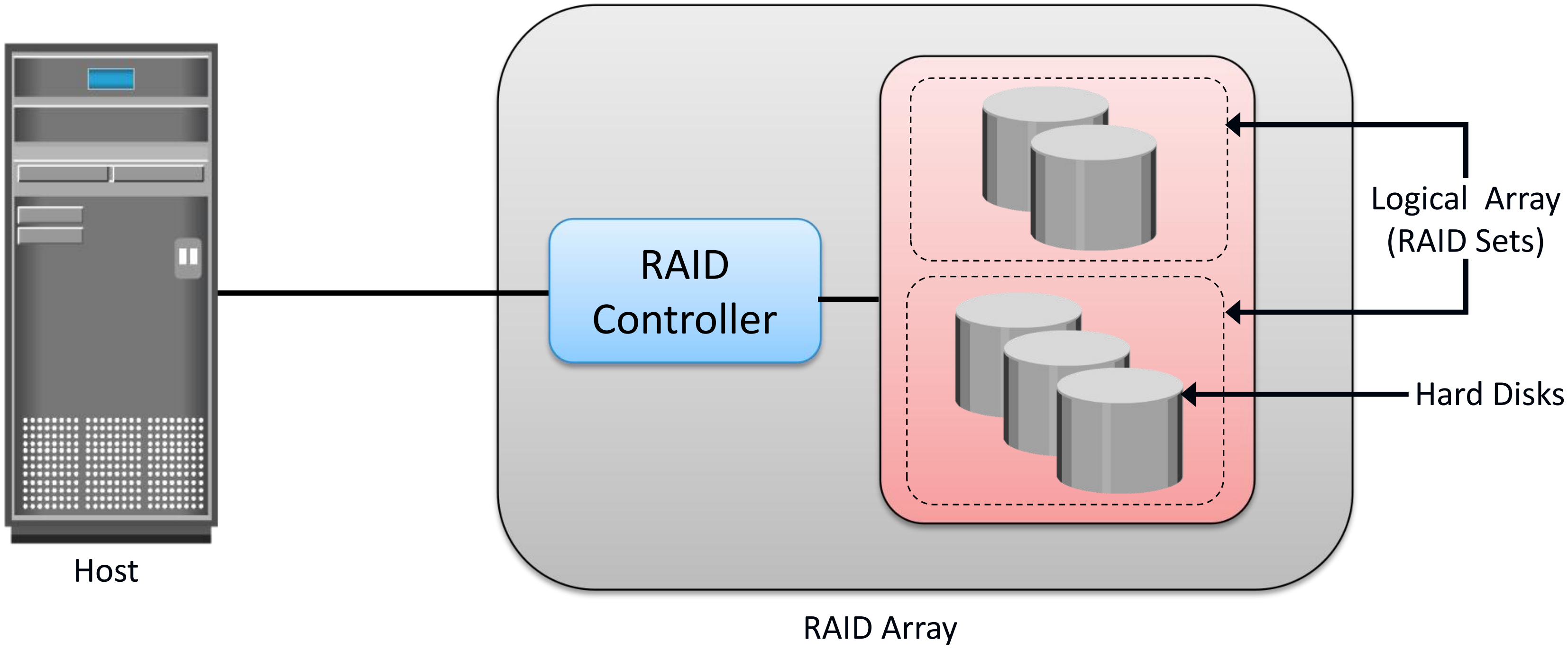
- Due to mechanical components in a disk drive it offers limited performance
- An individual drive has a certain life expectancy and is measured in **MTBF**（**Mean Time Between Failure**）：
  - For example: If the MTBF of a drive is 750,000 hours, and there are 1000 drives in the array, then the MTBF of the array is 750 hours (750,000/1000)
- RAID was introduced to mitigate these problems

Patterson,Gibson,Katz 《A Case for Redundant Arrays of Inexpensive Disks (RAID)》_University of California Berkeley,1987

RAID Implementation Methods

- Software RAID implementation
  - Uses host-based software to provide RAID functionality
  - Limitations
    - Use host CPU cycles to perform RAID calculations, hence impact overall system performance
    - Support limited RAID levels
    - RAID software and OS can be upgraded only if they are compatible
- Hardware RAID Implementation
  - Uses a specialized hardware controller installed either on a host or on an array

