

Capítulo 5

Entrada/Saída

- 5.1 Princípios do hardware de E/S
- 5.2 Princípios do software de E/S
- 5.3 Camadas do software de E/S
- 5.4 Discos

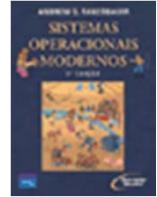
Princípios do Hardware de E/S



| Dispositivo | Taxa de dados |
|--|---------------|
| Teclado | 10 bytes/s |
| Mouse | 100 bytes/s |
| Modem 56 K | 7 KB/s |
| Canal telefônico | 8 KB/s |
| Linhas ISDN dual | 16 KB/s |
| Impressora a laser | 100 KB/s |
| Scanner | 400 KB/s |
| Ethernet clássica | 1,25 MB/s |
| USB (<i>universal serial bus</i> — barramento serial universal) | 1,5 MB/s |
| Câmera de vídeo digital | 4 MB/s |
| Disco IDE | 5 MB/s |
| CD-ROM 40x | 6 MB/s |
| Ethernet rápida | 12,5 MB/s |
| Barramento ISA | 16,7 MB/s |
| Disco EIDE (ATA-2) | 16,7 MB/s |
| FireWire (IEEE 1394) | 50 MB/s |
| Monitor XGA | 60 MB/s |
| Rede SONET OC-12 | 78 MB/s |
| Disco SCSI Ultra 2 | 80 MB/s |
| Ethernet Gigabit | 125 MB/s |
| Dispositivo de Fita Ultrium | 320 MB/s |
| Barramento PCI | 528 MB/s |
| Barramento da Sun Gigaplane XB | 20 GB/s |

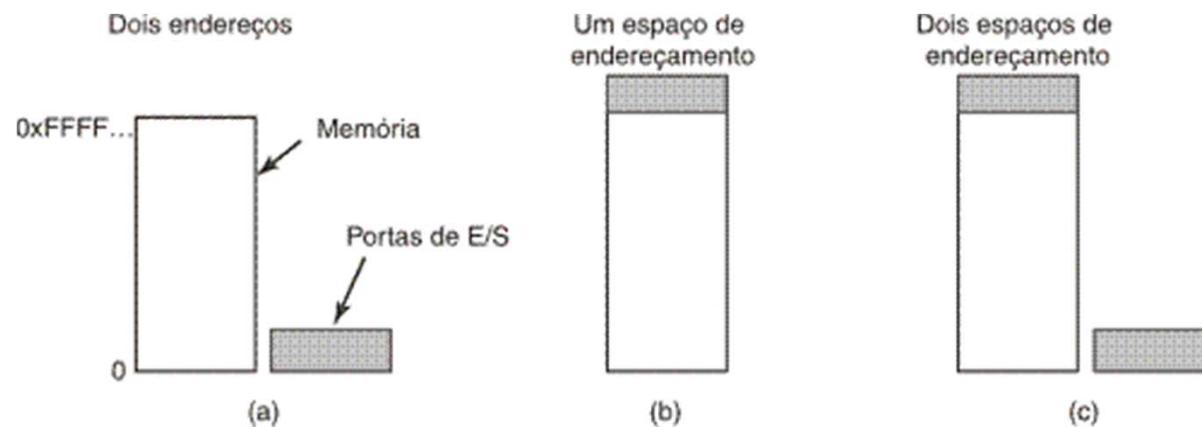
Taxas de dados típicas de dispositivos, redes e barramentos

Controladores de Dispositivos



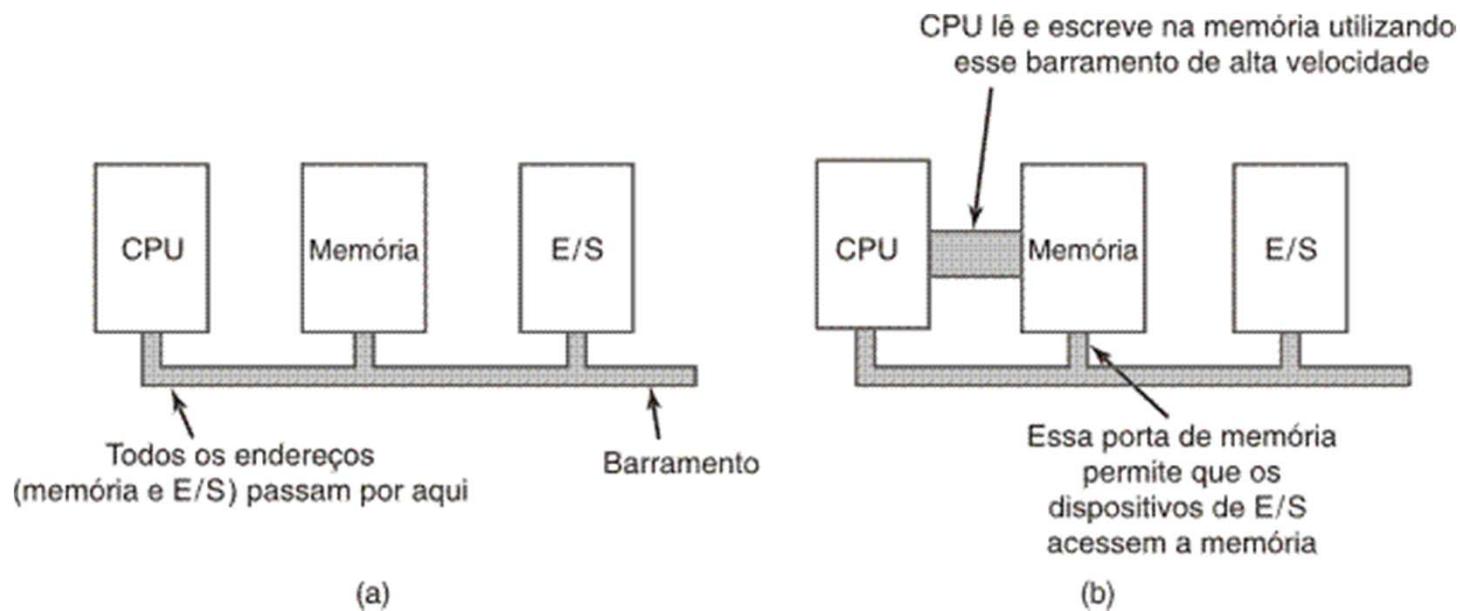
- Componentes de dispositivos de E/S
 - mecânico
 - eletrônico
- O componente eletrônico é o controlador do dispositivo
 - pode ser capaz de tratar múltiplos dispositivos
- Tarefas do controlador
 - converter fluxo serial de bits em bloco de bytes
 - executar toda correção de erro necessária
 - tornar o bloco disponível para ser copiado para a memória principal

E/S mapeada na memória (1)



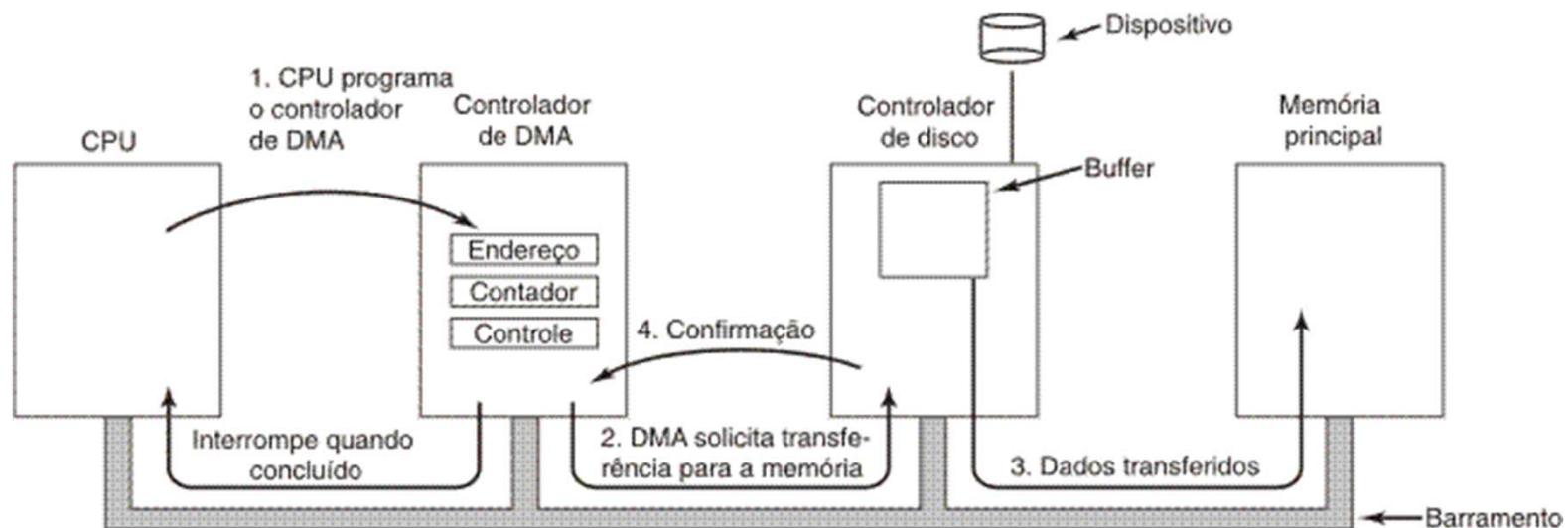
- a) Espaços de memória e E/S separados
- b) E/S mapeada na memória
- c) Híbrido

