Arquitetura de Computadores

MIEI – 2021/22

DI-FCT/UNL

Discos RAID

Sumário

- Discos RAID

Bibliografia:

Remzi H. Arpaci-Dusseau and Andrea C. Arpaci-Dusseau. Operating Systems: Three Easy Pieces. Cap 38

(https://pages.cs.wisc.edu/~remzi/OSTEP/file-devices.pdf).

Estruturas RAID

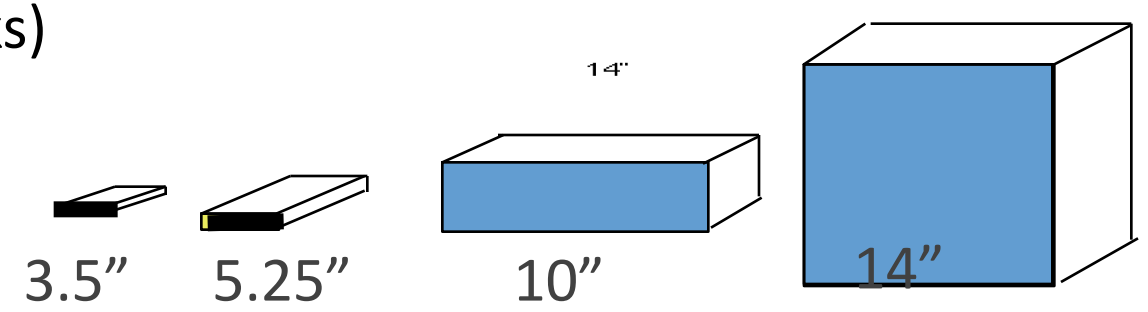
- **RAID** – múltiplas unidades de disco suportam elevada **fiabilidade** através de **redundância**.
- Inicialmente foram definidos 6 níveis RAID
  - RAID 0 : só assegura aumento da velocidade de acesso, porque permite acessos em paralelo
  - RAID 1, 2, 3, 4 e 5: permitem tolerância a falhas, porque há discos extra que asseguram redundância
- Outros níveis definidos mais recentemente
  - 6, 10 (ou 1+0) …
- A maior parte das vezes é um sistema complexo (exterior à caixa do computador) conhecido por disk array
  - Com um Sistema Operativo dedicado; memória não-volátil,…

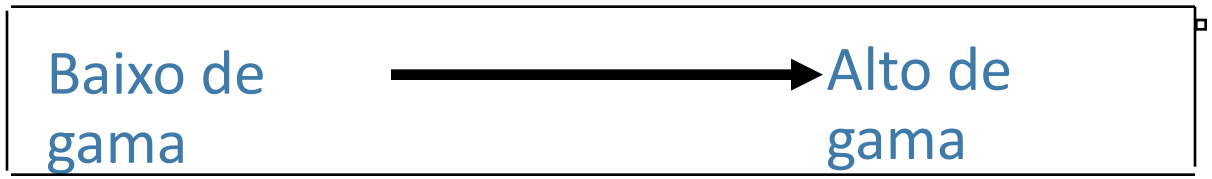
Proposta inicial (Patterson 1988)

SLED (Single Large Expensive Disk) VS

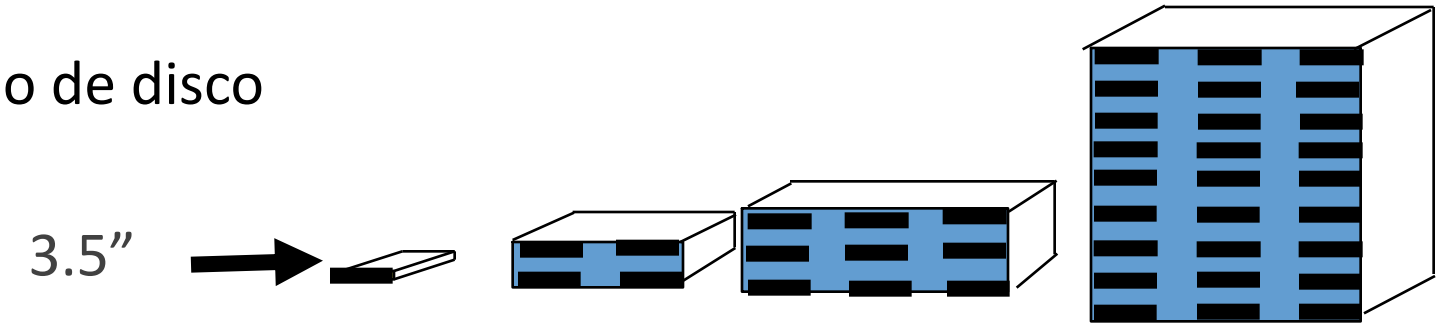
RAID (Redundant Array of Inexpensive Disks)

