FSO - 29/10/2018

Sumário:

Tecnologia de discos:

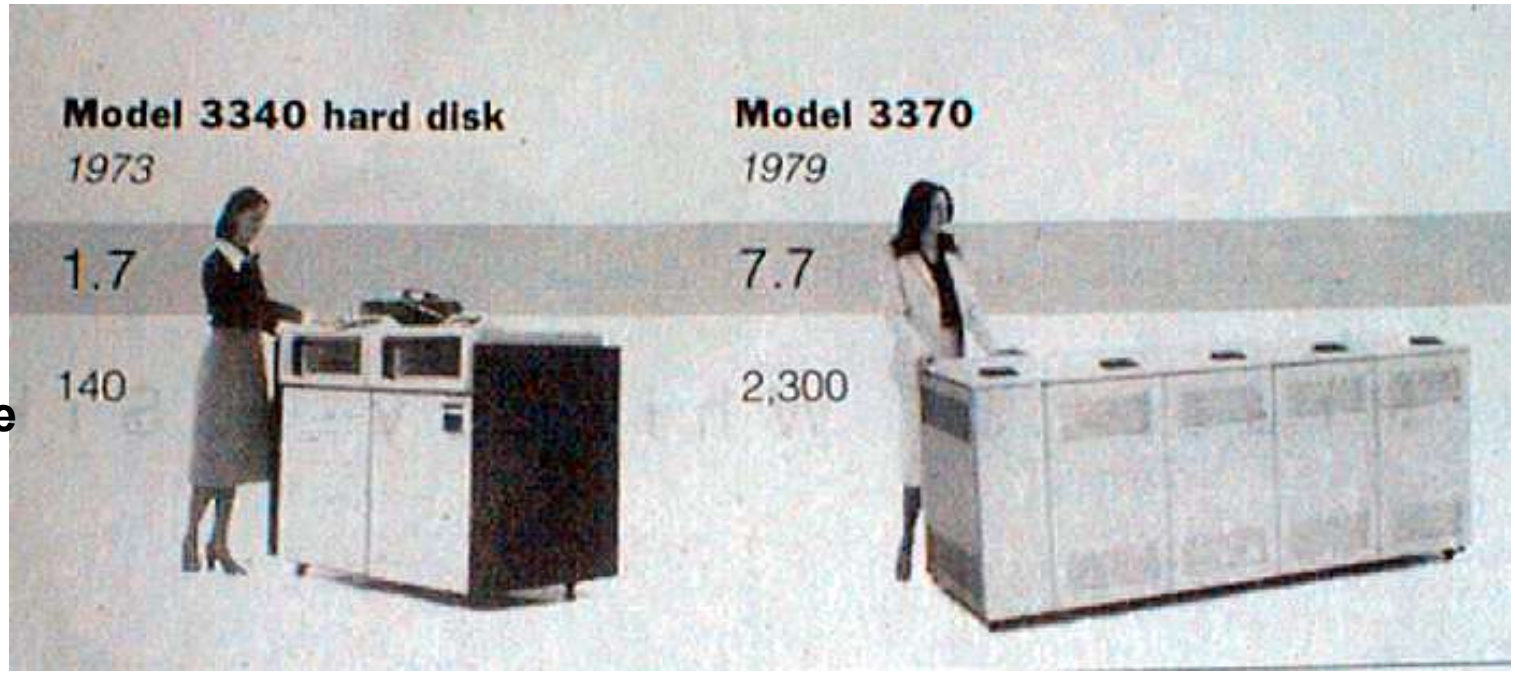
- Discos magnéticos
- RAID

Bibliografia: OSTEP 37, 38

História dos discos

Densidade Mbit/sq. in.