E/S mapeada na memória (2)



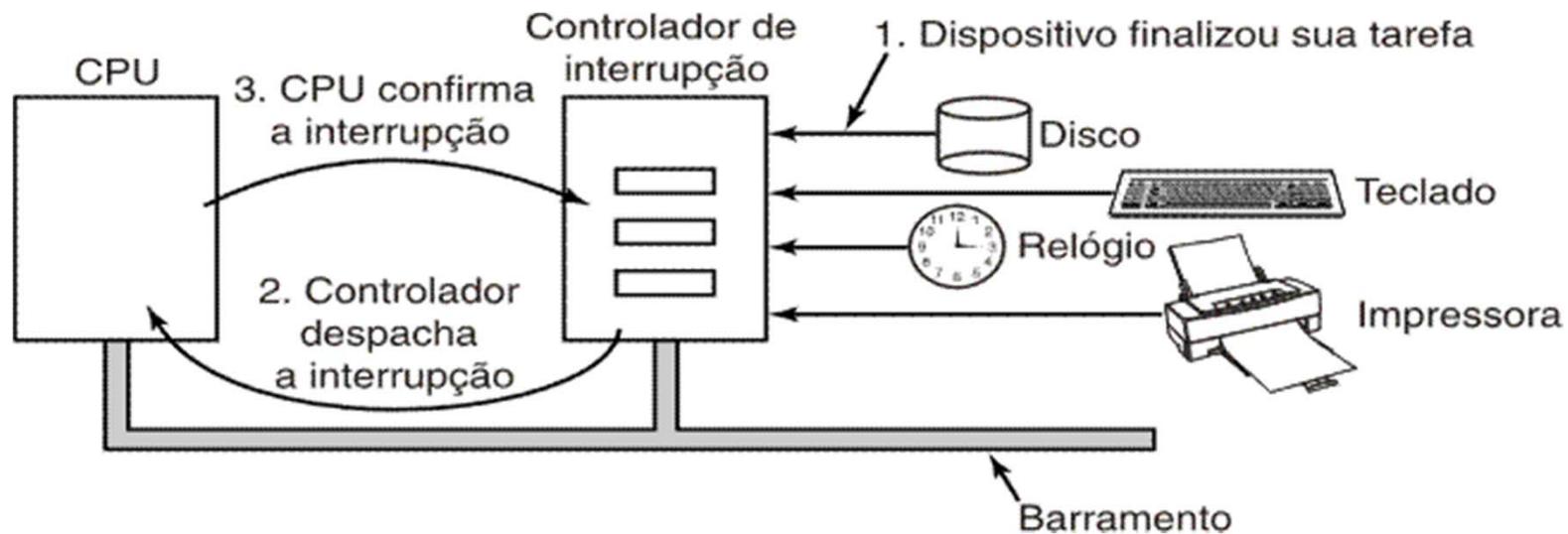
- (a) Arquitetura com barramento único
- (b) Arquitetura com barramento dual

Acesso Direto à Memória (DMA)



Operação de uma transferência com DMA

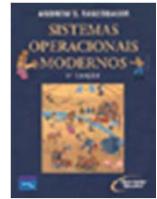
Interrupções Revisitadas



Como ocorre uma interrupção. Conexões entre dispositivos e controlador de interrupção usam linhas de interrupção no barramento em vez de fios dedicados

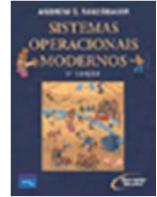
Princípios do Software de E/S

Objetivos do Software de E/S (1)



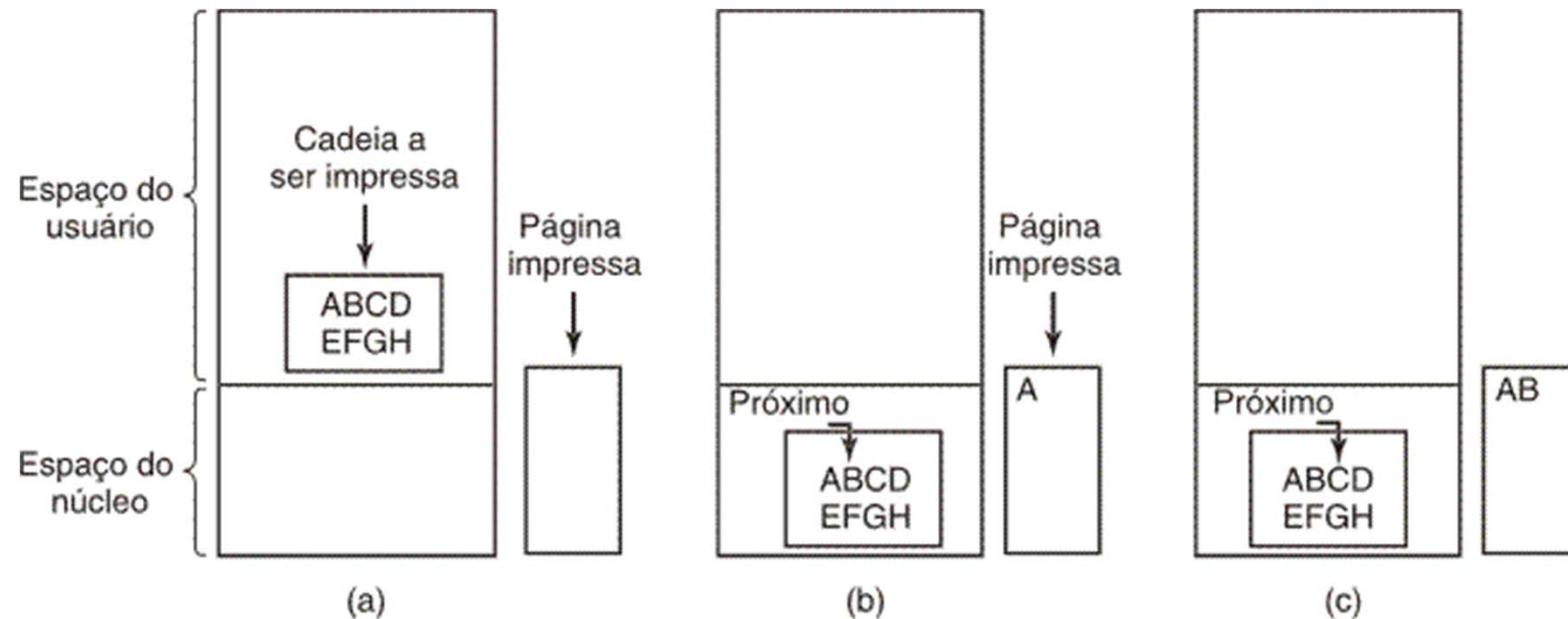
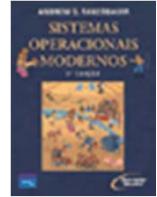
- Independência de dispositivo
 - Programas podem acessar qualquer dispositivo de E/S sem especificar previamente qual (disquete, disco rígido ou CD-ROM)
- Nomeação uniforme
 - Nome de um arquivo ou dispositivo pode ser uma cadeia de caracteres ou um número inteiro que é independente do dispositivo
- Tratamento de erro
 - Trata o mais próximo possível do hardware