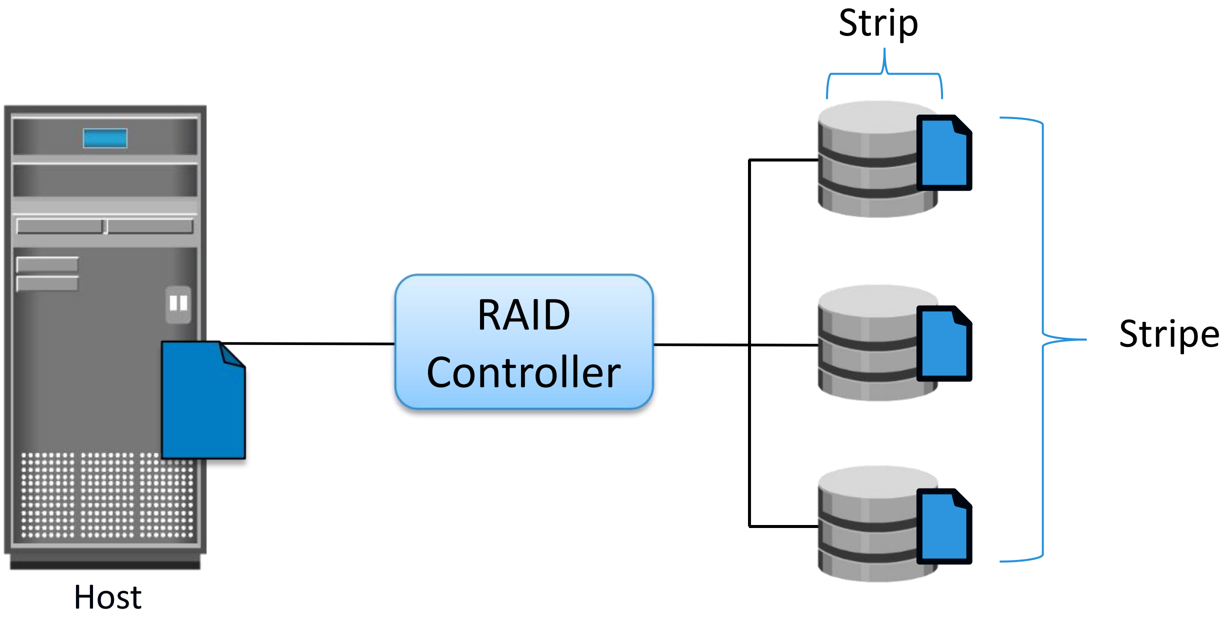
RAID Array Components



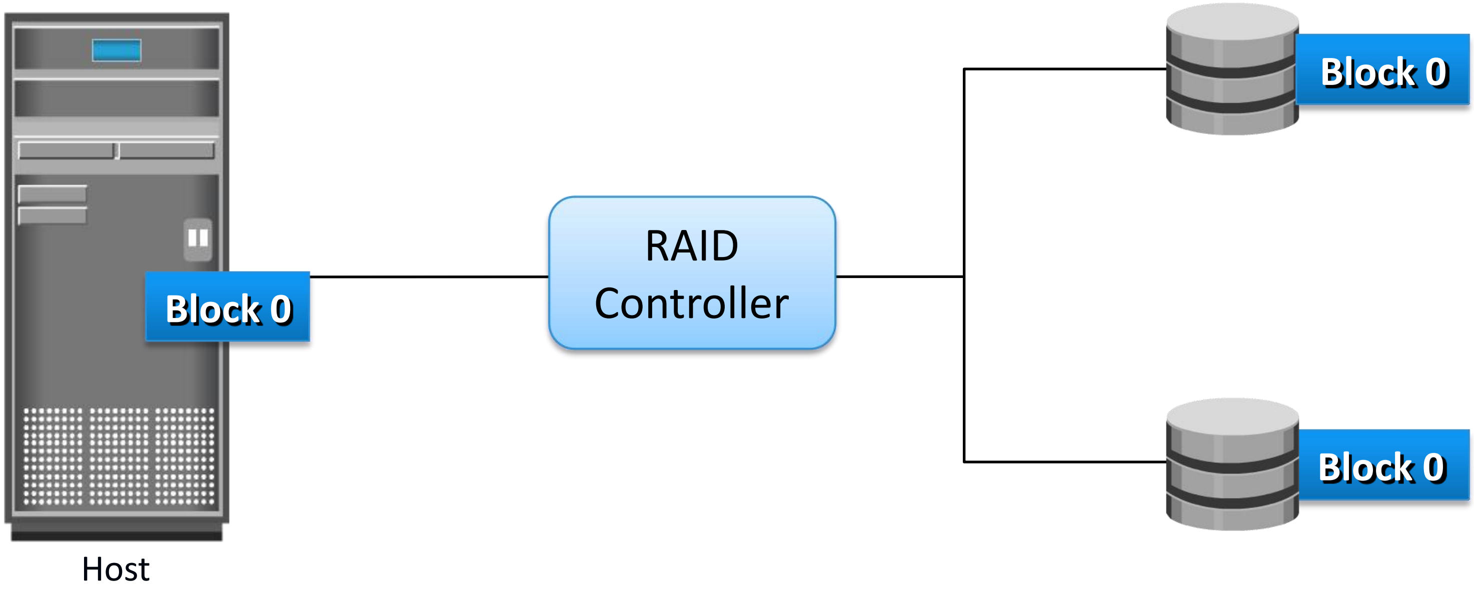
RAID Techniques

- Three key techniques used for RAID are:
  - Striping
  - Mirroring
  - Parity

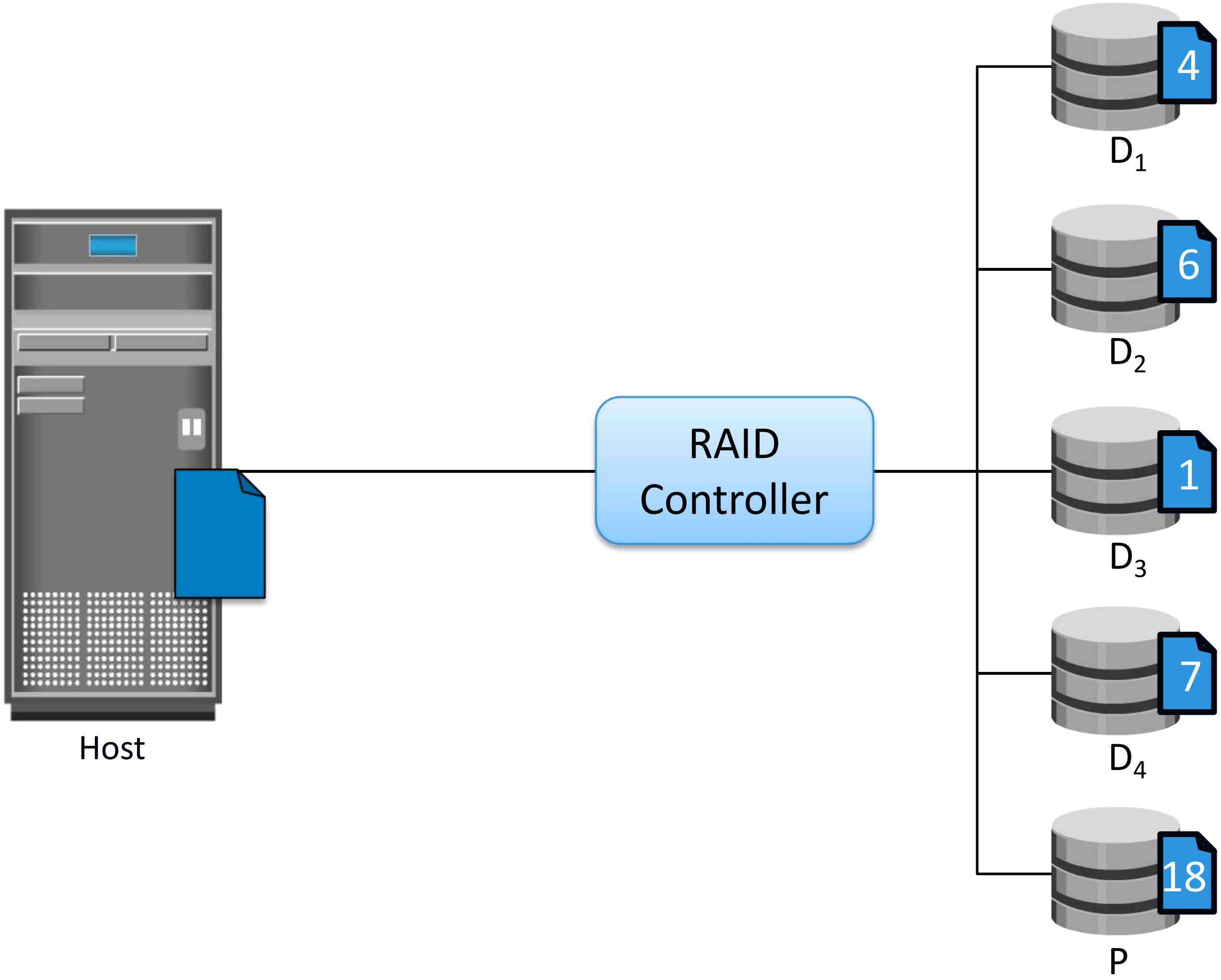
RAID Technique – Striping



RAID Technique – Mirroring

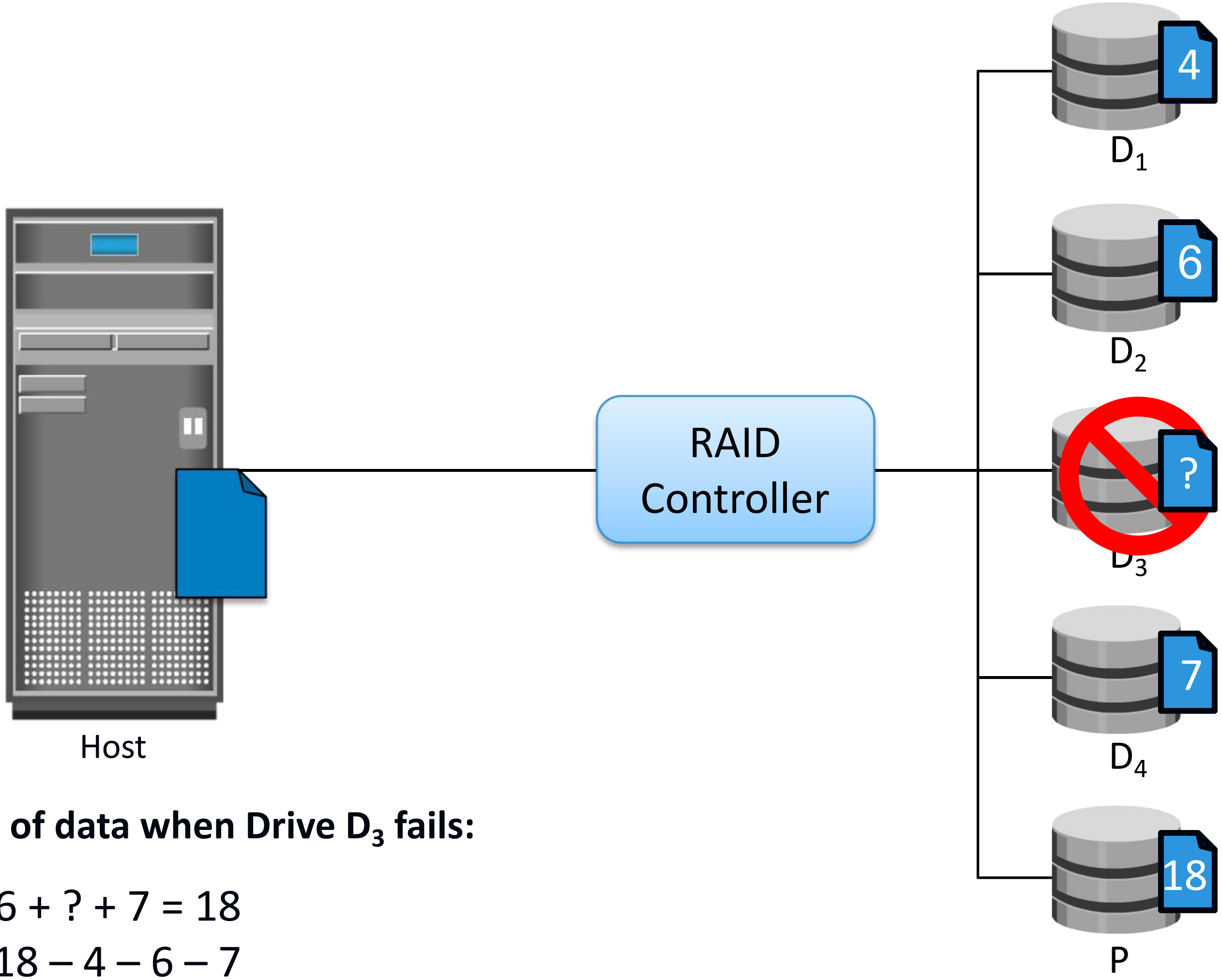


RAID Technique – Parity