Convencional: 4 tipos de discos





Disk Array: 1 só tipo de disco



Proposta inicial (Patterson 1988)

- Substituir um grande disco por muitos pequenos discos!

- Disk Arrays têm potencial para:
  - Grandes taxas de transferência
  - Maior capacidade por unidade de volume
  - Maior capacidade por KW
  - Fiabilidade?

Fiabilidade do Array

- Fiabilidade de N discos = Fiabilidade de 1 Disco ÷ N

  50,000 Horas ÷ 70 discos = 700 horas

- MTTF (mean time to failure) do sistema de discos desce de 6 anos para 1 mês!
- Arrays (sem redundância) demasiado pouco fiáveis para terem utilidade!

Suporte de “hot swap” com reconstrução em paralelo com a operação normal permite ter uma disponibilidade extremamente elevada

RAID = Redundant Arrays of Independent Disks

- Os blocos são distribuídos ("striped“) por vários discos
- Redundância garante alta disponibilidade dos dados
- Em caso de falha de um disco, o conteúdo é reconstruído a partir de dados redundantes armazenados no array
  - Existe uma penalização na capacidade de armazenamento
  - Existe uma penalização em bandwidth para actualização

Técnicas:
- Mirroring/Shadowing (elevado custo em capacidade)
- Códigos de Correção de Erros (paridade, outros)