Objetivos do Software de E/S (2)



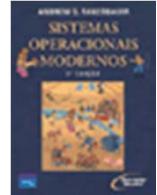
- Transferências Síncronas vs. Assíncronas
 - transferências bloqueantes vs. orientadas a interrupção
 - utilização de buffer para armazenamento temporário
 - dados provenientes de um dispositivo muitas vezes não podem ser armazenados diretamente em seu destino final
- Dispositivos Compartilháveis vs. Dedicados
 - discos são compartilháveis
 - unidades de fita não são

E/S Programada (1)



Passos da impressão de uma cadeia de caracteres

E/S Programada (2)

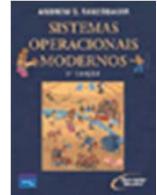


```
copy_from_user(buffer, p, cont);
for (i=0; i < count; i++) {
    while (*printer_status_reg !=READY) ;
    *printer_data_register = p[i];
}
return_to_user();
```

/* p é o buffer do núcleo */
/* executa o laço para cada caractere */
/* executa o laço até PRONTO */
/* envia um caractere para a saída */

Escrita de uma cadeia de caracteres para a impressora usando E/S programada

E/S Orientada à Interrupção



```
copy_from_user(buffer, p, count);
enable_interrupts();
while (*printer_status_reg != READY) ;
*printer_data_register = p[0];
scheduler();
```

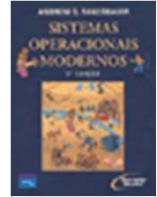
(a)

```
if (count == 0) {
    unblock_user();
} else {
    *printer_data_register = p[i];
    count = count - 1;
    i = i + 1;
}
acknowledge_interrupt();
return_from_interrupt();
```

(b)

- Escrita de uma cadeia de caracteres para a impressora usando E/S orientada à interrupção
 - a) Código executado quando é feita a chamada ao sistema para impressão
 - b) Rotina de tratamento de interrupção

E/S Usando DMA



```
copy_from_user(buffer, p, count);  
set_up_DMA_controller();  
scheduler();
```

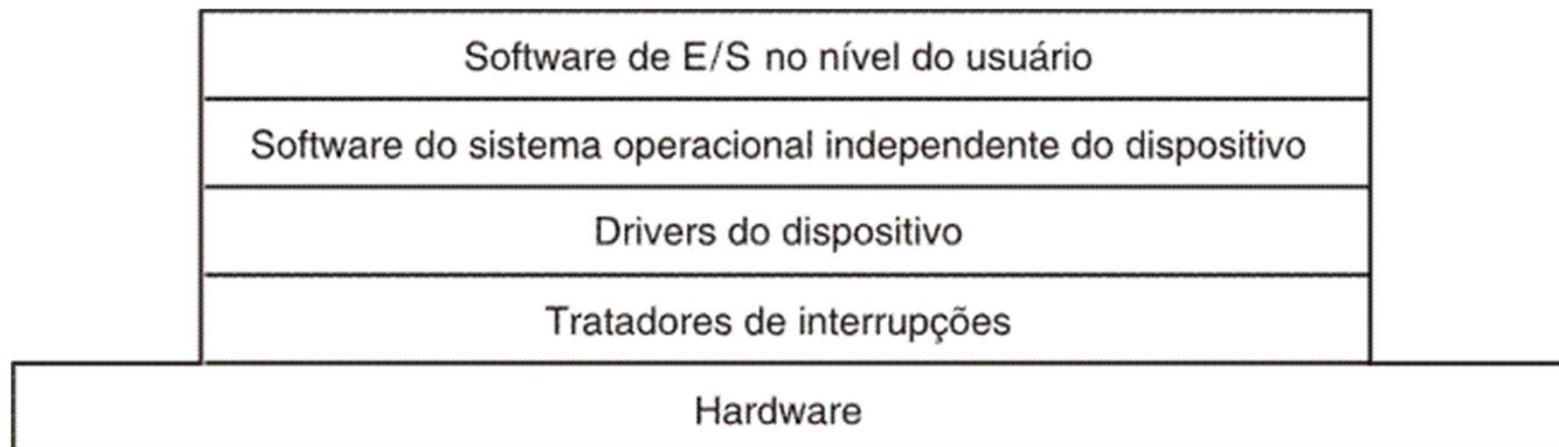
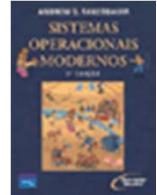
(a)

```
acknowledge_interrupt();  
unblock_user();  
return_from_interrupt();
```

(b)

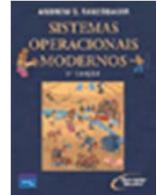
- Impressão de uma cadeia de caracteres usando DMA
 - a) Código executado quando é feita a chamada ao sistema para impressão
 - b) Rotina de tratamento de interrupção

Camadas do Software de E/S



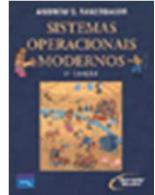
Camadas do sistema de software de E/S

Tratadores de Interrupção (1)



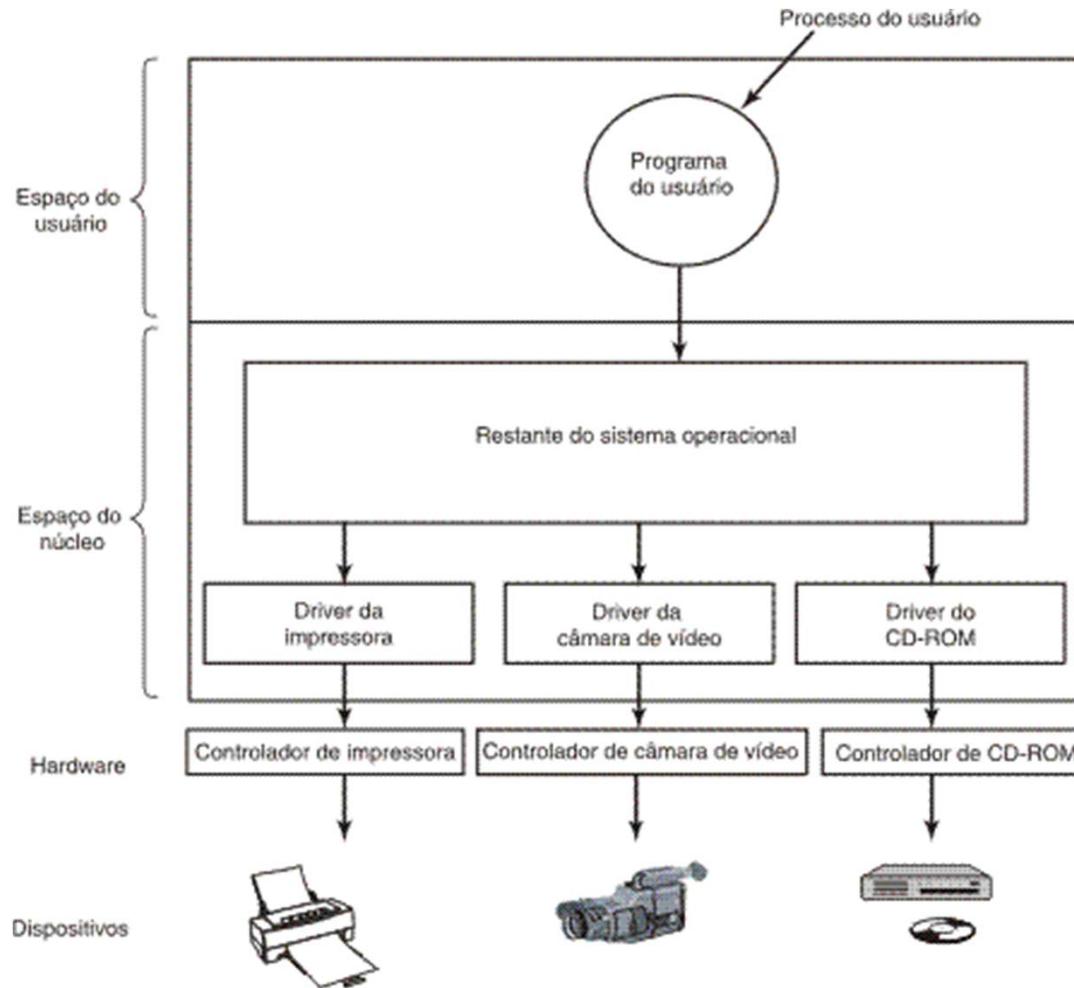
- As interrupções devem ser escondidas o máximo possível
 - uma forma de fazer isso é bloqueando o driver que iniciou uma operação de E/S até que uma interrupção notifique que a E/S foi completada
- Rotina de tratamento de interrupção cumpre sua tarefa
 - e então desbloqueia o driver que a chamou