Actual parity calculation is a bitwise XOR operation

Data Recovery in Parity Technique



Regeneration of data when Drive D$_3$ fails:

4 + 6 + ? + 7 = 18

? = 18 – 4 – 6 – 7

? = 1

Module 3: Data Protection – RAID

Lesson 2: RAID Levels

During this lesson the following topics are covered:

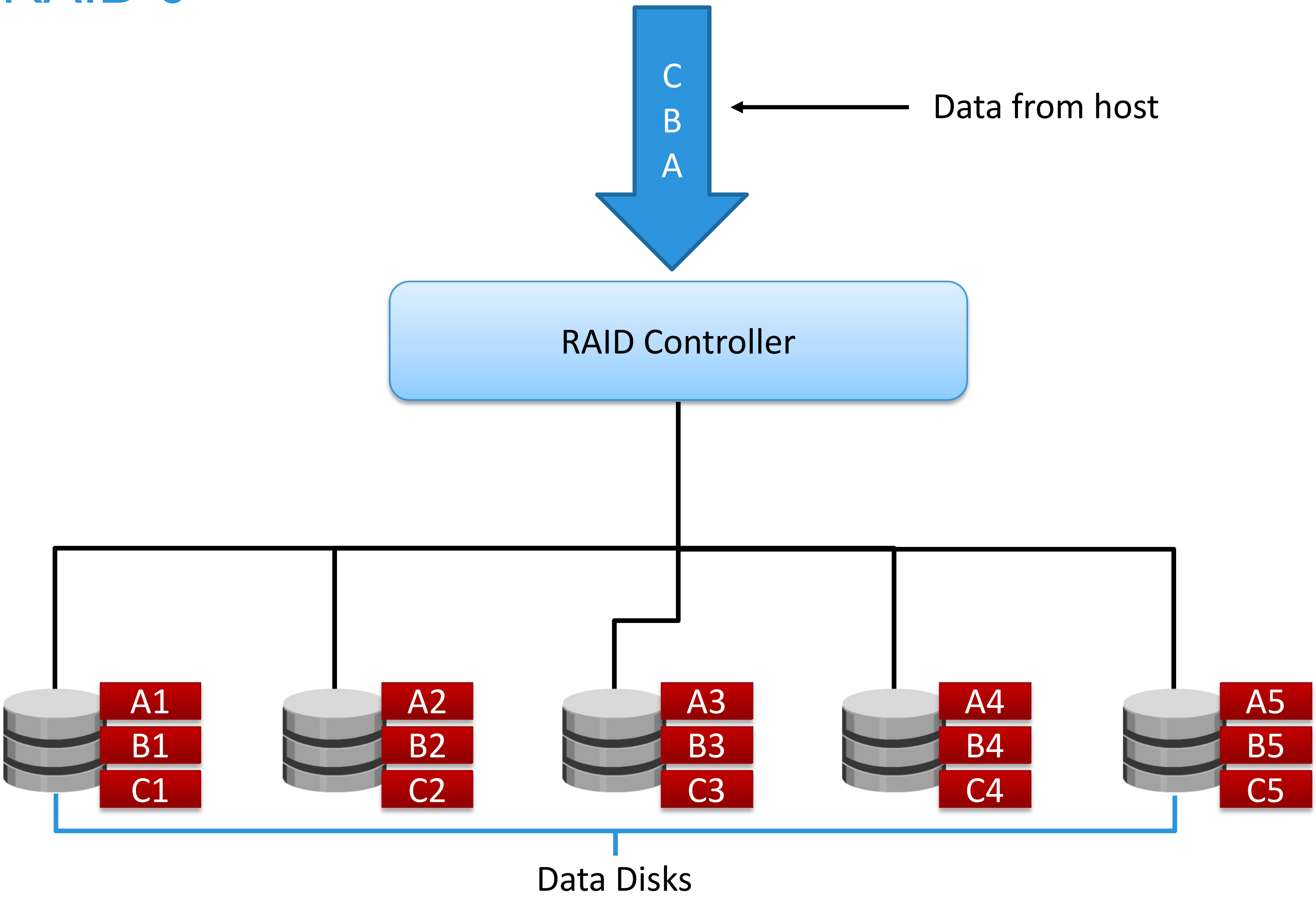
- Commonly used RAID levels
- RAID impacts on performance
- RAID comparison
- Hot spare