RAID

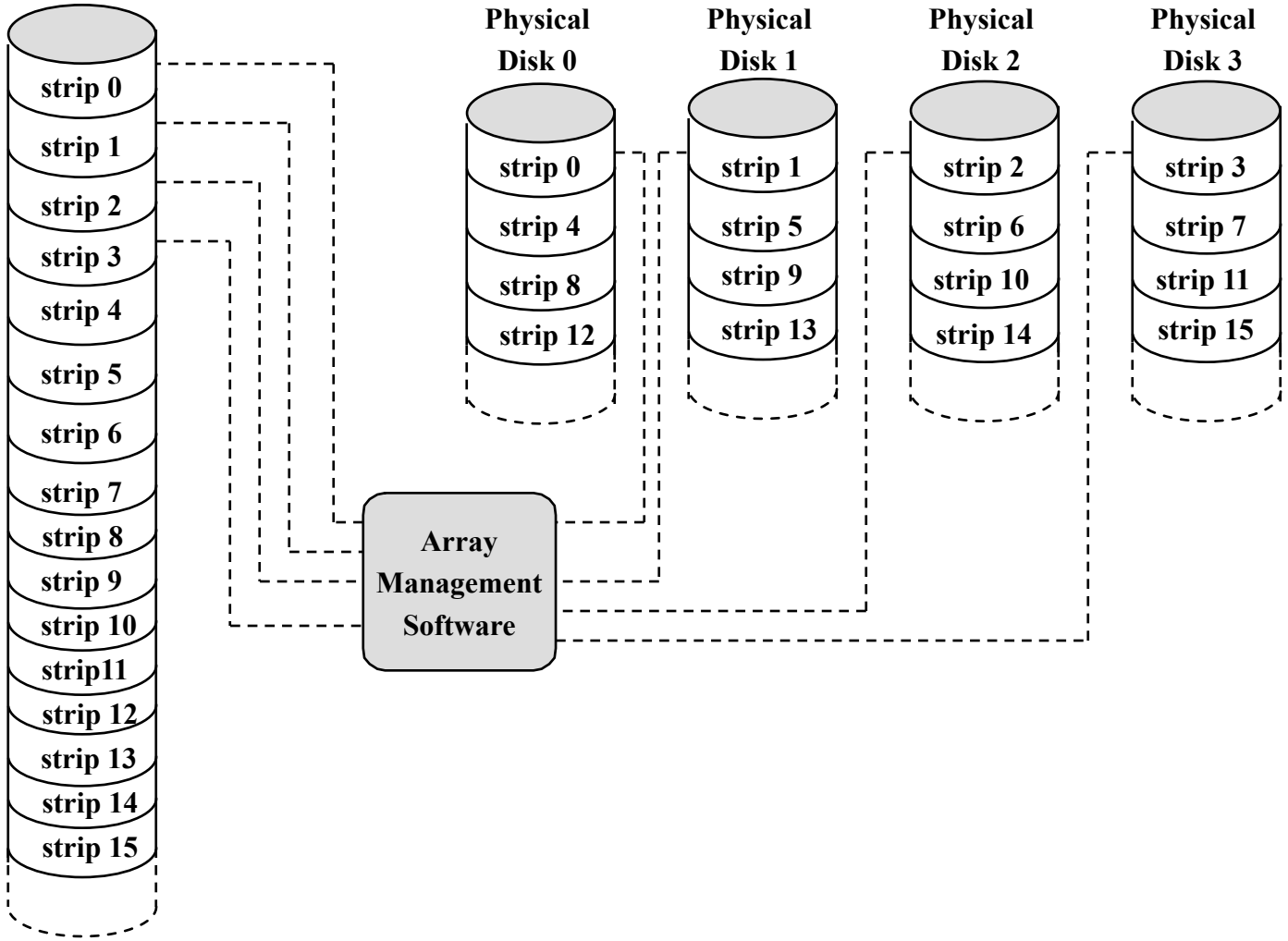
- Duas técnicas usadas:
  - “Disk striping” usa um grupo de discos como uma unidade lógica
    - diferentes partes dos dados são armazenados em discos diferentes
  - Aumento da velocidade de acesso e da fiabilidade através do armazenamento de dados redundantes.
    - Mirroring ou shadowing duplica discos inteiros.
    - Block interleaved parity usa muito menos redundância.

RAID Nível 0 (não redundante)

Maximiza o desempenho no acesso porque

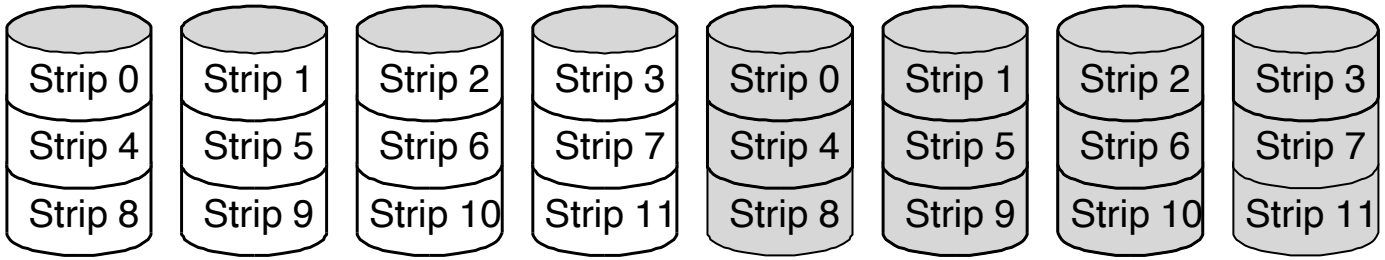
permite ler e escrever blocos em simultâneo