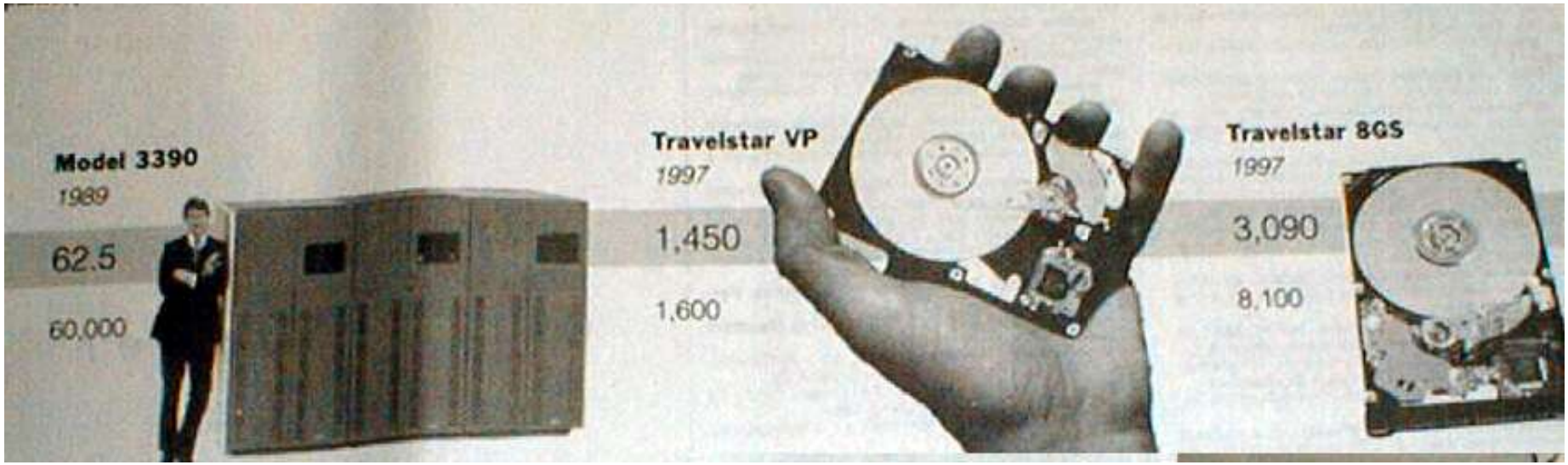
Capacidade em Megabytes



1973:
1. 7 Mbit/sq. in
140 MBytes

1979:
7. 7 Mbit/sq. in
2,300 MBytes

História dos discos



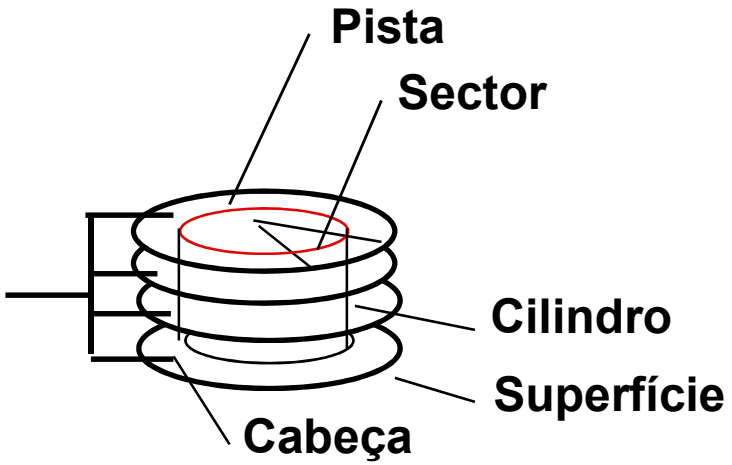
1989:
63 Mbit/sq. in
60,000 MBytes

1997:
1450 Mbit/sq. in
2300 MBytes

1997:
3090 Mbit/sq. in
8100 MBytes

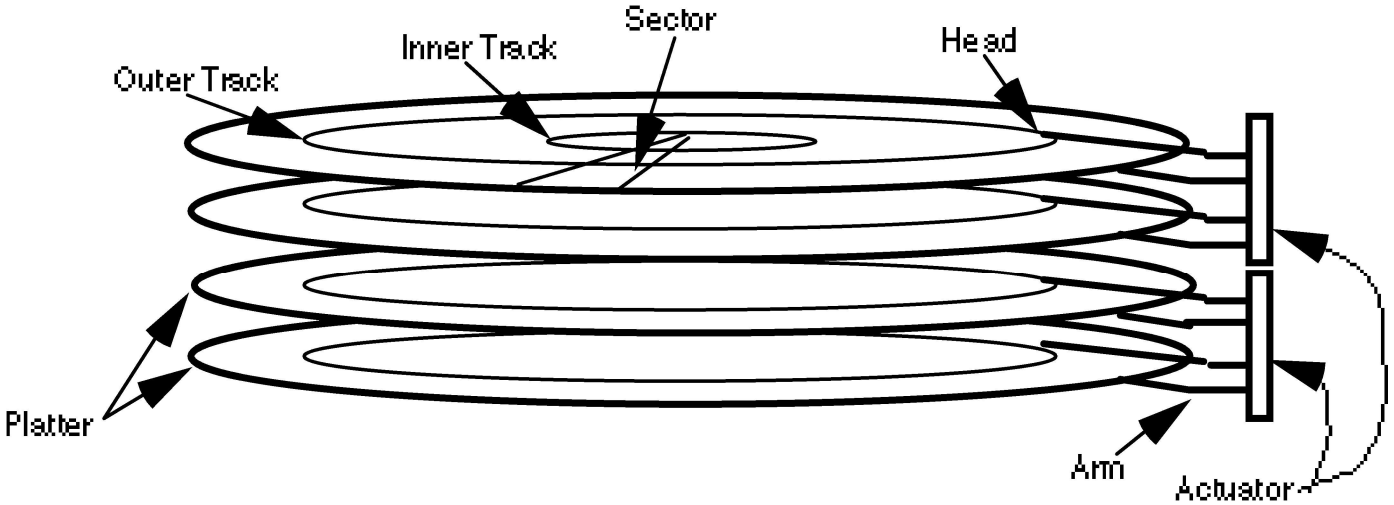
Discos magnéticos

- Características:
  - armazenamento não volátil a longo prazo
  - grande capacidade, nível mais lento na hierarquia de memória
  - Transferência por blocos
- Capacidade: gigabytes e quadruplica de 3 em 3 anos



Terminologia de discos

Latência do disco = Tempo de espera na fila + Tempo do controlador +
Tempo de Seek + Tempo de Rotação + Tempo de Transferência



Tempos aproximados para a transferência de um bloco (clusters) de 4K byte :

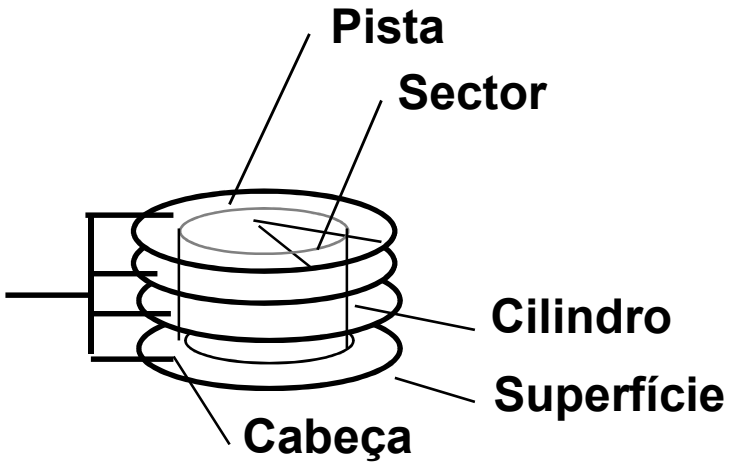
