Capítulo 5: Parte 2

- Hardware do Disco
- Estrutura RAID
- Formatação
- Escalonamento de Disco

Hardware do Disco

- Grande variedade de tipos de Discos. Mais comuns: discos magnéticos (rígidos e flexíveis), característica é acesso rápido. Discos ópticos : CD-ROM, DVDs, …
- Discos organizados em cilindros, cilindro contém trilhas, divididas em setores.
- Discos + antigos, pouca eletrônica, controlador faz todo trabalho.
- Discos IDE*(integrated drive eletronics),* SATA: serviço realizado pela unidade de disco (micro), controlador emite comandos de alto nível. Controlar cache, remapear blocos defeituosos.
- Figura compara floppy com HD (PC IBM)

O dispositivo disco

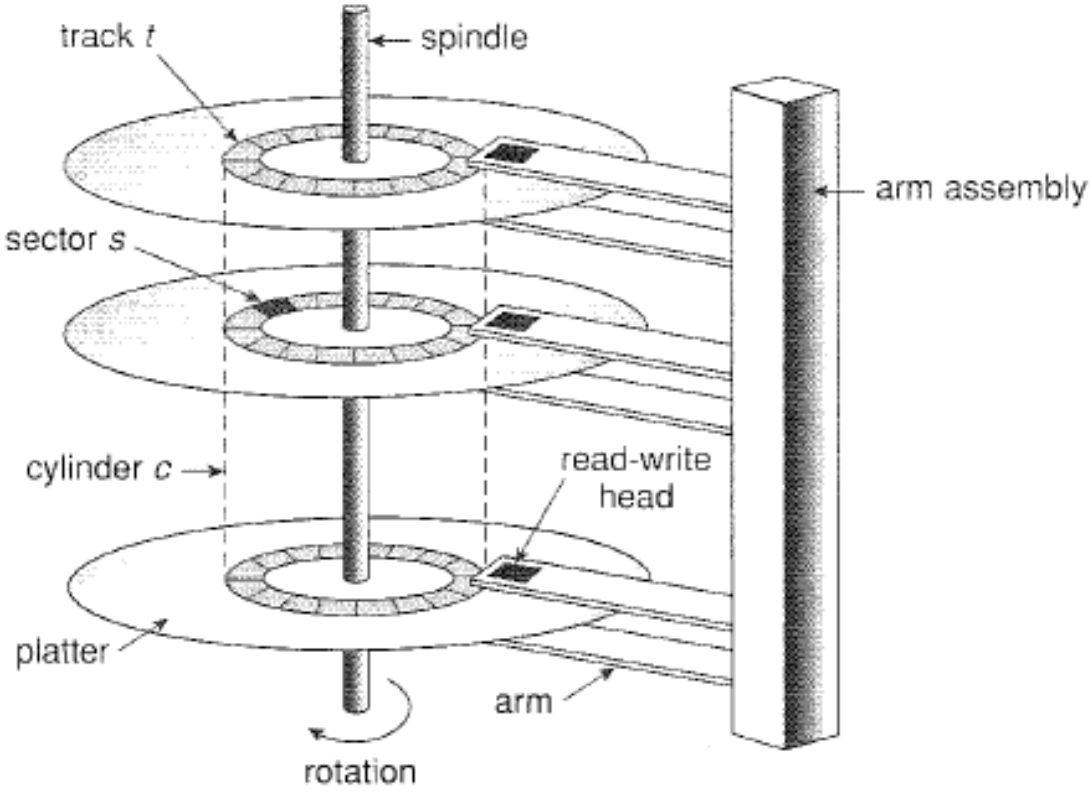


Figure 12.1 Moving-head disk mechanism.

Geometria do disco

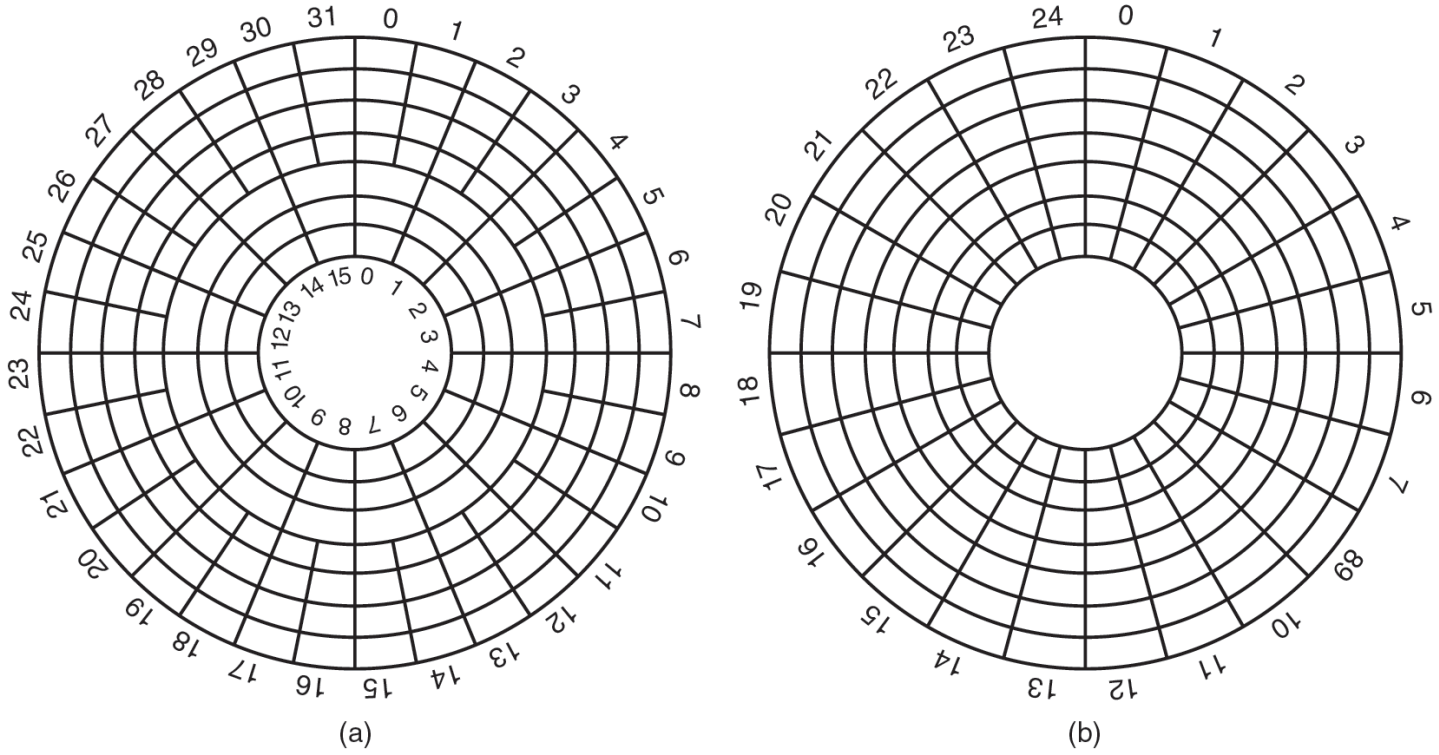


Figura 5.16 (a) Geometria física de um disco com duas zonas. (b) Uma possível geometria virtual para esse disco.