RAID Levels

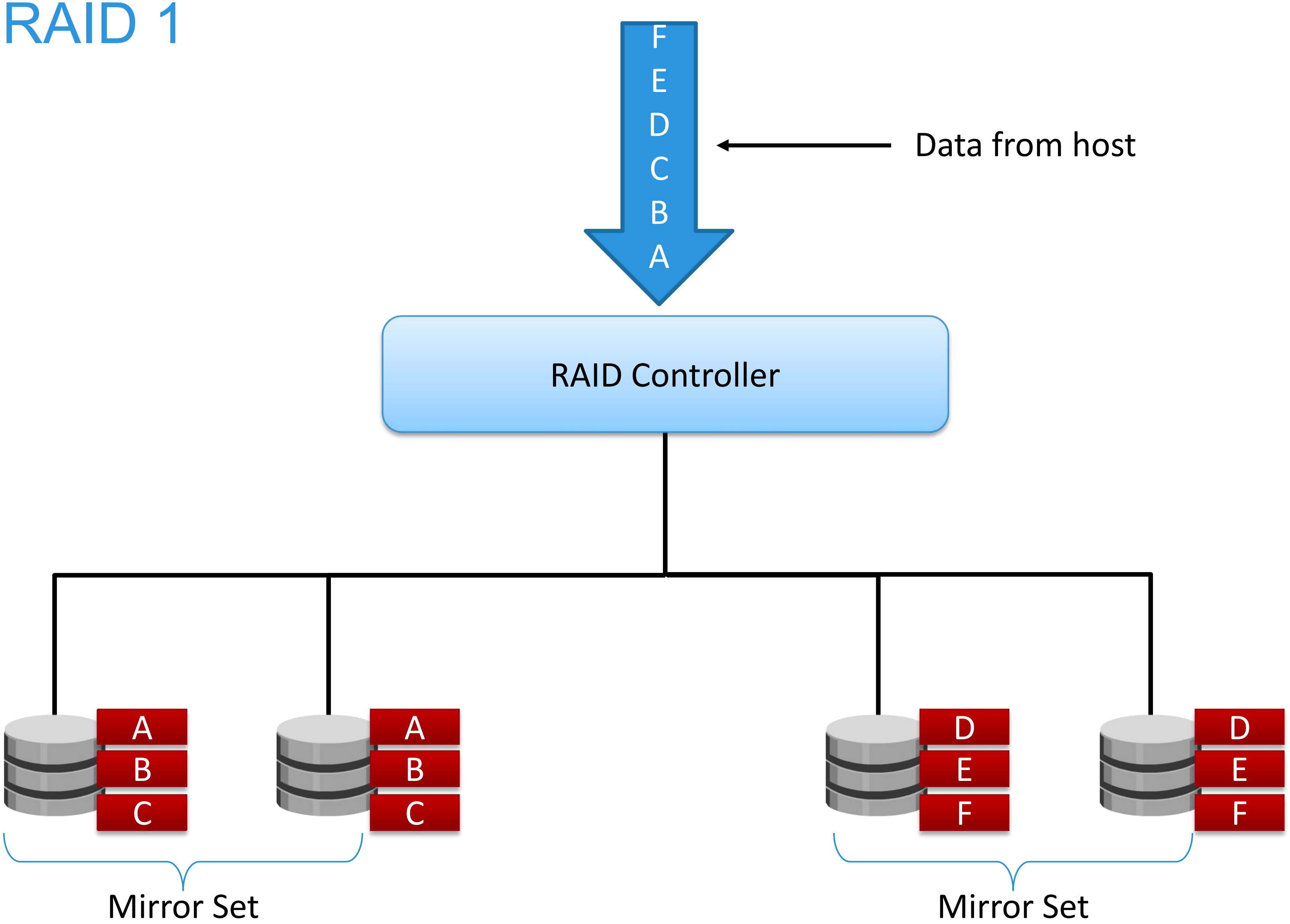
- Commonly used RAID levels are:
  - RAID 0 – Striped set with no fault tolerance
  - RAID 1 – Disk mirroring
  - RAID 1 + 0 – Nested RAID
  - RAID 3 – Striped set with parallel access and dedicated parity disk
  - RAID 5 – Striped set with independent disk access and a distributed parity
  - RAID 6 – Striped set with independent disk access and dual distributed parity

RAID 2
RAID 4
X ?