Seek: 8 ms ou menos

Rotação: 4.2 ms @ 7200 rpm

Transferência: 1 ms @ 7200 rpm

Discos magnéticos

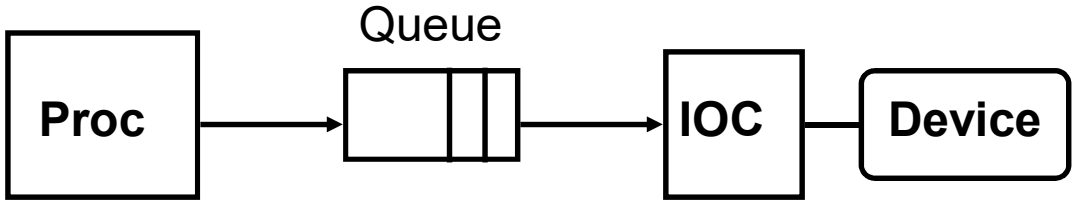
- Seek Time (~8/10 ms fabricante)
  - depende do número de pistas atravessado
  - valor anunciado pelo fabricante corresponde à travessia de um número de pistas fixo (ex. metade das pistas)
- Latência rotacional
  - Metade do tempo de rotação
- Tempo de transferência: aproximado por nº blocos por pista / tempo de rotação
- *Tempo de transferência controlador -> RAM* : tamanho_bloco / vel de transferência do Bus de entrada/saída
- Tempo de transferência *disco-> controlador* e *controlador-> RAM* (DMA) é desprezável face aos outros



7200 RPM = 120 RPS => 8 ms por rot.
Latência de rotação média = 4 ms
128 sectors por pista => 0.25 ms por sector
1 KB por sector => 16 MB / s