Hardware do Disco

Parâmetro	Unidade de disquete IBM PC 360 KB	Disco rígido Western Digital WD 18300
Número de cilindros	40	10601
Trilhas por cilindro	2	12
Setores por trilha	9	281 (em média)
Setores por disco	720	35742000
Bytes por setor	512	512
Capacidade do disco	360 KB	18,3 GB
Tempo de busca (cilindros adjacentes)	6 ms	0,8 ms
Tempo de busca (em média)	77 ms	6,9 ms
Tempo de rotação	200 ms	8,33 ms
Tempo para parada/início do motor	250 ms	20 ms
Tempo de transferência de um setor	22 ms	17 µs

Tabela 5.3 Parâmetros de disco para a unidade de disquete do IBM PC 360 KB e para o disco rígido do Western Digital WD 18300.

Estrutura do Disco

- Discos são endereçados como grandes arrays unidimensionais de *blocos lógicos*, onde o bloco lógico é a menor unidade de transferência.

- O array de blocos lógicos é mapeado nos setores do disco sequencialmente.
  - Setor 0 é o primeiro setor da primeira trilha no cilindro mais afastado do centro.
  - Mapeamento prossegue em ordem na trilha, depois no restante das trilhas daquele cilindro e depois no restante dos cilindros de fora para dentro.
- Meta → reduzir tempo de acesso
  - Entrelaçamento
  - Escalonamento
  - RAID – E/S paralela

Tempo de Acesso ao Disco

- Para realizar um acesso a um disco rígido, é necessário posicionar o cabeçote de leitura e escrita sob um determinado setor e trilha onde o dado será lido ou escrito. O tempo de acesso é definido por 3 fatores:
  - $T_{acesso} = T_{seek} + T_{rotacional} + T_{transferencia}$

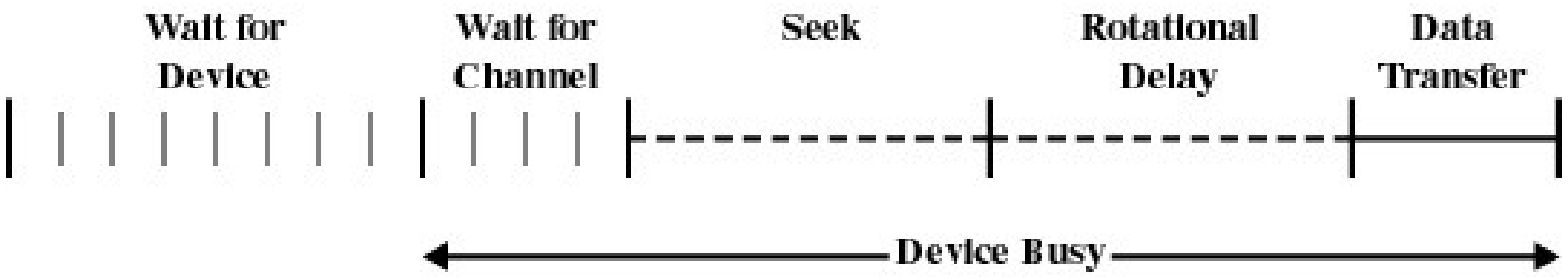


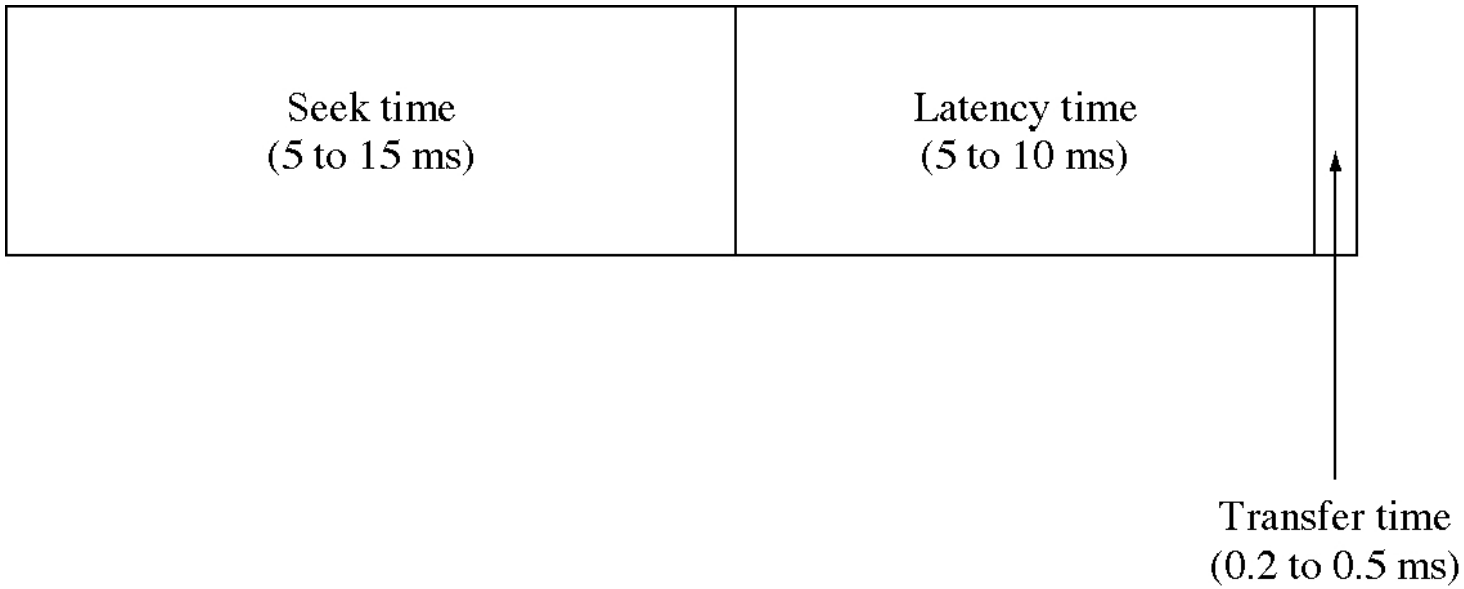
Figure 11.7 Timing of a Disk I/O Transfer

Tempos de E/S no Disco

- Tempo de posicionamento
  - Tempo necessário para deslocar o cabeçote de leitura e escrita até o cilindro correspondente à trilha a ser acessada
- Tempo rotacional / latência
  - Tempo necessário, uma vez que o cabeçote já está na trilha correta, para o setor a ser lido ou escrito, se posicionar sob o cabeçote de leitura e escrita no início do setor.
- Tempo de transferência
  - Corresponde ao tempo necessário á transferência dos dados, isto é, a leitura ou escrita dos dados.
- $T_{acesso} = T_{seek} + T_{rotacional} + T_{transferencia}$

- Largura de banda do Disco é o número total de bytes transferidos dividido pelo tempo total entre a primeira requisição de serviço e o término da última transferência.