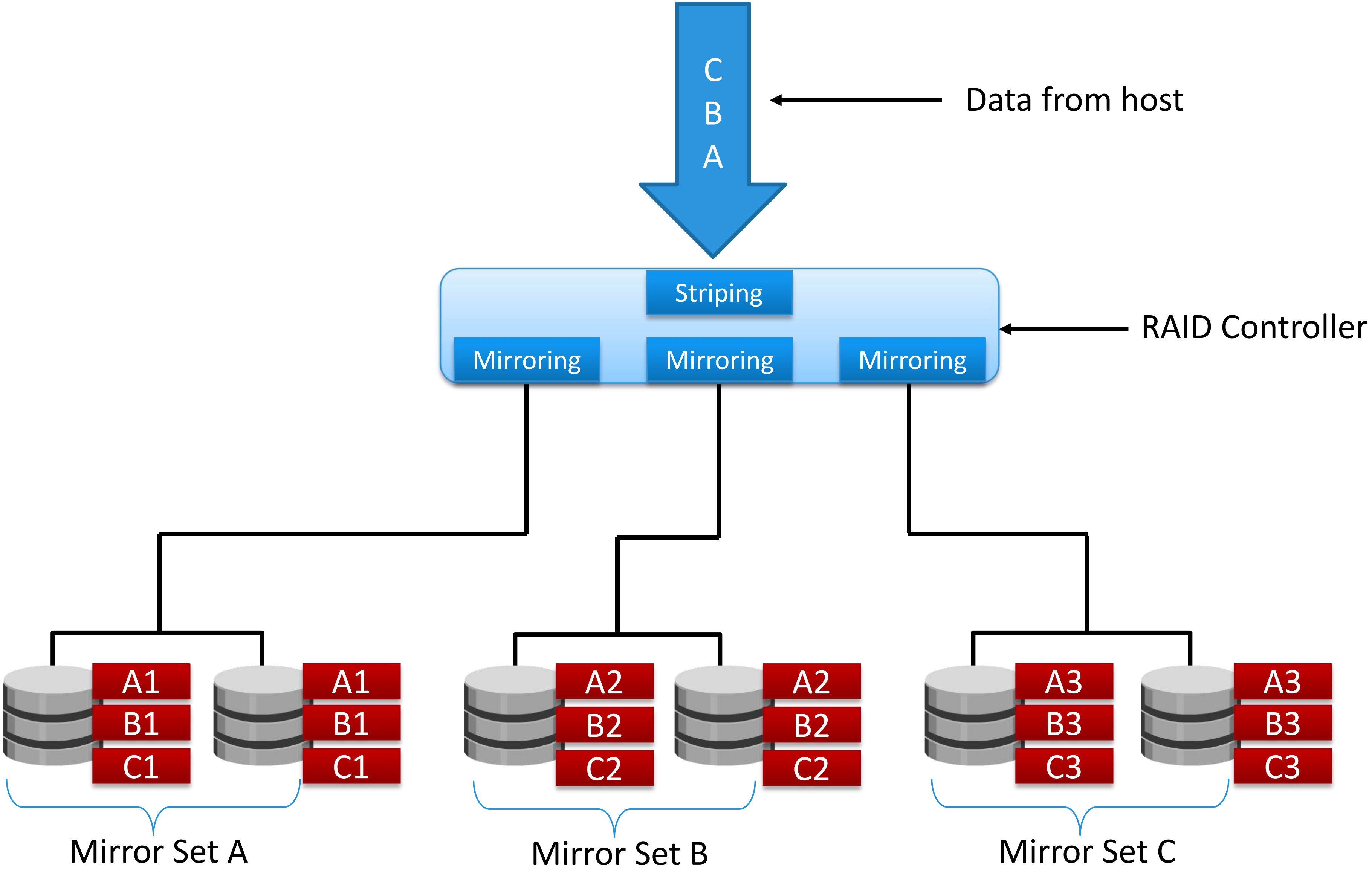
RAID 0


C
B
A
Data from host
RAID Controller
A1
B1
C1
A2
B2
C2
A3
B3
C3
A4
B4
C4
A5
B5
C5
Data Disks

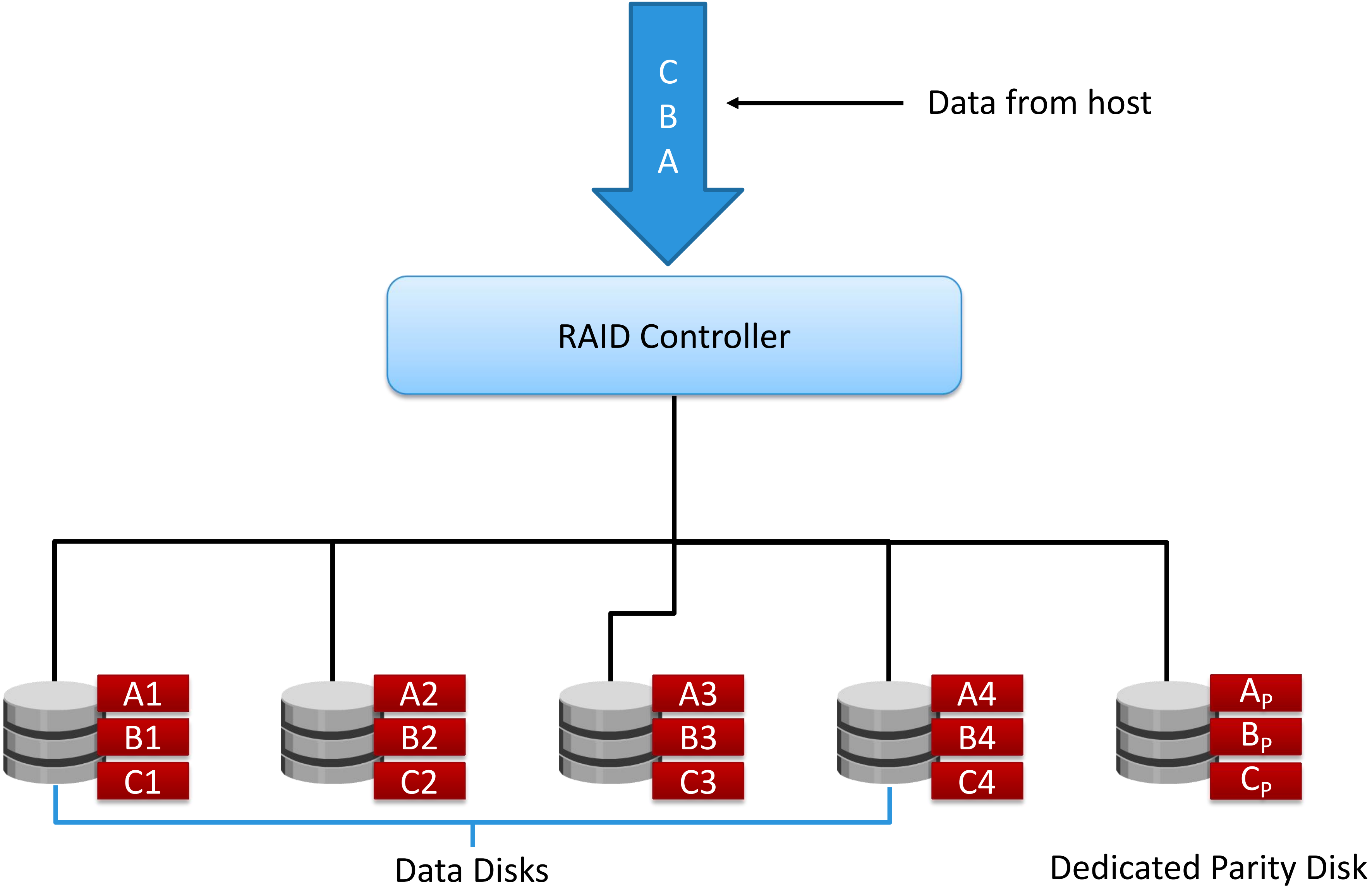
RAID 1


F
E
D
C
B
A
Data from host
RAID Controller
A
B
C
A
B
C
D
E
F
D
E
F
Mirror Set
Mirror Set