vários discos



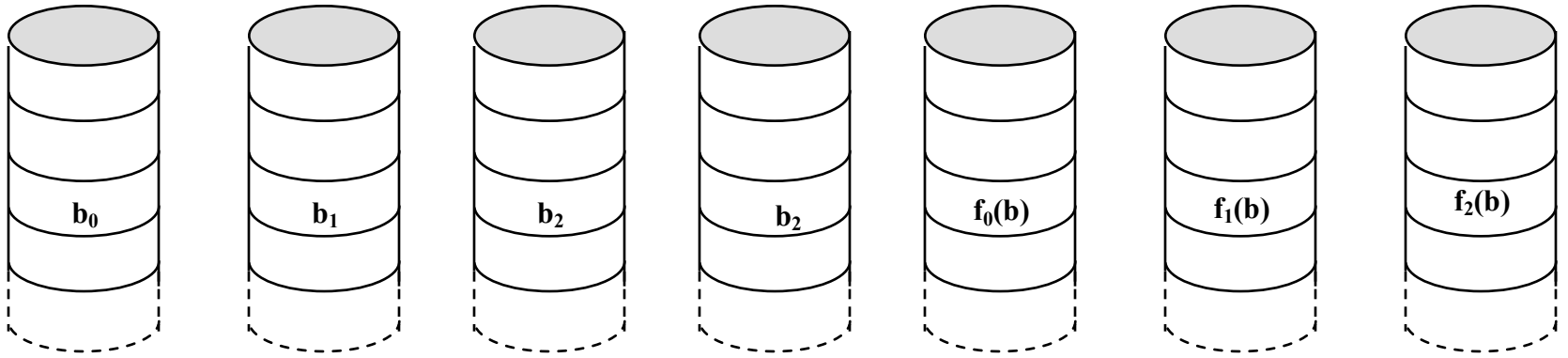
RAID 1 (Disk Mirroring/Shadowing)

- Cada disco é completamente duplicado no seu "shadow”
  - Pode ser atingida uma disponibilidade muito elevada
- Sacrifício da largura de banda em escrita:
  - uma escrita lógica → duas escritas físicas
- As leituras podem ser otimizadas
- A solução mais cara: 100% de custos extra em capacidade
- Usado em ambientes em que interessa alta disponibilidade



RAID 2 e 3

- Códigos de correção de erros ao nível do bit
  - RAID 2 – Log2 discos extra
  - RAID 3 – 1 disco extra
- Praticamente, não são usados



Exemplo de correção de erros

- Paridade
  - 1 disco extra
  - O conteúdo do disco extra é obtido fazendo P(b) = b_0 XOR b_1 XOR ... b_{j-1} XOR b_j

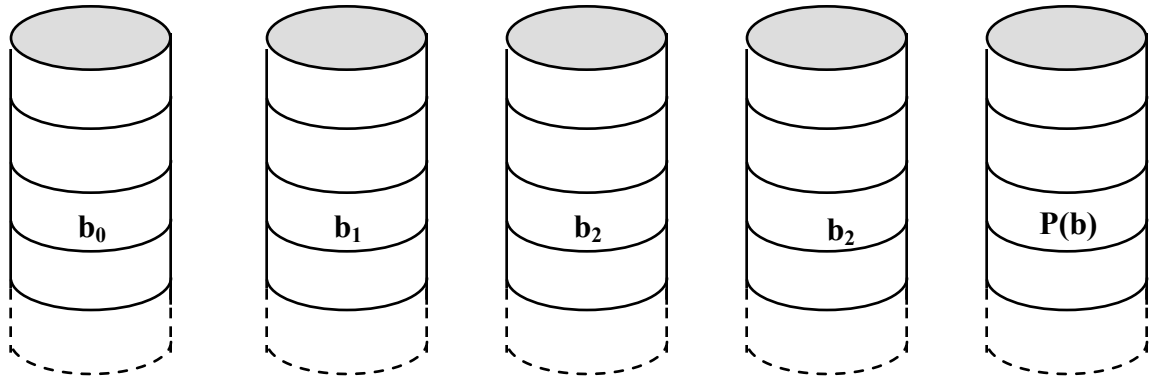
Exemplo:

- 0 1 1 1
- Bit paridade 1 = 0 xor 1 xor 1 xor 1
- Se o primeiro disco falhar os dados podem ser corrigidos aplicando novamente XORs:
  1 xor 1 xor 1 xor 1 = 0.