Tempos de disco



Formatação de disco

Formatação de baixo nível feita por software, criar trilhas concêntricas com setores formatados: preâmbulo (cilindro,setor,etc), dados (512bytes), ECC (recuperação de erro, 16 bytes).
Deslocamento de cilindro, para melhorar o desempenho, a posição do setor 0 em cada trilha é deslocada com relação à trilha anterior, em função da geometria do disco.
Como resultado da formatação a capacidade do disco é reduzida: depende dos tamanhos do preâmbulo, intervalo entre setores, ECC, no. Setores reserva (até 20% menor). Entrelaçamento.
Formatação lógica do disco (MBR+partições)Formatação de alto nível

Figura 5.22 Um setor de disco.

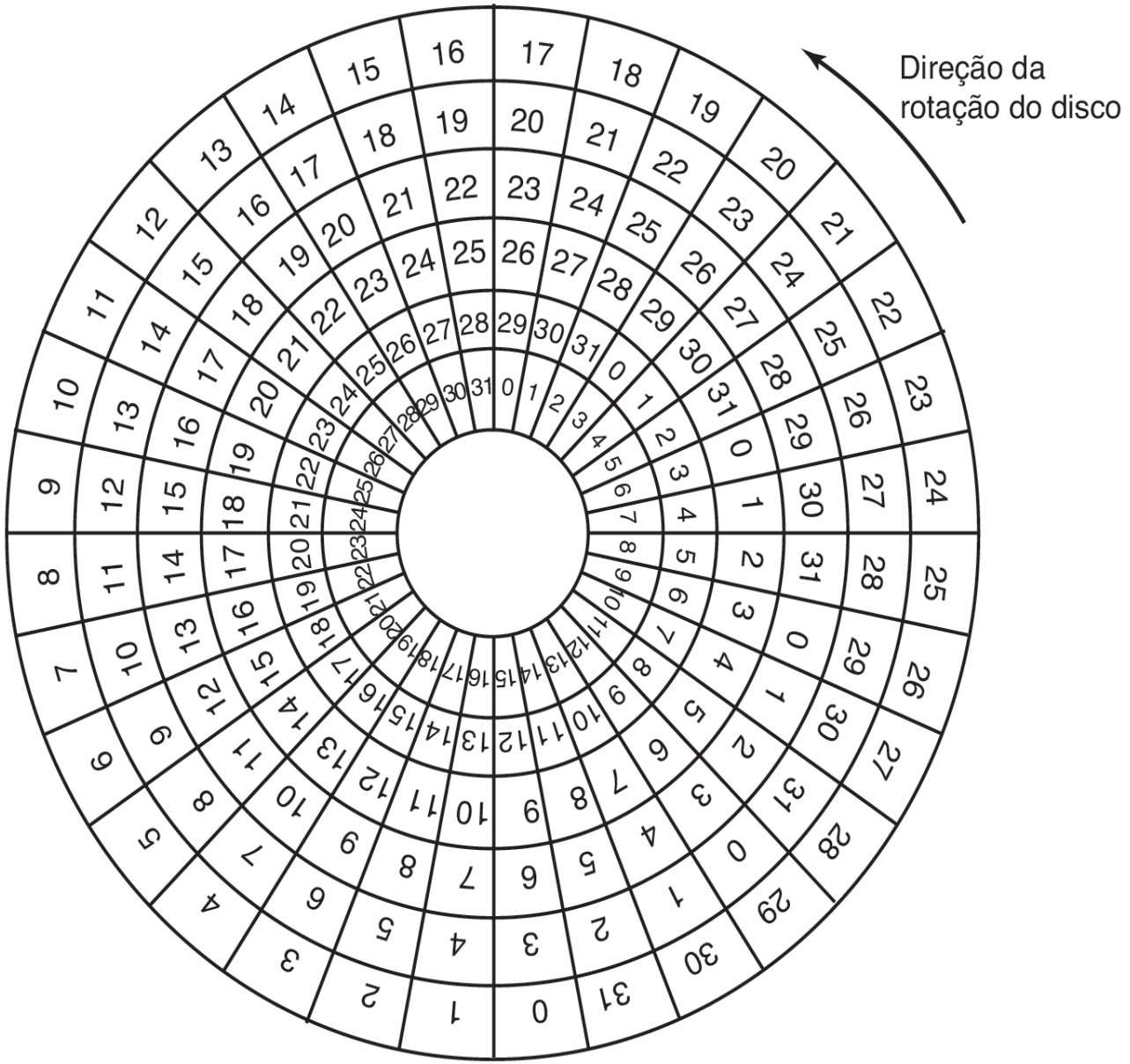


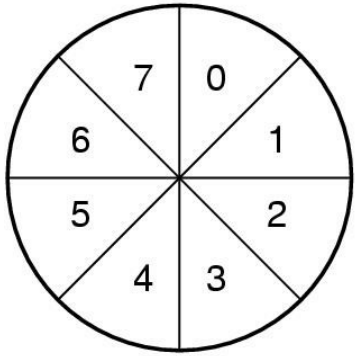
Figura 5.23 Ilustração de um deslocamento de cilindro.

Entrelaçamento

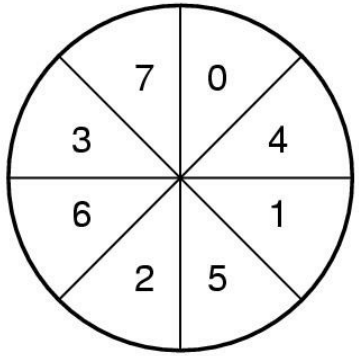
Considere um controlador com um buffer de 1 setor(512) para o qual foi passado um comando para leitura de 2 setores.

- Após a leitura do 1o., cálculo do ECC, dados devem ser transferidos para memória
- Enquanto transferência sendo feita cabeçote passa sobre setor seguinte
- Qdo cópia completa, controlador deverá esperar quase o tempo de rotação para acessar o segundo setor
- Isto pode ser eliminado a partir da numeração entrelaçada dos setores na formatação do disco:
  - Entrelaçamento simples
  - Entrelaçamento duplo

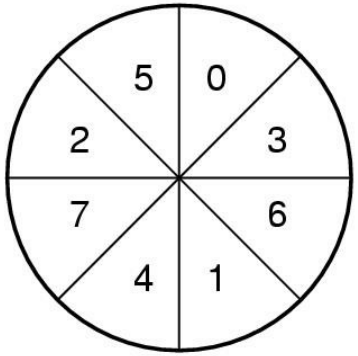
- Para evitar a necessidade de entrelaçamento, controlador capaz de colocar no buffer uma trilha inteira



(a)



(b)



(c)

Algoritmos de escalonamento de braço de disco

Fatores relacionados ao tempo de ler/escrever:

1. Tempo de posicionamento (o tempo para mover o braço para o cilindro correto).
2. Atraso na rotação (o tempo necessário para rotar o setor correto sob o cabeçote).
3. Tempo de transferência real dos dados.