Tratadores de Interrupção (2)



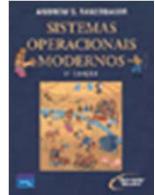
- **Passos que devem ser executados em software depois da interrupção ter sido concluída**
 1. salva registradores que ainda não foram salvos pelo hardware de interrupção
 2. estabelece contexto para rotina de tratamento de interrupção
 3. estabelece uma pilha para a rotina de tratamento de interrupção
 4. sinaliza o controlador de interrupção, reabilita as interrupções
 5. copia os registradores de onde eles foram salvos
 6. executa rotina de tratamento de interrupção
 7. escolhe o próximo processo a executar
 8. estabelece o contexto da MMU para o próximo processo a executar
 9. carrega os registradores do novo processo
 10. começa a executar o novo processo

Drivers dos Dispositivos



- Posição lógica dos drivers dos dispositivos
- A comunicação entre os drivers e os controladores de dispositivos é feita por meio do barramento

Software de E/S Independente de Dispositivo (1)



Interface uniforme para os drivers dos dispositivos

Armazenamento em buffer

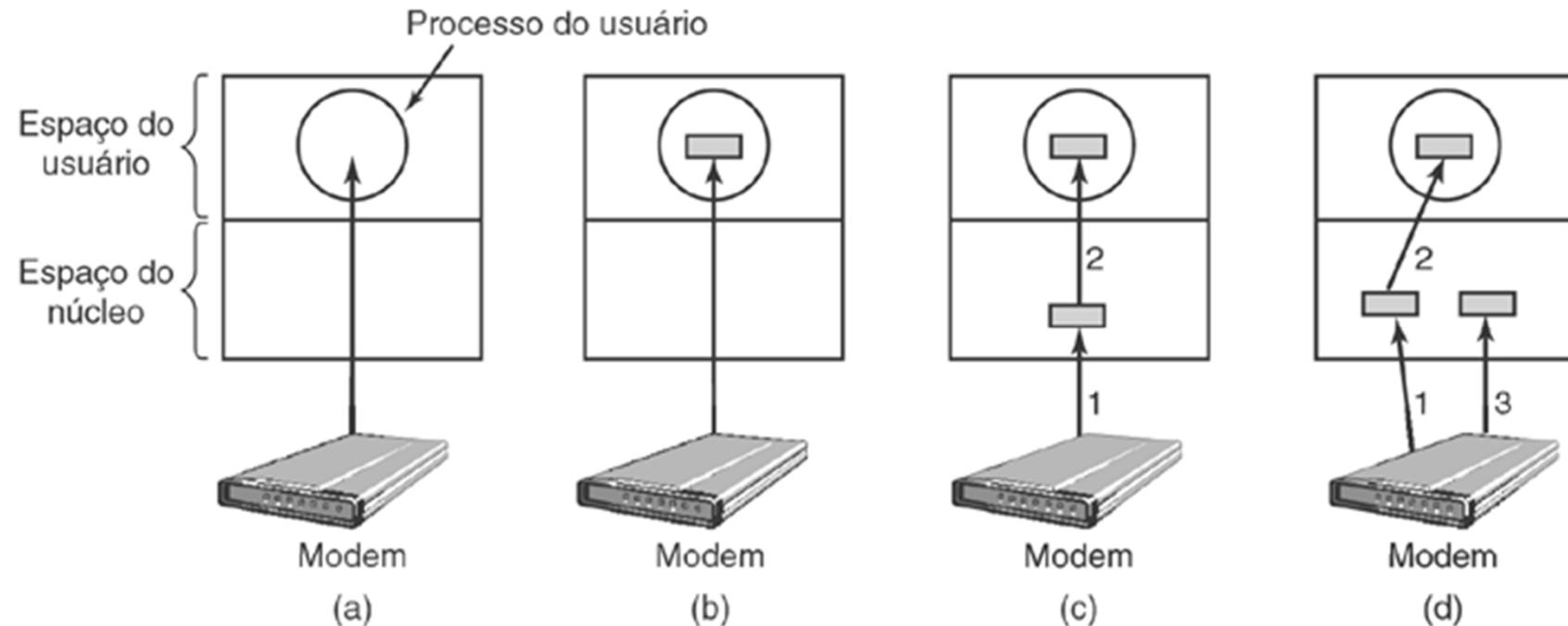
Relatório dos erros

Alocação e liberação de dispositivos dedicados

Fornecimento de tamanho de bloco independente
de dispositivo

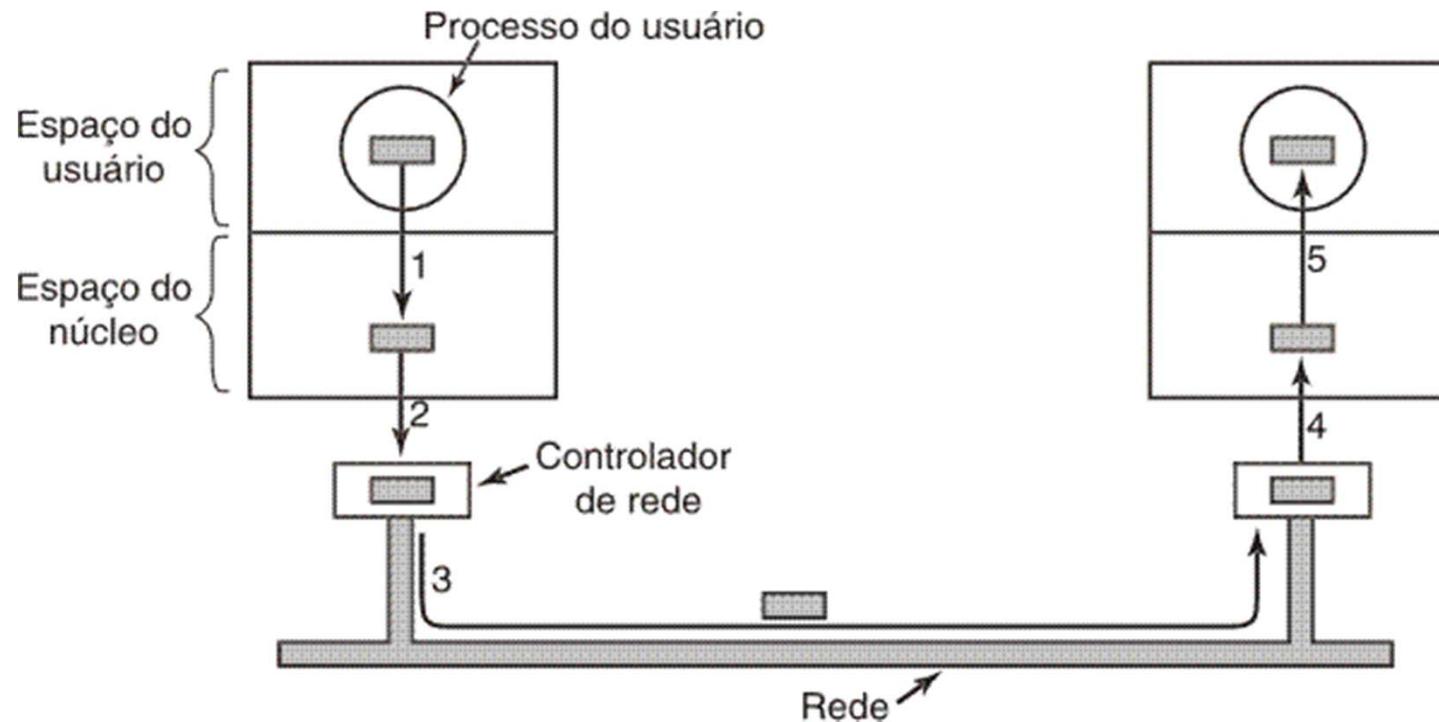
Funções do software de E/S independente de dispositivo