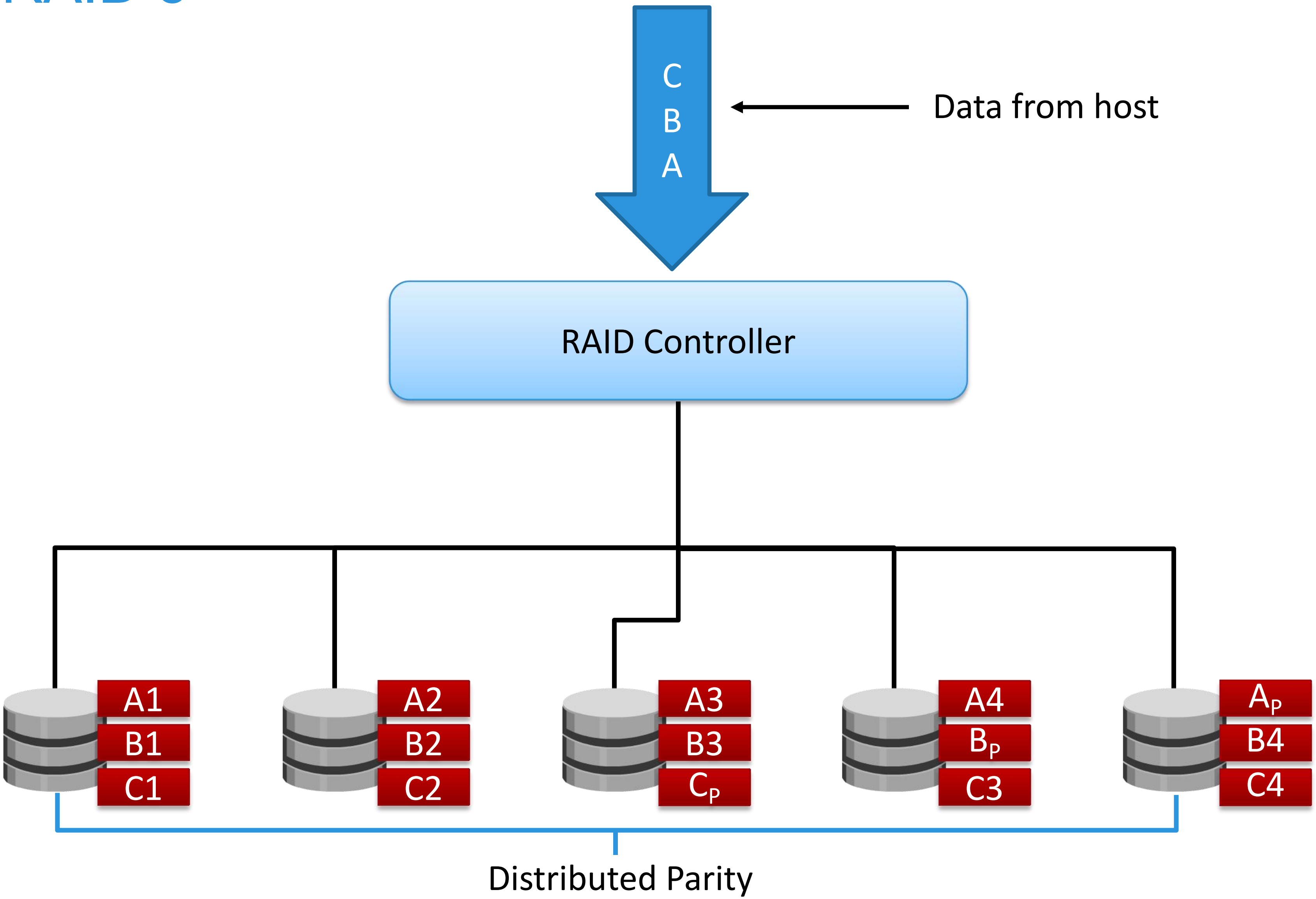
Nested RAID – 1+0



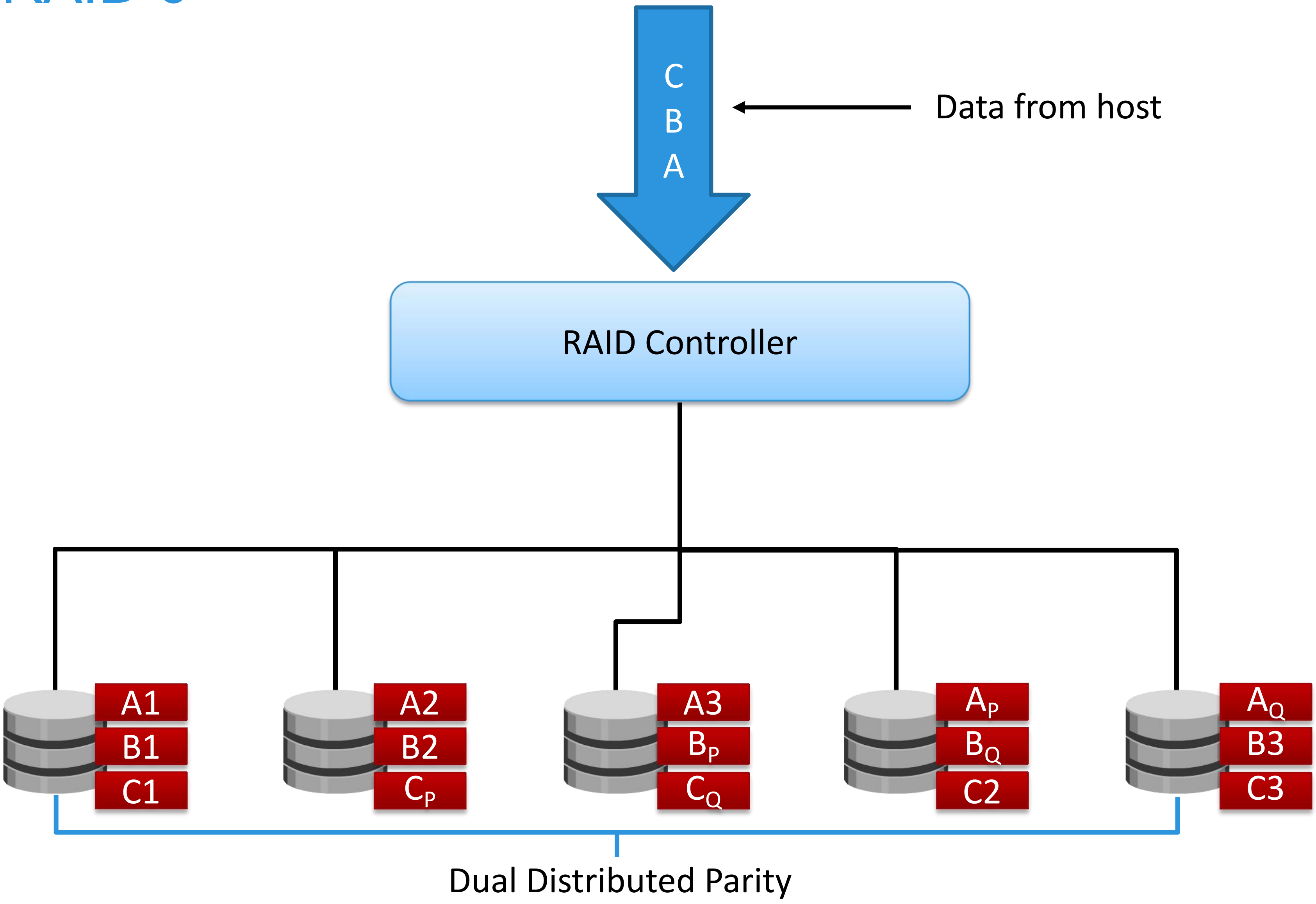
RAID 3


C
B
A
Data from host
RAID Controller
A1
B1
C1
A2
B2
C2
A3
B3
C3
A4
B4
C4
A_P
B_P
C_P
Data Disks
Dedicated Parity Disk