Escalonamento do Disco

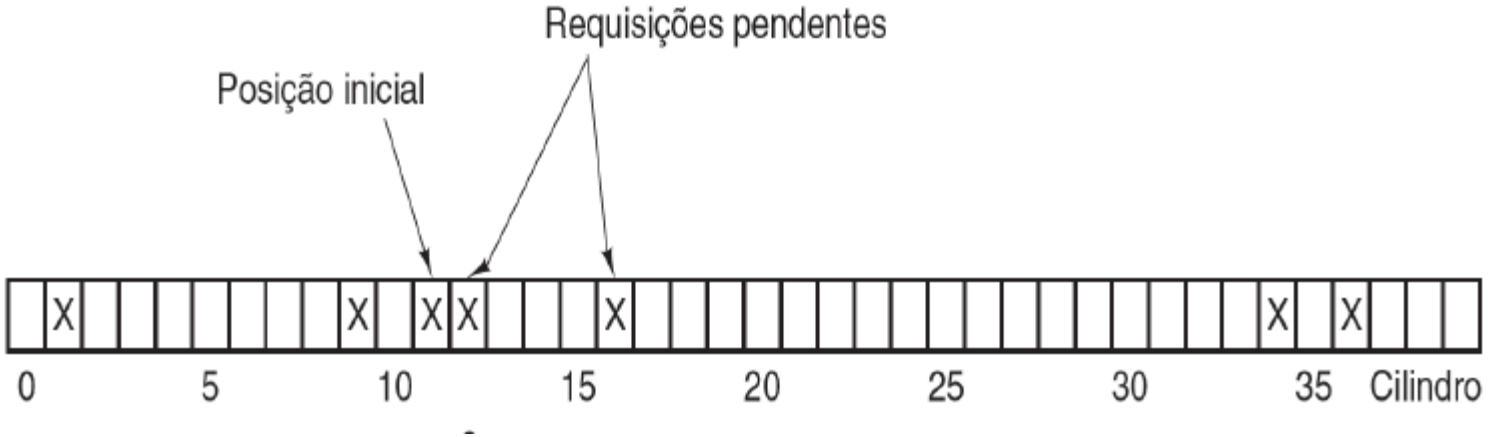
- Se existem 2 ou mais requisições de disco pendentes, qual deveria ser atendida primeiro?
- O tempo necessário a uma operação de E/S com disco é fortemente influenciado pelo tempo de acesso ao disco. Assim, minimizar os movimentos da cabeça de leitura/escrita e maximizar a transferência de bytes, atendendo mais requisições em menos tempo é a meta.
- Estratégias
  - First-come, first-served (FCFS)
  - Shortest-seek-time-first (SSTF)
  - Scan

Escalonamento de Disco

- FCFS
  - Se o driver do disco recebe requisições sequencialmente, uma após a outra e atende a todas na ordem que elas foram recebidas, ou seja, “a primeira que chegar será a primeira a ser atendida”, quase nada pode ser feito para otimizar o tempo de posicionamento.

FIFO

Considere um disco com 40 cilindros: Requisição para leitura de bloco no cilindro 11, enquanto isso, outras requisições 1, 36, 16, 34, 9 e 12 chegam. Usando FCFS qual seria a ordem de tratamento? Qual a distância total percorrida?



Escalonamento de Disco

- Menor tempo de serviço primeiro (Shortest Seek Time First)
  - Seleciona a requisição que necessita o menor movimento do braço do disco a partir da posição corrente
  - Sempre escolhe o tempo mínimo de seek

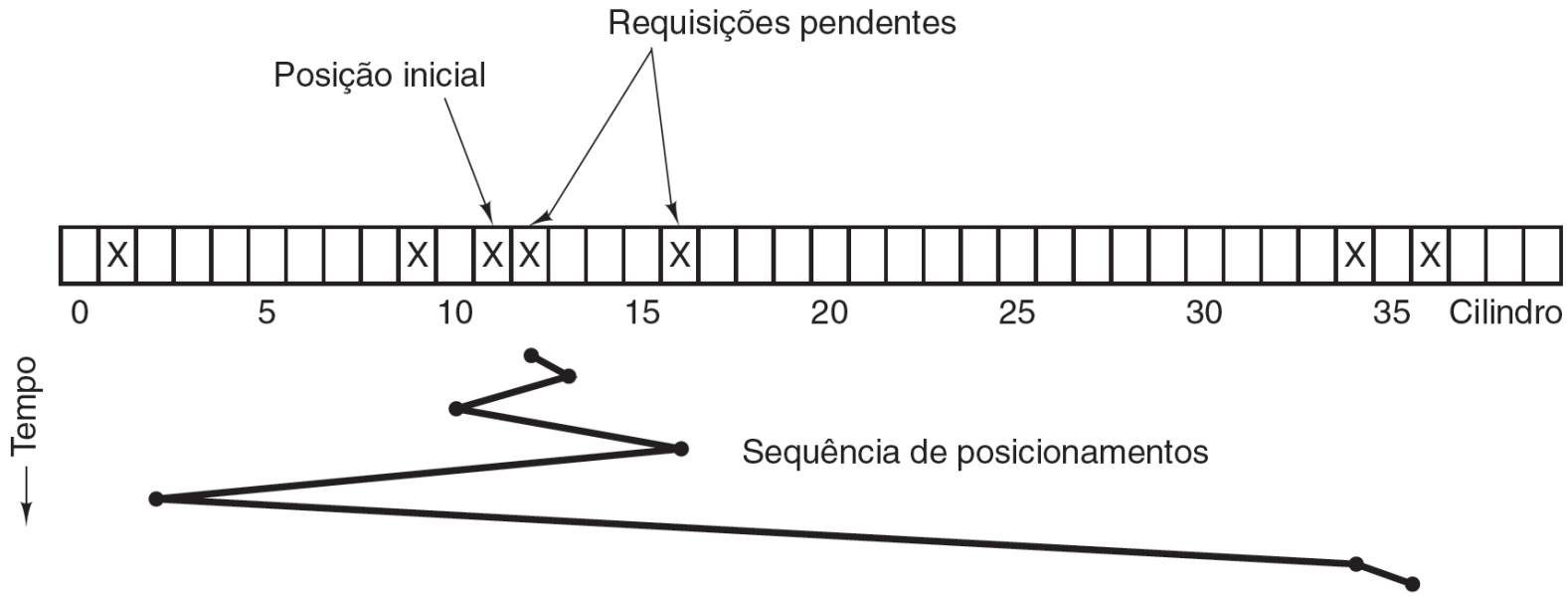


Figura 5.25 Algoritmo de escalonamento 'posicionamento mais curto primeiro' (SSF).