Software de E/S Independente de Dispositivo (3)



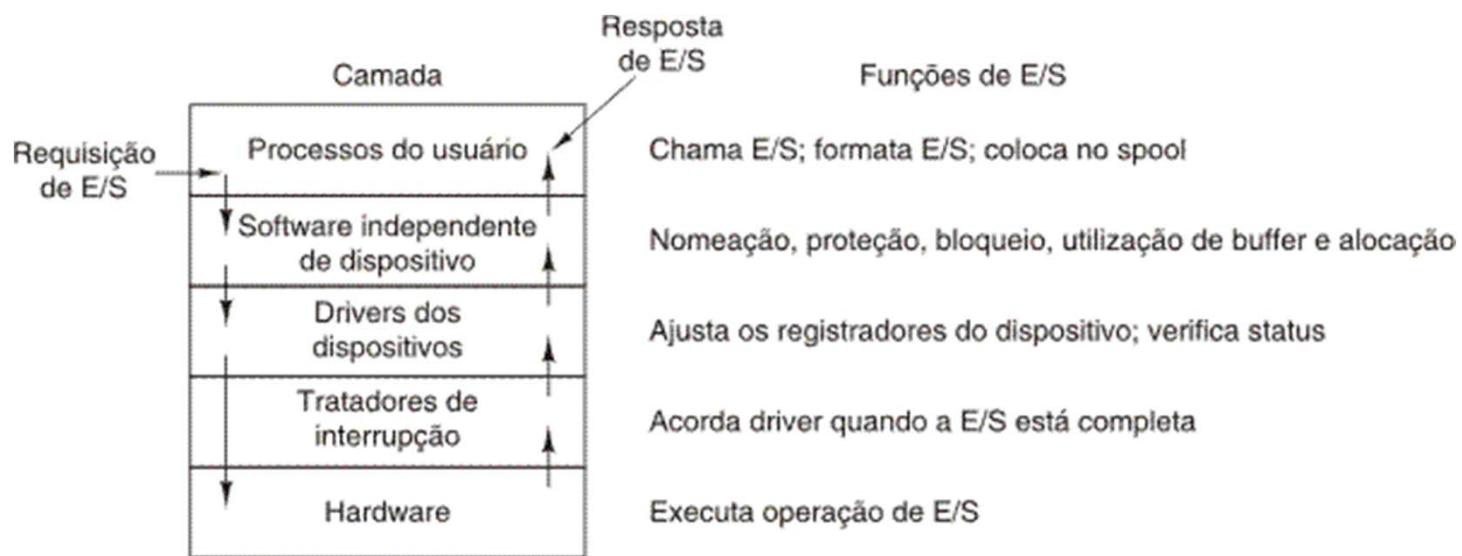
- Entrada sem utilização de buffer
- Utilização de buffer no espaço do usuário
- Utilização de buffer no núcleo seguido de cópia para o espaço do usuário
- Utilização de buffer duplo no núcleo

Software de E/S Independente de Dispositivo (4)



A operação em rede pode envolver muitas cópias de um pacote

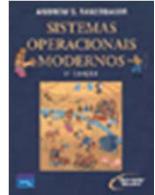
Software de E/S no Espaço do Usuário



Camadas do sistema de E/S e as principais funções de cada camada

Discos

Hardware do Disco (1)

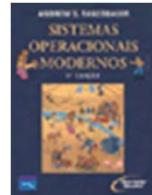


| Parâmetro | Disco flexível IBM 360 KB | Disco rígido WD 18300 |
|--|---------------------------|-----------------------|
| Número de cilindros | 40 | 10 601 |
| Trilhas por cilindro | 2 | 12 |
| Setores por trilha | 9 | 281 (avg) |
| Setores por disco | 720 | 35 742 000 |
| Bytes por setor | 512 | 512 |
| Capacidade do disco | 360 KB | 18,3 GB |
| Tempo de posicionamento (cilindros adjacentes) | 6 ms | 0,8 ms |
| Tempo de posicionamento (caso médio) | 77 ms | 6,9 ms |
| Tempo de rotação | 200 ms | 8,33 ms |
| Tempo de pára/inicia do motor | 250 ms | 20 s |
| Tempo de transferência para um setor | 22 ms | 17 µs |

Parâmetros de disco para o disco flexível original do IBM PC e o disco rígido da Western Digital WD 18300

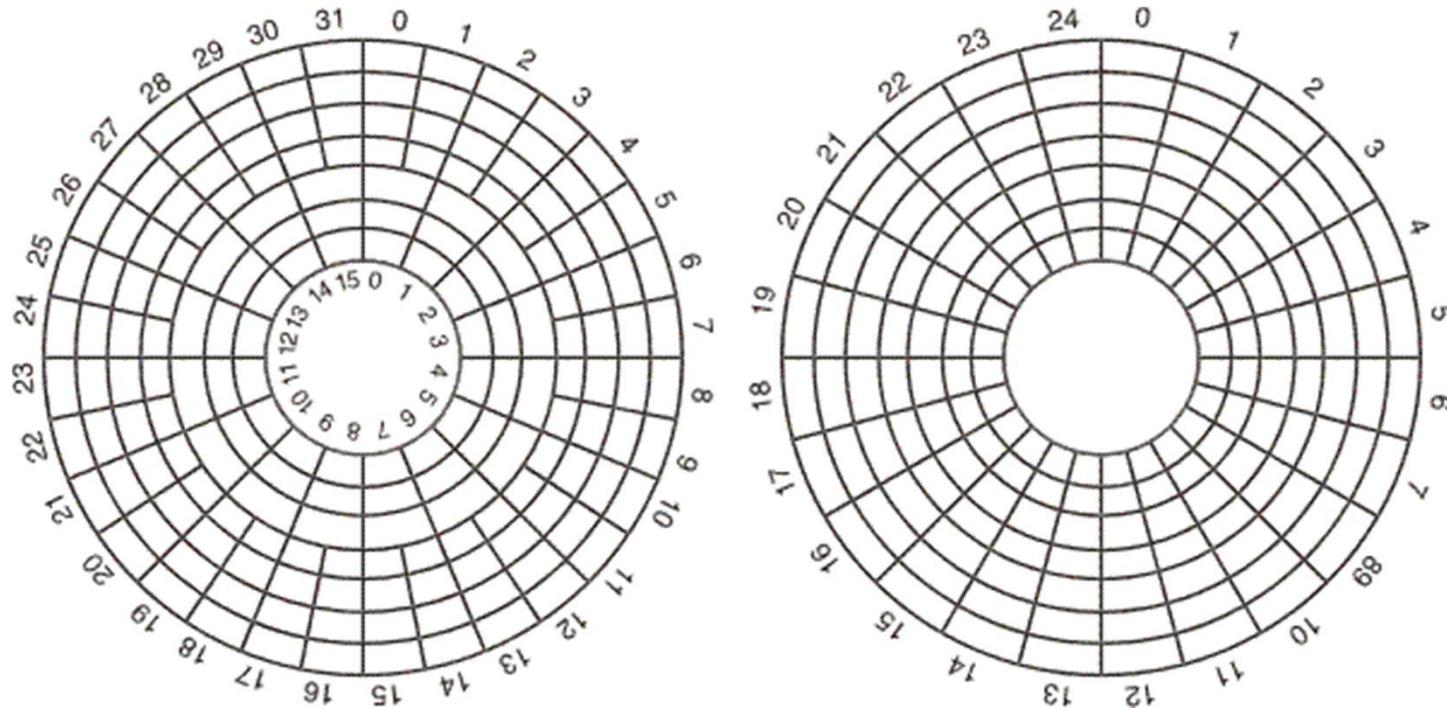
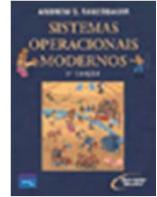
Discos