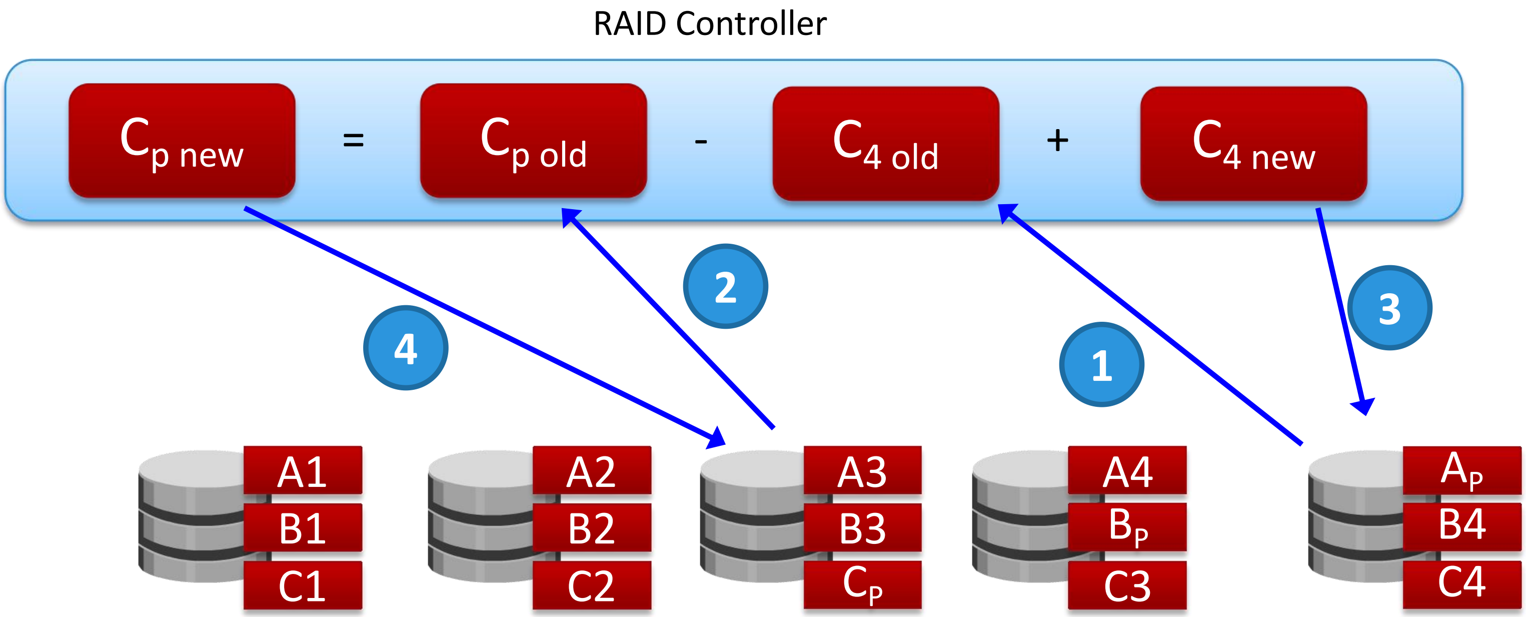
RAID 5


C
B
A
Data from host
RAID Controller
A1
B1
C1
A2
B2
C2
A3
B3
C_P
A4
B_P
C3
A_P
B4
C4
Distributed Parity

RAID 6


C
B
A
Data from host
RAID Controller
A1
B1
C1
A2
B2
C_P
A3
B_P
C_Q
A_P
B_Q
C2
A_Q
B3
C3
Dual Distributed Parity

RAID Impacts on Performance



- In RAID 5, every write (update) to a disk manifests as four I/O operations (2 disk reads and 2 disk writes)
- In RAID 6, every write (update) to a disk manifests as six I/O operations (3 disk reads and 3 disk writes)
- In RAID 1, every write manifests as two I/O operations (2 disk writes)

RAID Penalty Calculation Example

- Total IOPS(Input/Output Per Second) at peak workload is 1200
- Read/Write ratio 2:1
- Calculate disk load at peak activity for:
  - RAID 1/0
  - RAID 5

Solution: RAID Penalty

- For RAID 1/0, the disk load (read + write)

  = (1200 x 2/3) + (1200 x (1/3) x 2)

  = 800 + 800

  = 1600 IOPS

- For RAID 5, the disk load (read + write)

  = (1200 x 2/3) + (1200 x (1/3) x 4)