Escalonamento de Disco

- SCAN
  - Braço move apenas em uma direção, satisfazendo todas as requisições até encontrar a última trilha naquela direção
  - Direção é revertida
  - Conhecido como algoritmo do elevador

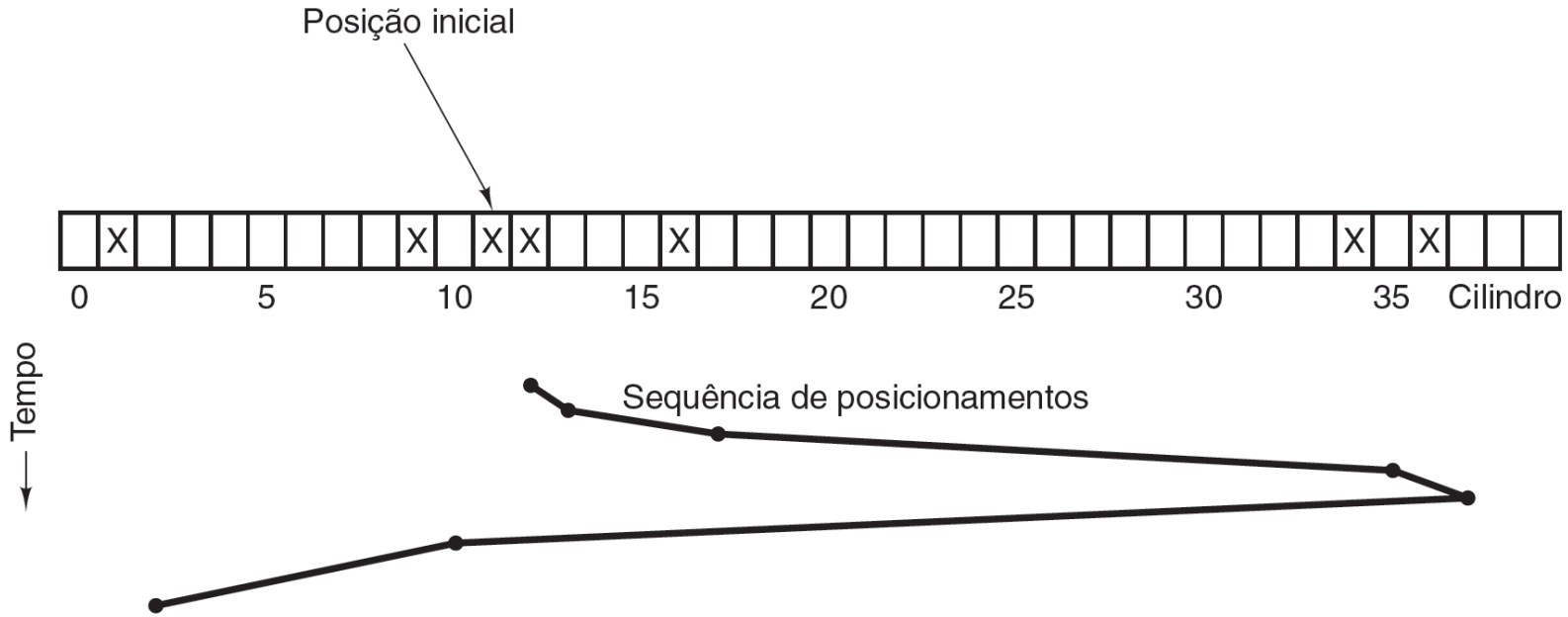
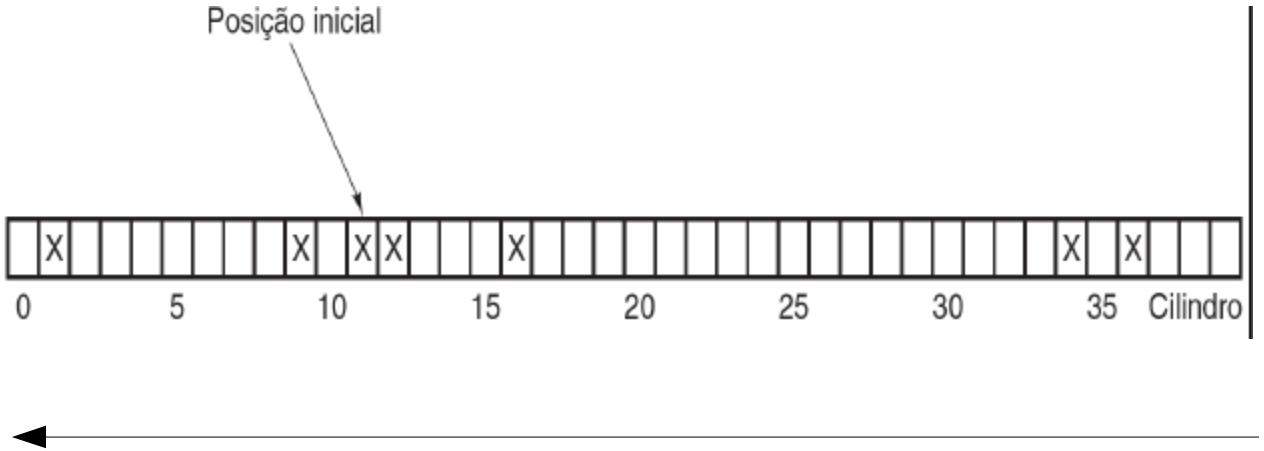


Figura 5.26 O algoritmo do elevador para escalonamento de solicitações do disco.

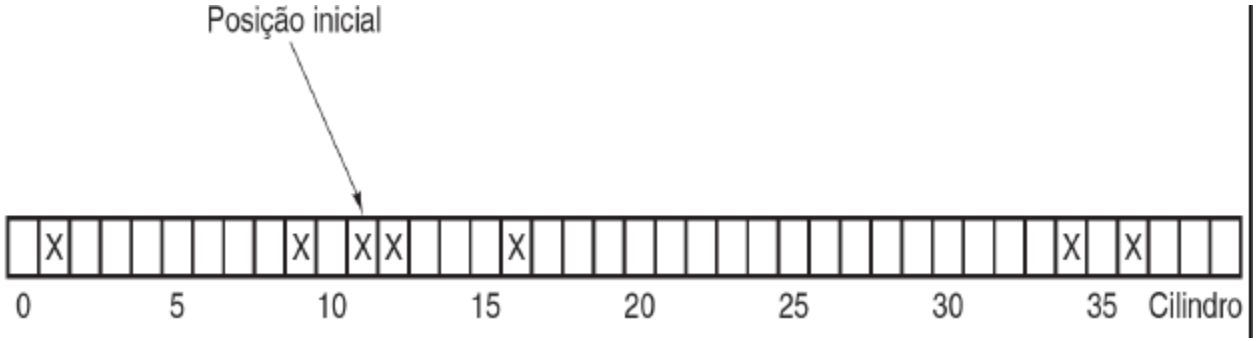
Escalonamento de Disco

- C-SCAN
  - Fornece tempo de espera mais uniforme que SCAN
  - Restringe busca em uma direção apenas
  - Quando a última trilha foi visitada em uma direção, o braço é retornado para o lado oposto do disco e a busca inícia novamente
  - Trata os cilindros como uma lista circular



Escalonamento de Disco

- C-Look
  - Versão do C-SCAN
  - Braço vai até a última requisição em cada direção, depois reverte a direção imediatamente, sem primeiro ir até o final do disco.