Isto permite:

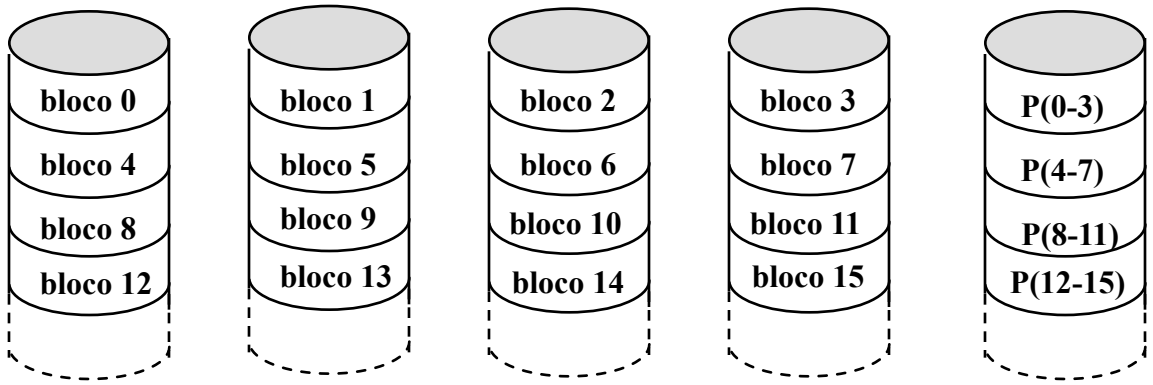
Operar com um disco estragado

Introduzir um disco e refazer o seu conteúdo



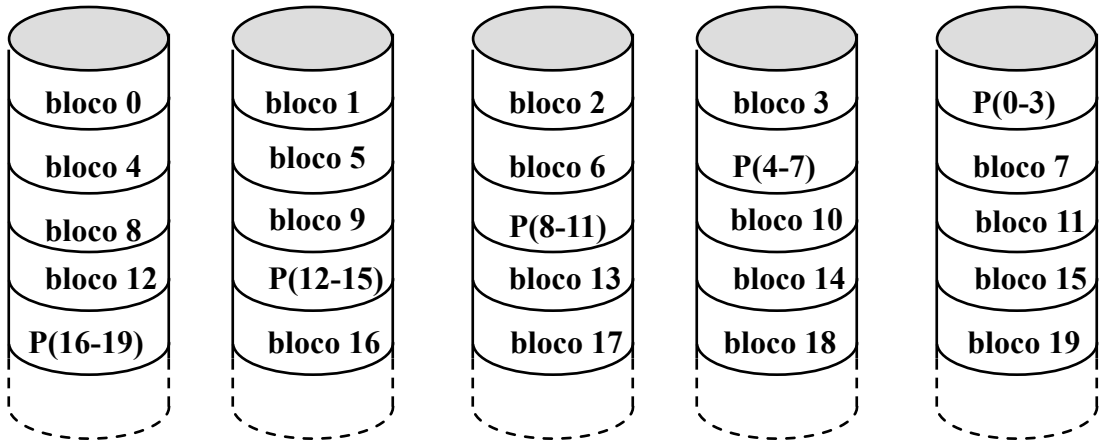
RAID 4 (paridade a nível do bloco)

- Semelhante ao anterior.
  - A diferença está na unidade do interleaving, que passa do bit para um bloco de dados.



RAID 5 (paridade a nível do bloco distribuída)

- A informação da paridade é distribuída pelos vários discos
- Elimina o problema do número de acessos ao disco de paridade.



Escritas em RAID 5

- RAID 5 é um bom compromisso velocidade/espaço desperdiçado para redundância
- Em RAID 5 uma operação de escrita inclui:
  - Escrever um novo conteúdo Y um bloco de dados D que continha X:
    - Ler o bloco D
    - Ler o bloco de paridade correspondente com o conteúdo Z
    - Escrever Y em D
    - Se fizermos (X xor Y) temos a 1 os bits diferentes
    - o novo conteúdo do bloco é Z xor (X xor Y); troca os bits de paridade nas posições em que há diferenças
  - 2 leituras e duas escritas, mas pode haver paralelismo entre as duas leituras e as duas escritas