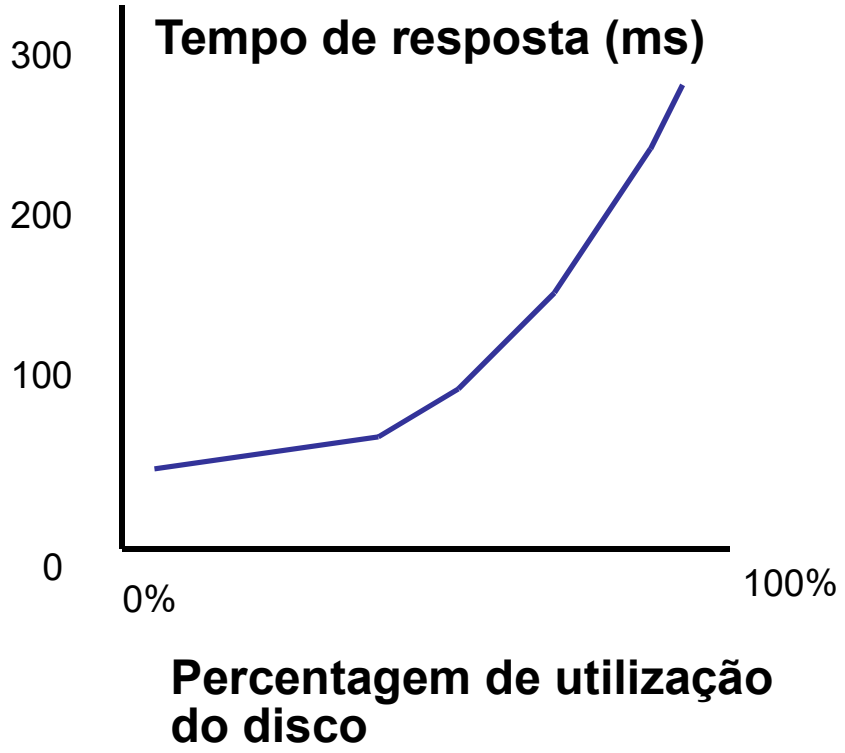
Performance do I/O de disco

Tempo de resposta = Tempo na Fila + Tempo de Serviço do Disco



Teoria das filas de espera
Tr – tempo de resposta
Ts – tempo de serviço
Tq – tempo de espera na fila
U - taxa de utilização
U = nº de pedidos/s * Ts

Tr = Ts/ (1- U)



Tempo de resposta de um disco

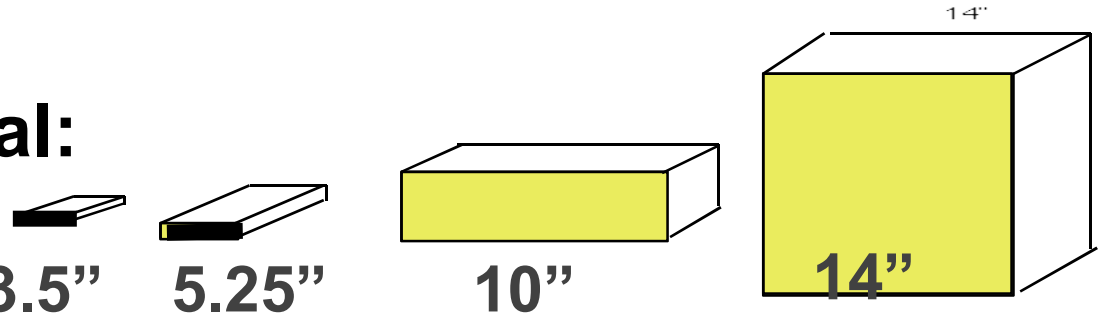
- Parâmetros do disco:
  - Tamanho do bloco é 8K bytes
  - Seek time publicitado é de 12 ms
  - Disco roda a 7200 RPM
  - taxa de transferência é 4 MB/sec
- Overhead no controlador de 2 ms
- Suponhamos que o disco está sempre livre - não há tempo de espera
- Qual é o tempo médio de acesso a um sector?
  - T. seek médio + rot delay médio + tempo transf. + ov. controlador
  - 12 ms + 0.5/(7200 RPM/60) + 8 KB/4 MB/s + 2 ms
  - 12 + 4.15 + 2 + 2 = 20 ms
- O seek time assume que não há localidade: o real é tipicamente 1/4 a 1/3 do anunciado: 20 ms => 12 ms