Escalonamento de Disco

- SSTF, SCAN e C-SCAN é possível que processos com taxas de acesso + rápidas monopolizem o disco
- N-step-SCAN
  - Segmenta a fila de requisições do disco em sub filas de tamanho N
  - Sub filas são processadas uma de cada vez, usando SCAN
  - Novas requisições adicionadas a outra fila quando fila é processada
- FSCAN
  - Duas sub-filas são usadas
  - Quando o SCAN começa todas as requisições estão em uma das filas, sendo a outra fila vazia, recebe novas requisições durante o SCAN
  - Serviço de novas requisições é postergado até todas requisições antigas serem processadas

Exercício

- Considere um disco com 200 trilhas (0 a 199) e a seguinte fila de requisições (98, 186, 37, 122, 14, 124, 65, 67). A cabeça de leitura está posicionada inicialmente na trilha 53. Faça análise deste cenário usando os algoritmos FIFO, SSTF, SCAN e C-SCAN e indique qual o numero de cilindros percorridos em cada uma das políticas, considerando:
  - A cabeça se move na direção decrescente de trilhas

Estrutura RAID

- Desempenho CPU aumento exponencial, não ocorre com disco, de 50 a 100ms (1970) a cerca de 10ms atual. Diferença desempenho CPU/disco acentuada. Processamento paralelo acelerar desempenho da CPU. E/S paralela pode ser uma saída. Em 1998, Paterson et al., sugerem organizações (6) para os discos, objetivo melhorar desempenho/confiabilidade – RAID.
- **RAID** – Redundant Array of Inexpensive(Paterson) / Independent(indústria) Disks, multiplos discos fornecem **confiabilidade** via **redundância**.
  - Idéia básica vários discos na mesma “caixa”, comandados por um controlador RAID, para o SO um RAID se parece como um disco único.
  - Propriedade de os dados serem distribuídos pelos dispositivos, permitindo operações em paralelo.
  - Esquemas definidos por Paterson chamados de RAID 0 a RAID 5
- RAID combina vários discos em uma estrutura lógica, propósito de armazenar informações de forma redundante para permitir a recuperação de dados em caso de falha de um disco.
- Desempenho através da escrita em paralelo nos diferentes discos, forma de escrita e acesso (*stripping*) define níveis/organizações de RAID (6), *strip* (faixa) pode ser : bloco físico, setor ou outra unidade.

Estrutura RAID

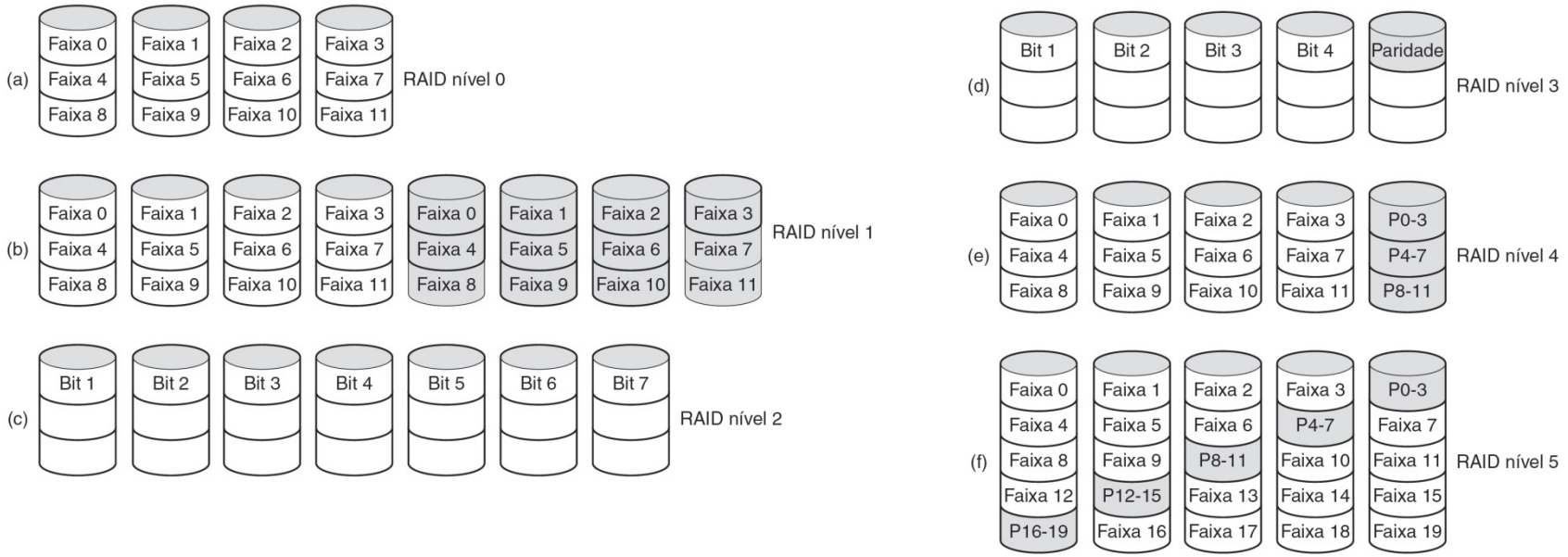


Figura 5.17 RAID níveis 0 a 5. Os discos de cópia de segurança e paridade estão sombreados.

RAID Nível 0

RAID 0 (5.17a) – visualização de um único disco virtual. Divisão em faixas de k setores cada (0 a k-1; k a 2k-1), gravadas de forma alternada (round-robin).

Leitura de um bloco nas 4 faixas, 4 comandos (E/S paralelo).

Controlador responsável pela partição da requisição (figura a seguir).

SOs com requisição de um setor por vez, desempenho inferior.

Confiabilidade é menor que um SLED(single large expensive disk).

Não é considerado RAID de fato, porque não tem redundância.