Hardware do Disco (1)



| Especificações | 4 TB | 3 TB | 2 TB | 1 TB | 500 GB |
|--|------------------------|------------------------|------------------------|------------------------|------------------------|
| Número do modelo | WD4001FAEX | WD3001FAEX | WD2002FAEX | WD1002FAEX | WD5003AZEX |
| Interface | SATA 6 Gb/s |
| Capacidade formatada ¹ | 4.000.787 MB | 3.000.592 MB | 2.000.398 MB | 1.000.204 MB | 500.107 MB |
| Setores de usuários por unidade | 7.814.037.168 | 5.860.533.168 | 3.907.029.168 | 1.953.525.169 | 976.773.168 |
| Fator de forma | 3,5 polegadas (6,3 cm) |
| Formato Avançado (FA) | Sim | Sim | Sim | Sim | Sim |
| Cumpre as normas RoHS ² | Sim | Sim | Sim | Sim | Sim |
| Desempenho | | | | | |
| Taxa de transferência de dados (máx) | | | | | |
| Buffer para host | 6 Gb/s |
| Host para/desde drive (mantido) | 154 MB/s | 154 MB/s | 138 MB/s | 126 MB/s | 150 MB/s |
| Cachê (MB) | 64 | 64 | 64 | 64 | 64 |
| Velocidade de rotação (RPM) | 7200 | 7200 | 7200 | 7200 | 7200 |
| Confiabilidade/Integridade dos dados | | | | | |
| Ciclos de carga/descarga ³ | 300.000 | 300.000 | 300.000 | 300.000 | 300.000 |
| Erros de leitura irrecuperáveis por bits lidos | <1 em 10 ¹⁴ |
| Garantia limitada (anos) ⁴ | 5 | 5 | 5 | 5 | 5 |
| Gerenciamento de energia | | | | | |
| Requisitos médios de energia (W) | | | | | |
| Leitura/gravação | 10,4 | 10,4 | 10,7 | 6,8 | 6,8 |
| Ociooso | 8,1 | 8,1 | 8,2 | 6,1 | 6,1 |
| Standby/Descanso | 1,2 | 1,2 | 1,3 | 0,7 | 0,8 |
| Especificações ambientais⁵ | | | | | |
| Temperatura (°C) | | | | | |
| Operacional | 0 a 60 |
| Não-operacional | -40 a 70 |
| Choque (Gs) | | | | | |
| Operacional (2ms, gravação) | 30 | 30 | 30 | 30 | 30 |
| Operacional (2ms, leitura) | 65 | 65 | 65 | 65 | 65 |
| Não operacional (2ms) | 300 | 300 | 300 | 300 | 350 |
| Acústica (dBA) ⁶ | | | | | |
| Ociooso | 29 | 29 | 29 | 28 | 29 |
| Pesquisa (média) | 34 | 34 | 34 | 33 | 30 |
| Dimensões físicas | | | | | |
| Altura (pol./mm, máx) | 1,028/25,4 | 1,028/25,4 | 1,028/25,4 | 1,028/25,4 | 1,028/25,4 |
| Comprimento (pol./mm, máx) | 5,787/147 | 5,787/147 | 5,787/147 | 5,787/147 | 5,787/147 |
| Largura (pol./mm, ± 0,01 pol.) | 4/101,6 | 4/101,6 | 4/101,6 | 4/101,6 | 4/101,6 |
| Peso (lb/kg, ± 10%) | 1,72/0,78 | 1,72/0,78 | 1,66/0,75 | 1,52/0,69 | 0,97/0,44 |

Hardware do Disco (2)



- Geometria física de um disco com duas zonas
- Uma possível geometria virtual para esse disco

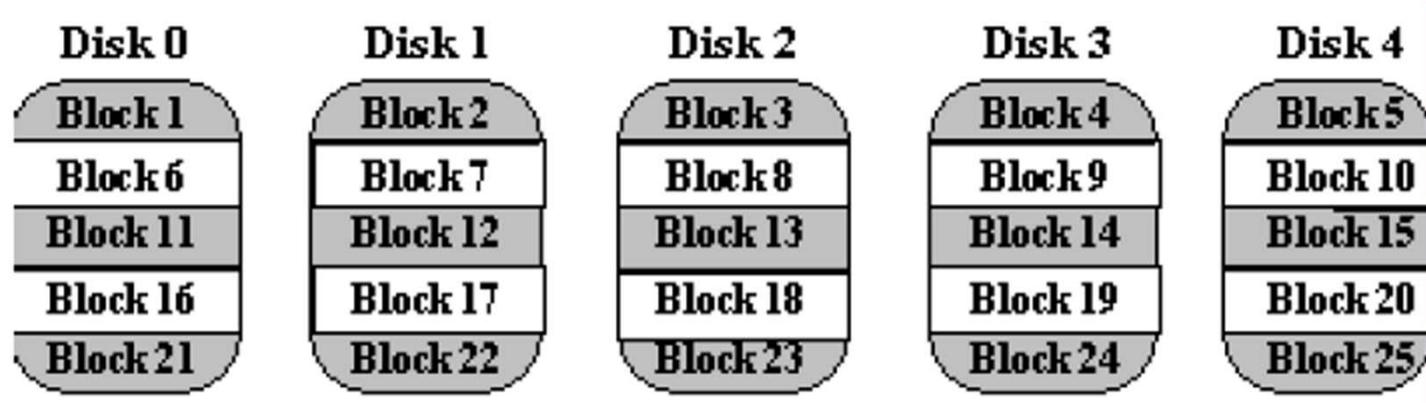
Hardware do Disco (3)

- Embora presente hoje também em discos IDE, a técnica RAID foi utilizada inicialmente em discos SCSI
- **RAID = Redundant Array of Independent Disks** (Arranjo Redundante de Discos Independentes)
- Anteriormente, RAID significava **Redundant Array of Inexpensive Disks**
- Há vários tipos de RAID, cada qual com sua finalidade

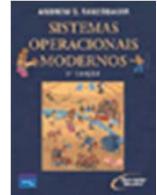


RAID 0

- Voltado para a melhoria do desempenho
- Dados escritos em seções seqüenciais dos discos
- Vários dispositivos acessados de uma só vez ☺
- Não é voltado para tolerância à falhas ☹

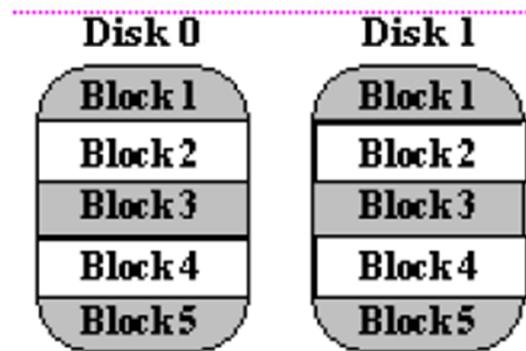


Hardware do Disco (3)

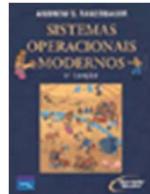


RAID 1

- Dados são escritos em um ou mais discos
- Também chamado de espelhamento
- Redundância provê tolerância a falhas ☺
- Desempenho ruim se comparado ao RAID 0 ☹

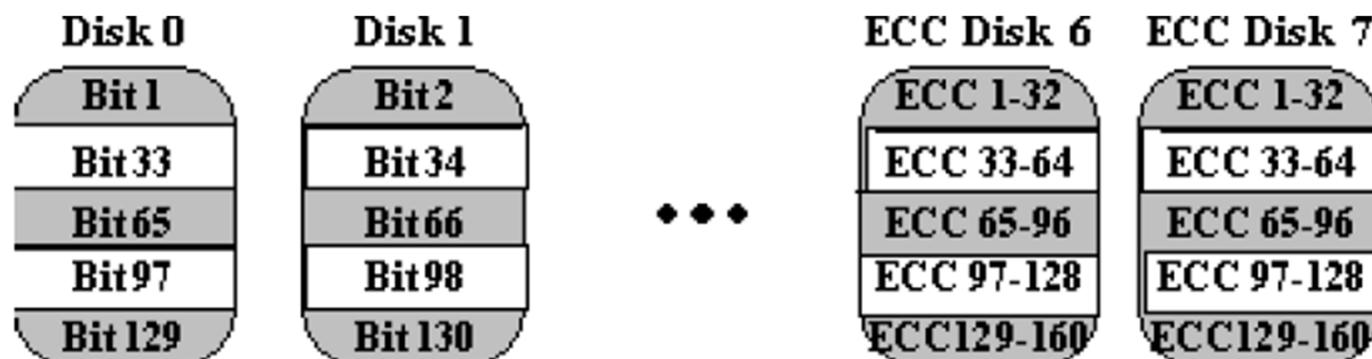


Hardware do Disco (3)