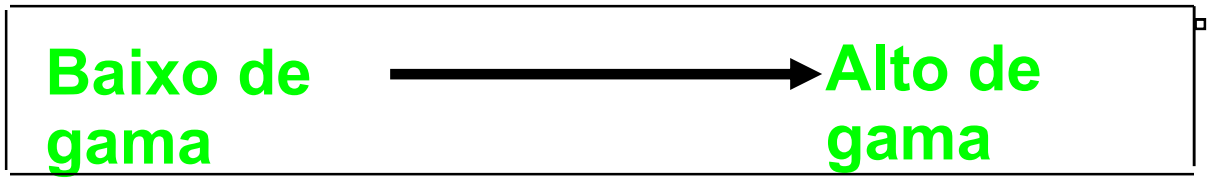
Estruturas RAID

- **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: permite tolerância a falhas, porque há discos extra que asseguram redundância
- Outros níveis definidos mais recentemente
  - 6, 10 (ou 1+0) …

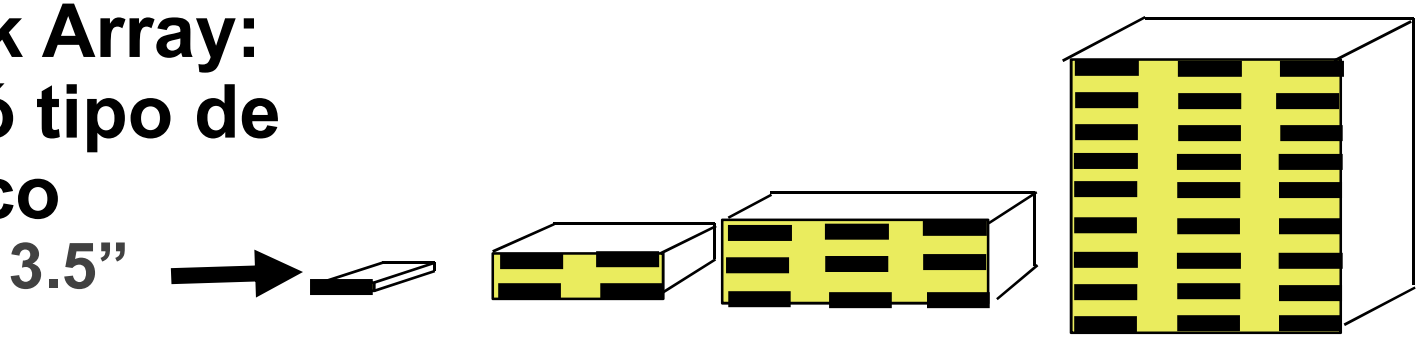
Vantagens do fabrico de Disk Arrays

Convencional: 4 tipos de discos





Disk Array: 1 só tipo de disco



Substituir um grande disco por muitos pequenos discos! (Patterson 1988)

	IBM 3390 (K)	IBM 3.5" 0061	*x70*
Capacidade	20 GBytes	320 MBytes	*23 GBytes*
Volume	97 cu. ft.	0.1 cu. ft.	*11 cu. ft.*
Potência	3 KW	11 W	*1 KW*
Ritmo de transf	15 MB/s	1.5 MB/s	*120 MB/s*
Ritmo de ops I/O	600 I/Os/s	55 I/Os/s	*3900 IOs/s*
MTTF	250 KHrs	50 KHrs	*??? Hrs*
Custo	$250K	$2K	*$150K*

Disk Arrays têm potencial para

- Grandes taxas de transferência
- high MB por cu. ft., high MB por KW
- fiabilidade?

Fiabilidade do Array

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

  50,000 Horas ÷ 70 discos = 700 horas

  MTTF 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 tem uma disponibilidade extremamente elevada

RAID=Redundant Arrays of Disks

- **Os ficheiros são distribuídos ("striped“) por várias unidades de disco**
- **Redundância garante alta disponibilidade dos dados**

Em caso de falha de um disco, o conteúdo é reconstruído a partir de dados redundadntes armazenados no array

→ **Perde-se capacidade para os armazenar**

→ **Existe uma penalização em bandwidth para actualização**

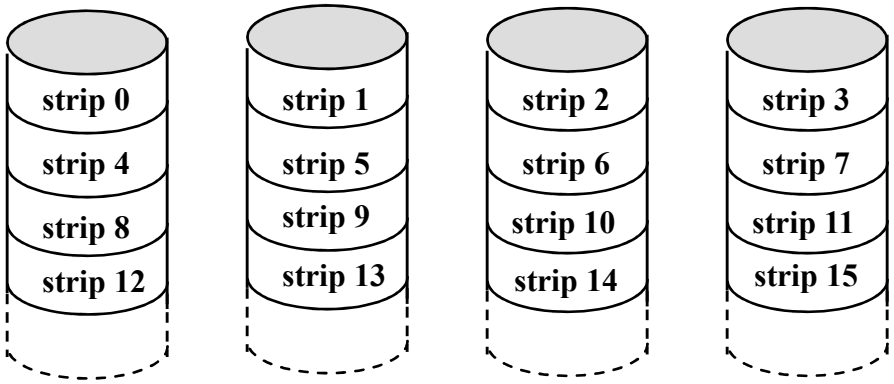
Técnicas:

- **Mirroring/Shadowing (elevado custo em capacidade)**
- **Códigos de Hamming Horizontais (idem)**
- **Paridade & Códigos de Reed-Solomon**

RAID (cont)

- 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 0 (não-redundante)



RAID 1
Disk Mirroring/Shadowing

- **Cada disco é completamente duplicado no seu "shadow"**
  Pode ser atingida uma disponilidade muito elevada

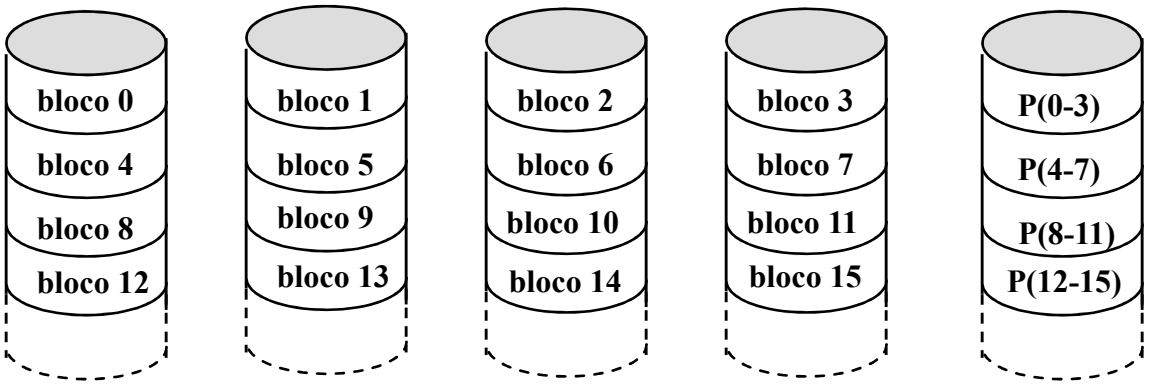
- **Sacrifício da largura de banda em escrita:**
  uma escrita lógica = duas escritas físicas

- **As leituras podem ser optimizadas**

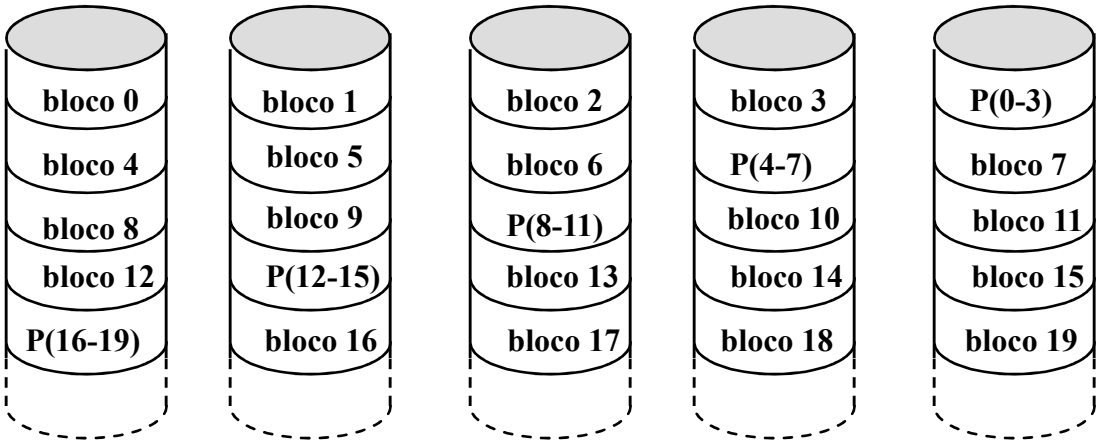
- **A solução mais cara: 100% de custos extra em capacidade**

Usado em ambientes em que interssa alta disponibilidade
e uma taxa elevada de operações de IO

RAID 4 (paridade a nível do bloco)



RAID 5 (paridade a nível do bloco distribuída)



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