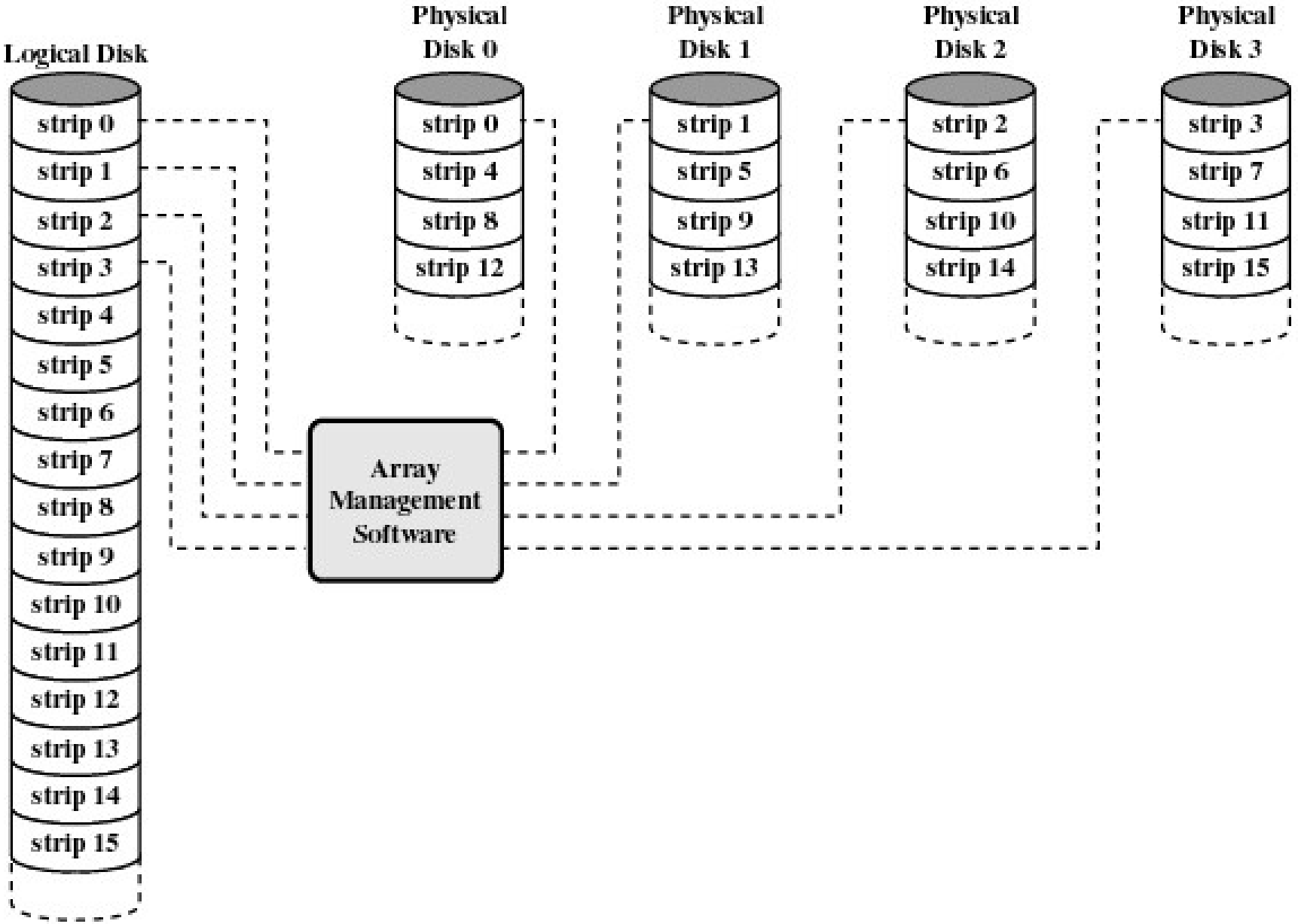


Figure 11.10 Data Mapping for a RAID Level 0 Array [MASS97]

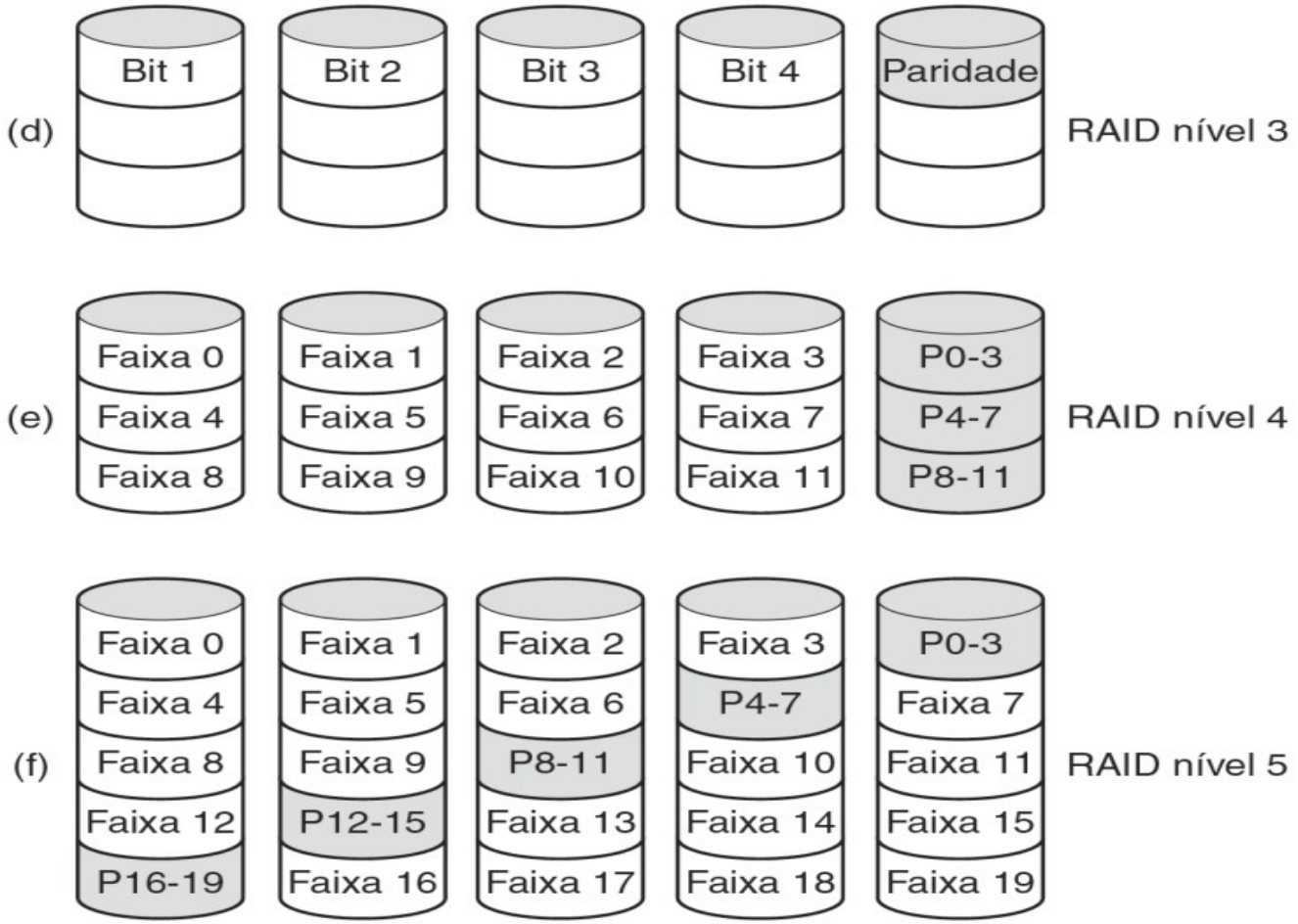
RAID 1

- RAID 1 (5.17b) – é considerada uma verdadeira organização RAID, conhecida como espelhamento, onde o dado é escrito em disco primário e secundário (espelho), a redundância é conseguida através da duplicação.
- Durante uma escrita (faixa escrita 2x), durante leitura qualquer cópia. Desempenho da escrita não é melhor que uma única cópia, leitura até 2x melhor.
- Tolerância a falhas muito bom. Recuperação= instalar disco novo transferindo cópia de segurança.
- Necessário espaço físico = dobro da capacidade de armazenamento.