RAID 2

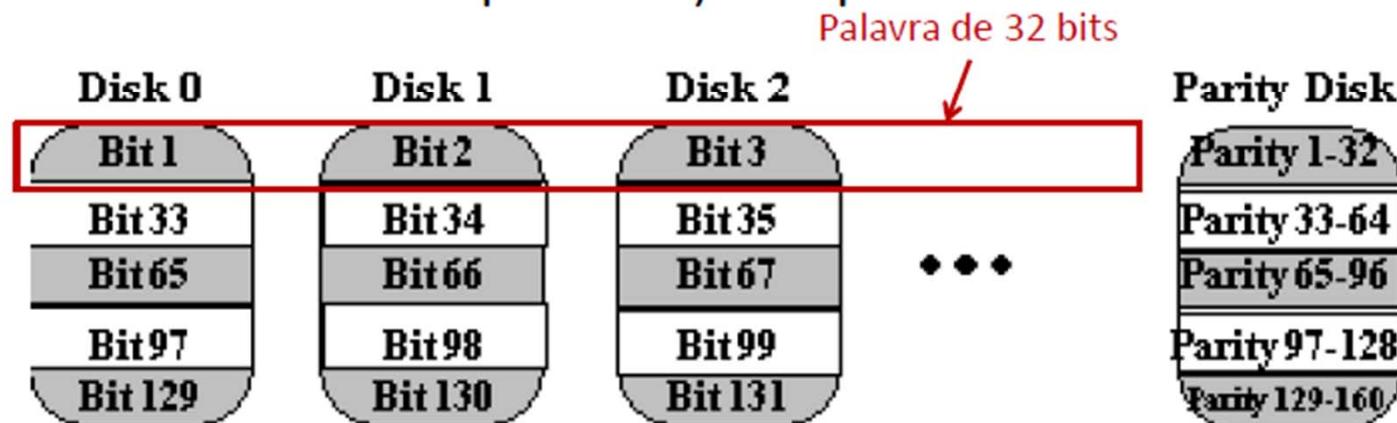
- Dados são escritos em seções seqüenciais dos discos, no nível de bit
- Discos extras contêm Códigos de Hamming, no nível de bit, para detecção e correção de erros
- São necessários vários discos de ECC ☹



Hardware do Disco (3)

RAID 3

- Dados são escritos em seções seqüenciais dos discos, no nível de bit
- Há apenas um disco extra, contendo um bit de paridade ☺
- Atendimento a solicitações simultâneas (leitura/escrita de duas ou mais palavras) é impossível ☹

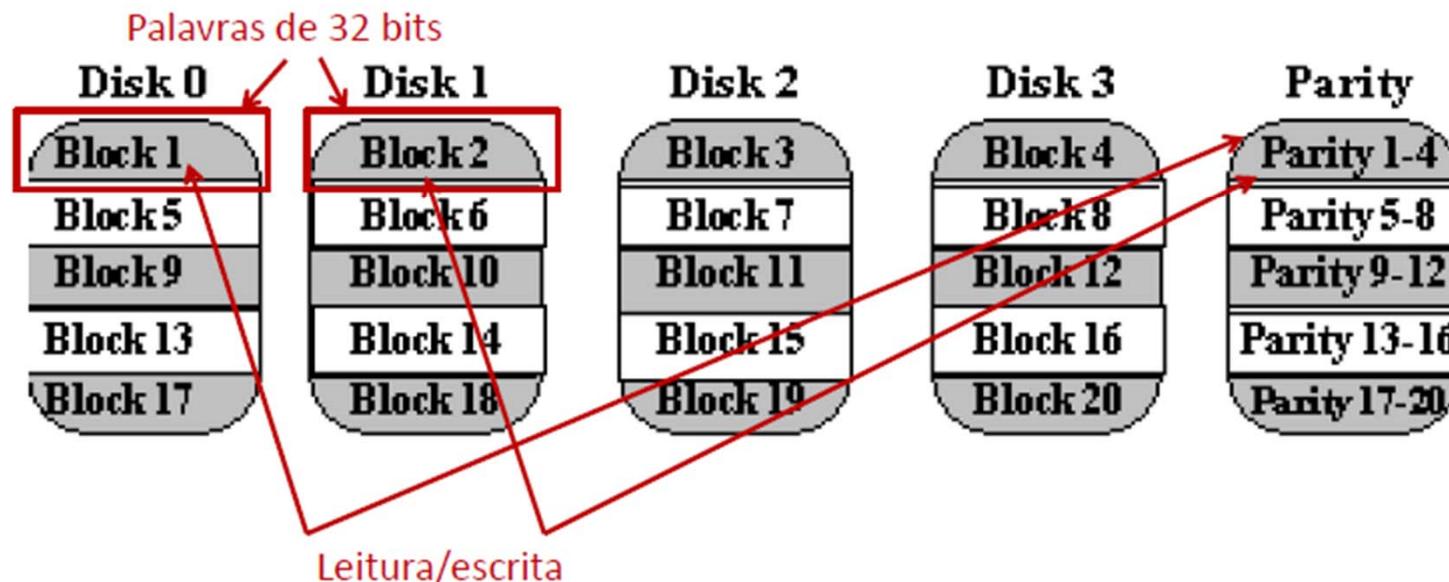


Hardware do Disco (3)



RAID 4

- Similar ao RAID 3, mas agora os discos de dados são organizados no nível de bloco
- Disco de paridade é um gargalo para atendimento a solicitações simultâneas ☹

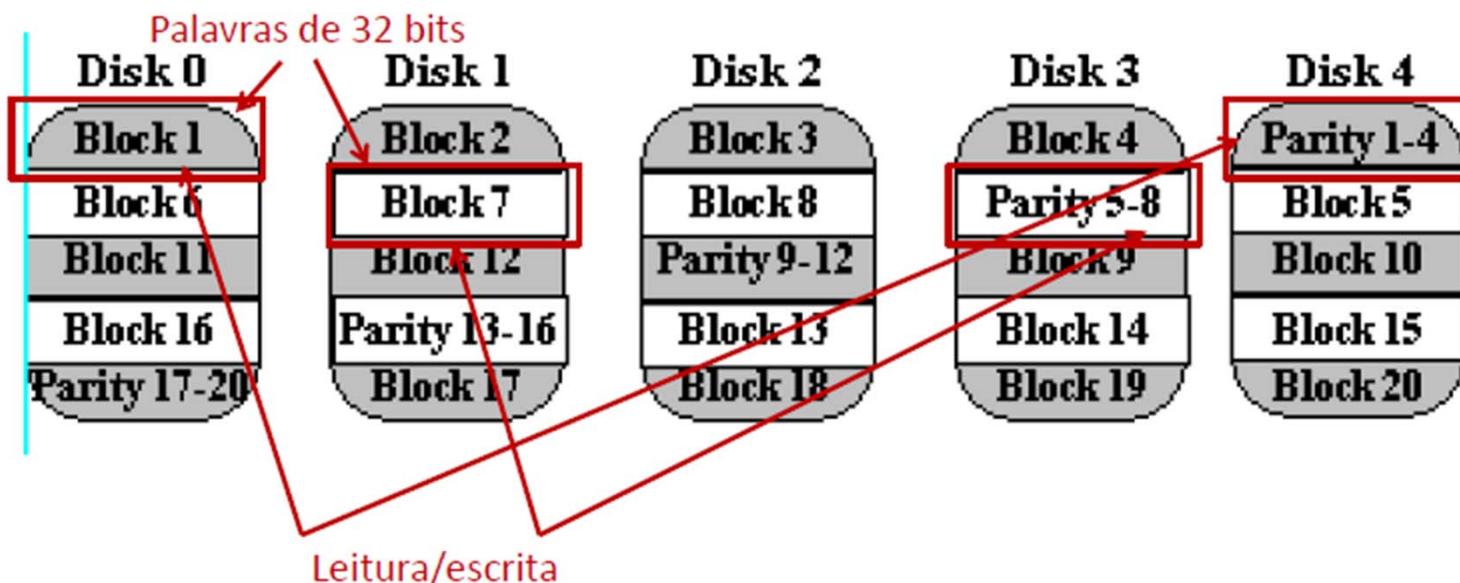


Hardware do Disco (3)