  = 800 + 1600

  = 2400 IOPS

RAID Comparison

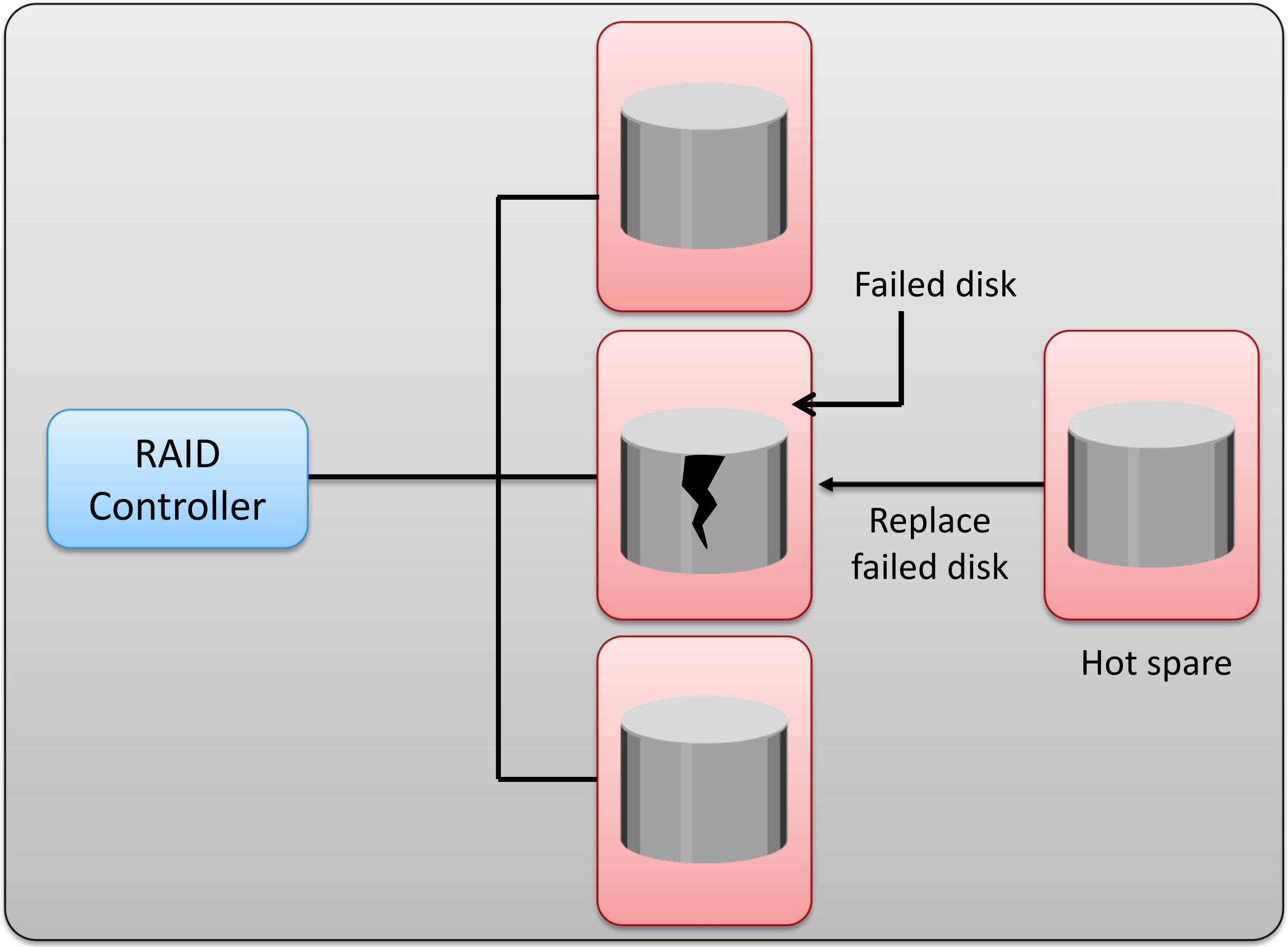
RAID level	Min disks	Available storage capacity (%)	Read performance	Write performance	Write penalty	Protection
1	2	50	Better than single disk	Slower than single disk, because every write must be committed to all disks	Moderate	Mirror
1+0	4	50	Good	Good	Moderate	Mirror
3	3	[(n-1)/n]*100	Fair for random reads and good for sequential reads	Poor to fair for small random writes fair for large, sequential writes	High	Parity (Supports single disk failure)
5	3	[(n-1)/n]*100	Good for random and sequential reads	Fair for random and sequential writes	High	Parity (Supports single disk failure)
6	4	[(n-2)/n]*100	Good for random and sequential reads	Poor to fair for random and sequential writes	Very High	Parity (Supports two disk failures)

where n = number of disks

Suitable RAID Levels for Different Applications

- RAID 1+0
  - Suitable for applications with small, random, and write intensive (writes typically greater than 30%) I/O profile
  - Example: OLTP, RDBMS – Temp space
- RAID 3
  - Large, sequential read and write
  - Example: data backup and multimedia streaming
- RAID 5 and 6
  - Small, random workload (writes typically less than 30%)
  - Example: email, RDBMS – Data entry

Hot Spare



Lab1：RAID

Key points :

- 简单卷（分区？）
- 跨区卷（JBOD）
- 带区卷（RAID 0）
- 镜像卷 (RAID 1)
- RAID 5

Module 3: Summary

Key points covered in this module:

- RAID implementation methods and techniques
- Common RAID levels
- RAID write penalty
- Compare RAID levels based on their cost and performance

Exercise 1: RAID

- A company is planning to reconfigure storage for their accounting application for high availability
  - Current configuration and challenges
    - Application performs 15% random writes and 85% random reads
    - Currently deployed with five disk RAID 0 configuration
    - Each disk has an advertised formatted capacity of 200 GB
    - Total size of accounting application’s data is 730 GB which is unlikely to change over 6 months
    - Approaching end of financial year, buying even one disk is not possible
- Task
  - Recommend a RAID level that the company can use to restructure their environment fulfilling their needs
  - Justify your choice based on cost, performance, and availability

Exercise 2: RAID

- A company (same as discussed in exercise 1) is now planning to reconfigure storage for their database application for HA
  - Current configuration and challenges
    - The application performs 40% writes and 60% reads
    - Currently deployed on six disk RAID 0 configuration with advertised capacity of each disk being 200 GB
    - Size of the database is 900 GB and amount of data is likely to change by 30% over the next 6 months
    - It is a new financial year and the company has an increased budget
- Task
  - Recommend a suitable RAID level to fulfill company's needs
  - Estimate the cost of the new solution (200GB disk costs $1000)
  - Justify your choice based on cost, performance, and availability

知识测验 – 1

- 关于软件 RAID 实现，以下哪项描述是正确的？

  A. 操作系统升级不需要验证与 RAID 软件的兼容性

  B. 其成本高于硬件 RAID 实现

  C. 支持所有 RAID 级别

  D. 使用主机 CPU 周期执行 RAID 计算 ♥

- 一个应用程序生成 400 个小型随机 IOPS，读写比为 3:1。用于 RAID 5 的磁盘上 RAID 更正的 IOPS 是多少？

  A. 400

  B. 500

  C. 700 ♥

  D. 900

知识测验 – 2

- 用于小型随机 I/O 的 RAID 6 配置中的写性能损失是多少？
  A. 2
  B. 3
  C. 4
  D. 6 ♥
- 以下哪个应用程序可通过使用 RAID 3 获得最大效益？
  A. 备份 ♥
  B. OLTP
  C. 电子商务
  D. 电子邮件

知识测验 – 3

- 一个具有 64 KB 条块大小且包含五个磁盘的奇偶校验 RAID 5 集的条带大小是多少？
  - A. 64 KB
  - B. 128 KB
  - [✓ C. 256 KB]
  - D. 320 KB
- 假如有3块73G SAS磁盘，2块146G磁盘组成RAID5阵列最后逻辑磁盘的总容量是多少？
  - [✓ A. 292]
  - B. 365
  - C. 511
  - D. 438

作业

Scenario:

一个业务场景，实际IOPS是4800，读cache命中率是30%，读写比：3：2；磁盘个数为60，计算采用RAID5与RAID10磁盘的IOPS，分析那种方案更合适该场景。