RAID 2

- RAID 2 (5.17 c) - Diferente de 0 e 1 (faixas de setores) 2 (palavras/bytes):
  - Quebra byte em pares de 4bits, adiciona Hamming para formar 7bits, sendo 1-2-4 bits de paridade
  - Sete discos sincronizados: posicionamento de braço e rotação.
    - Escrever palavra 7bits (Hamming) nos 7 discos, um bit/disco
- Ex.:CM-2 - 32discos de dados e 6 de paridade – +bit = 39 discos. Ganho enorme, no tempo de acesso setor, escreve 32 setores de dados. Perda de um disco não é problema. Desvantagem: todos discos sincronizados, numero substancial de discos (ex. Sobrecarga 19%), exige bastante do controlador, fazer verificação de erro do código de H a cada chegada de bit.

RAID 3, 4 e 5


(d)
Bit 1
Bit 2
Bit 3
Bit 4
Paridade
RAID nível 3
(e)
Faixa 0
Faixa 4
Faixa 8
Faixa 1
Faixa 5
Faixa 9
Faixa 2
Faixa 6
Faixa 10
Faixa 3
Faixa 7
Faixa 11
P0-3
P4-7
P8-11
RAID nível 4
(f)
Faixa 0
Faixa 4
Faixa 8
Faixa 12
P16-19
Faixa 1
Faixa 5
Faixa 9
P12-15
Faixa 16
Faixa 2
Faixa 6
P8-11
Faixa 13
Faixa 17
Faixa 3
P4-7
Faixa 10
Faixa 14
Faixa 18
P0-3
Faixa 7
Faixa 11
Faixa 15
Faixa 19
RAID nível 5

RAID 3

- RAID 3 (5.17 d) - Versão simplificada do RAID 2. Um único bit de paridade é calculado para cada palavra de dados, requer apenas 1 disco de paridade.
- Os discos devem estar sincronizados, as palavras de dados individuais são distribuídas nos vários discos.
- Requisições tratadas não melhor que um único disco.

RAID 4

- RAID 4 e 5, trabalham com faixas de setores e não necessitam de sincronização nos discos.
- Raid 4 (5.17 e) - A paridade entre as faixas é escrita em um disco extra. Se cada faixa tem k bytes de tamanho todas as faixas são processadas juntas por meio de um OU EXCLUSIVO, resultando em uma faixa de paridade de k bytes. Se um disco quebra, bytes perdidos são recalculados a partir do disco de paridade.
- Protege contra a perda de um disco, mas não funciona muito bem para pequenas atualizações. Disco de paridade se torna gargalo.

RAID 5

- RAID 5 (5.17 f) – Eliminar o gargalo de RAID 4, distribui de forma uniforme os bits de paridade em todos os discos, de modo circular.
- No caso de quebra, reconstrução é um processo complexo.

Tratamento de erros

Defeitos de fabricação causam setores defeituosos, que demandam tratamento. É possível trata-los via controlador ou SO. A substituição de setores defeituosos é feita de duas maneiras: remapeamento (b) e deslocamento (c).
Erro de posicionamento – braço chega no destino verifica cilindro atual do preâmbulo do setor seguinte, erro se o local esta errado.

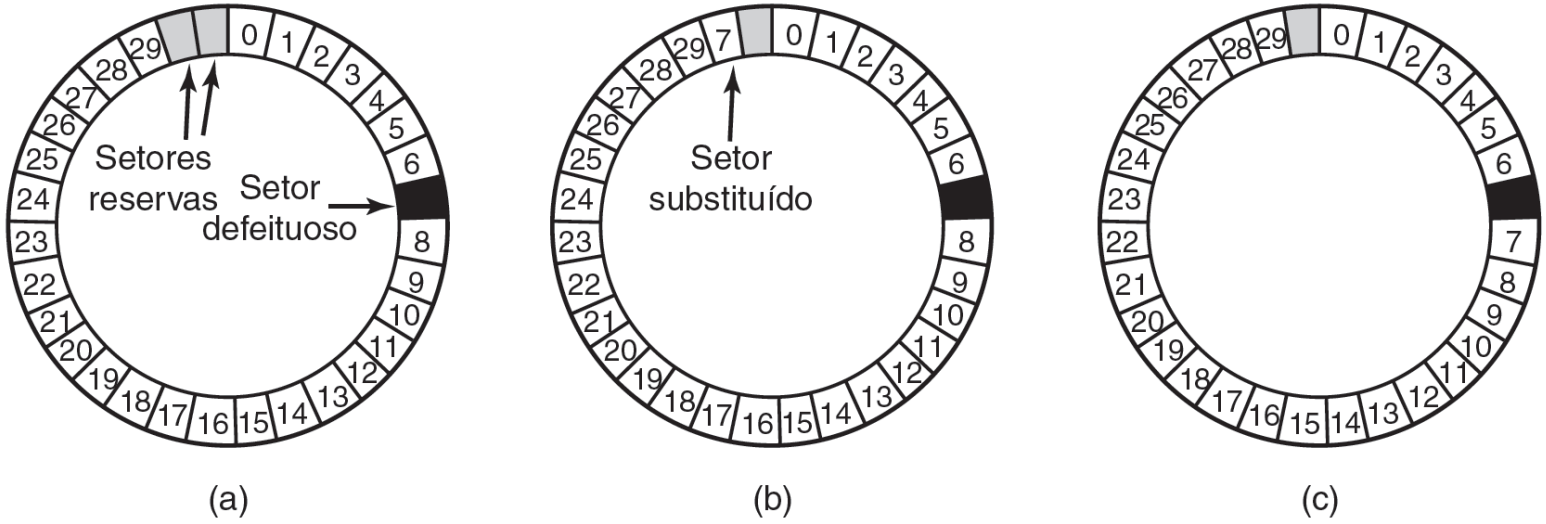


Figura 5.27 (a) Uma trilha de disco com setor defeituoso. (b) Substituição do setor defeituoso por um setor reserva. (c) Deslocamento de todos os setores para pular o setor defeituoso.