RAID 5

- Bits de paridade são distribuídos de maneira uniforme por todos os discos
- Diminui a ocorrência de gargalos no atendimento a solicitações simultâneas 😊

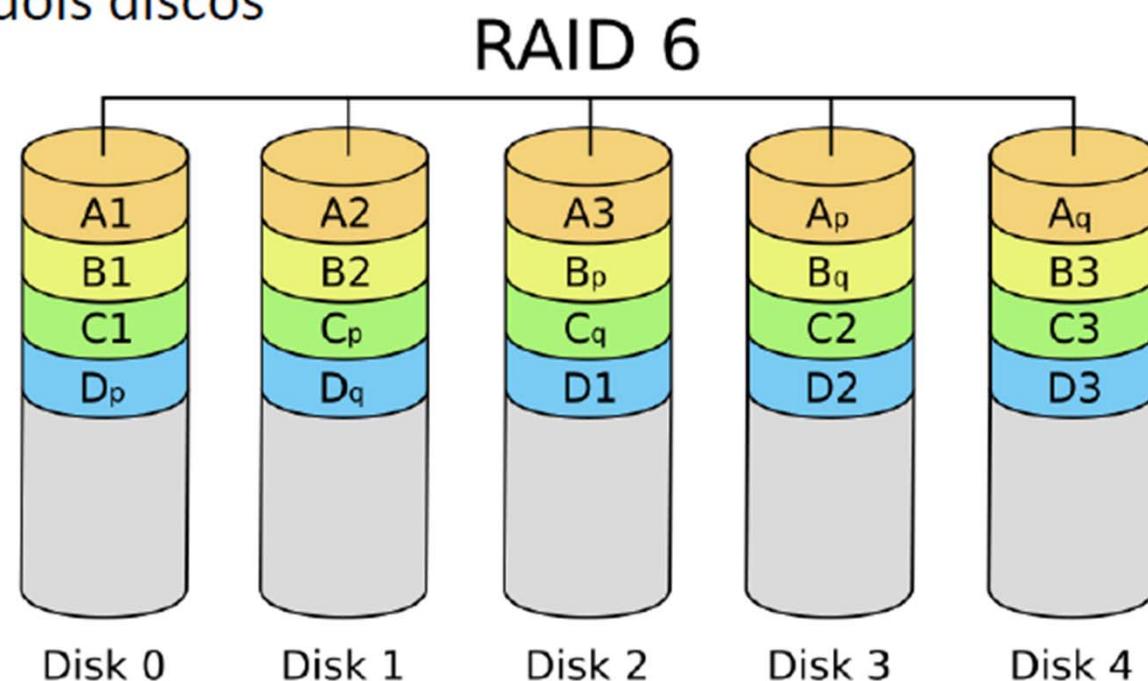


Hardware do Disco (3)



RAID 6

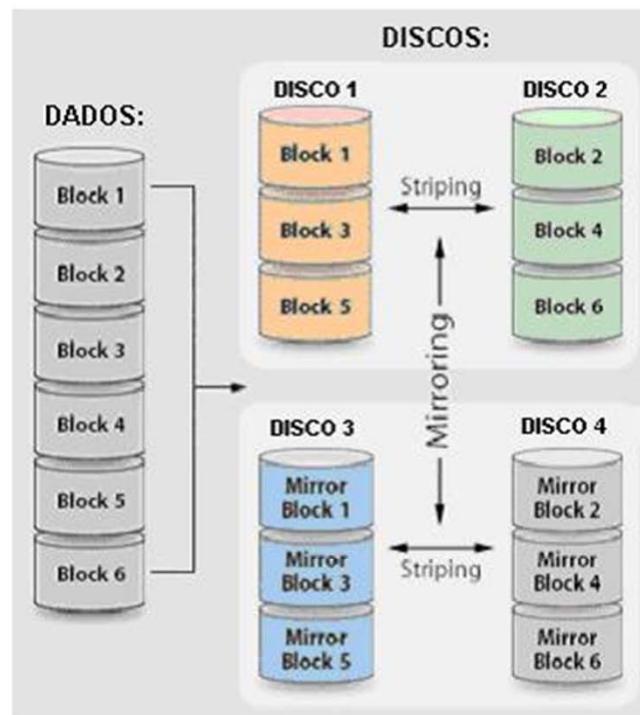
- Provê um bloco de paridade adicional
- Usa Códigos Reed-Solomon para proteção contra erros em dois discos



Hardware do Disco (3)

RAID 01 (ou RAID 0+1) – Striping e Mirroring

Em uma implementação RAID 0+1, os dados são espelhados através de grupos de discos segmentados, isto é, os dados são primeiramente segmentados e para cada segmento criados é feito um espelho, como demonstrado na figura abaixo:

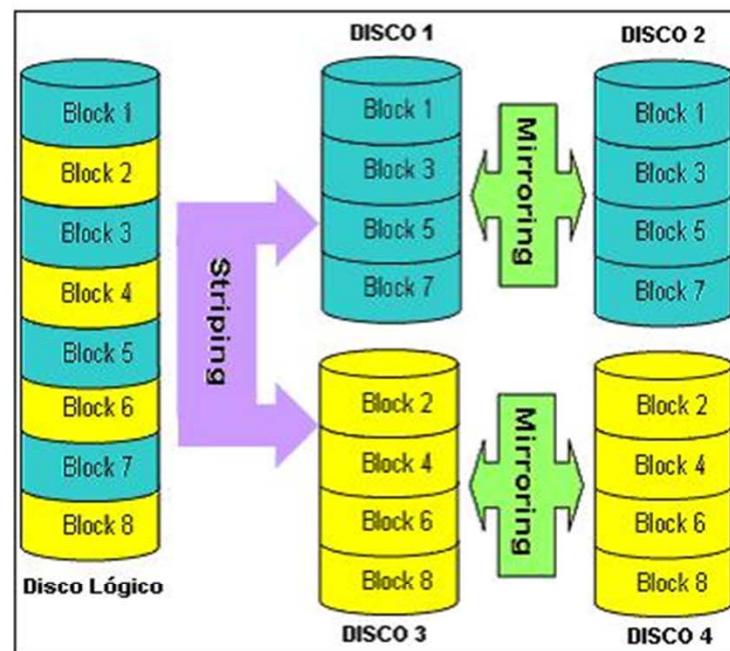


Na figura acima vemos que o discos 1 e 2 formam um RAID 0 sendo depois espelhados pelo discos 3 e 4 também em RAID 0, formando assim RAID 1 sobre RAID 0. Apesar de ser uma configuração que proporciona alta performance, se perdermos um disco em um dos lados, praticamente **teremos uma configuração em RAID 0**, porque em uma configuração RAID 0 se um disco falha todo o conjunto falhará. Neste caso, se o disco 1 falhar, então o disco 2 que está intacto ficará inutilizado, restando assim os discos 3 e 4 em RAID 0.

Hardware do Disco (3)

RAID 10 (ou RAID 1+0) - Mirroring e Striping

Em uma implementação RAID 1+0, os dados são segmentados através de grupos de discos espelhados, isto é, os dados são primeiro espelhados e para depois serem segmentados como demonstrado na figura abaixo:



Na figura acima vemos que o discos 1 e 2 formam um RAID 1 e os discos 3 e 4 também sendo após segmentados em RAID 0, formando assim RAID 0 sobre RAID 1. Além de ser uma configuração que proporciona o mesmo nível de performance proporcionado pelo RAID 01, o RAID 10 proporciona mais tolerância à falhas que o RAID 01 porque poderíamos ter uma falha simultânea dos discos 1 e 3 e ainda assim o conjunto estaria intacto, pois teríamos os espelhos em perfeito funcionamento. No meu ponto de vista, este conjunto é o mais indicado nos casos onde necessitamos aliar performance e redundância, como é o caso, por exemplo, de bancos de dados Oracle de alta performance.

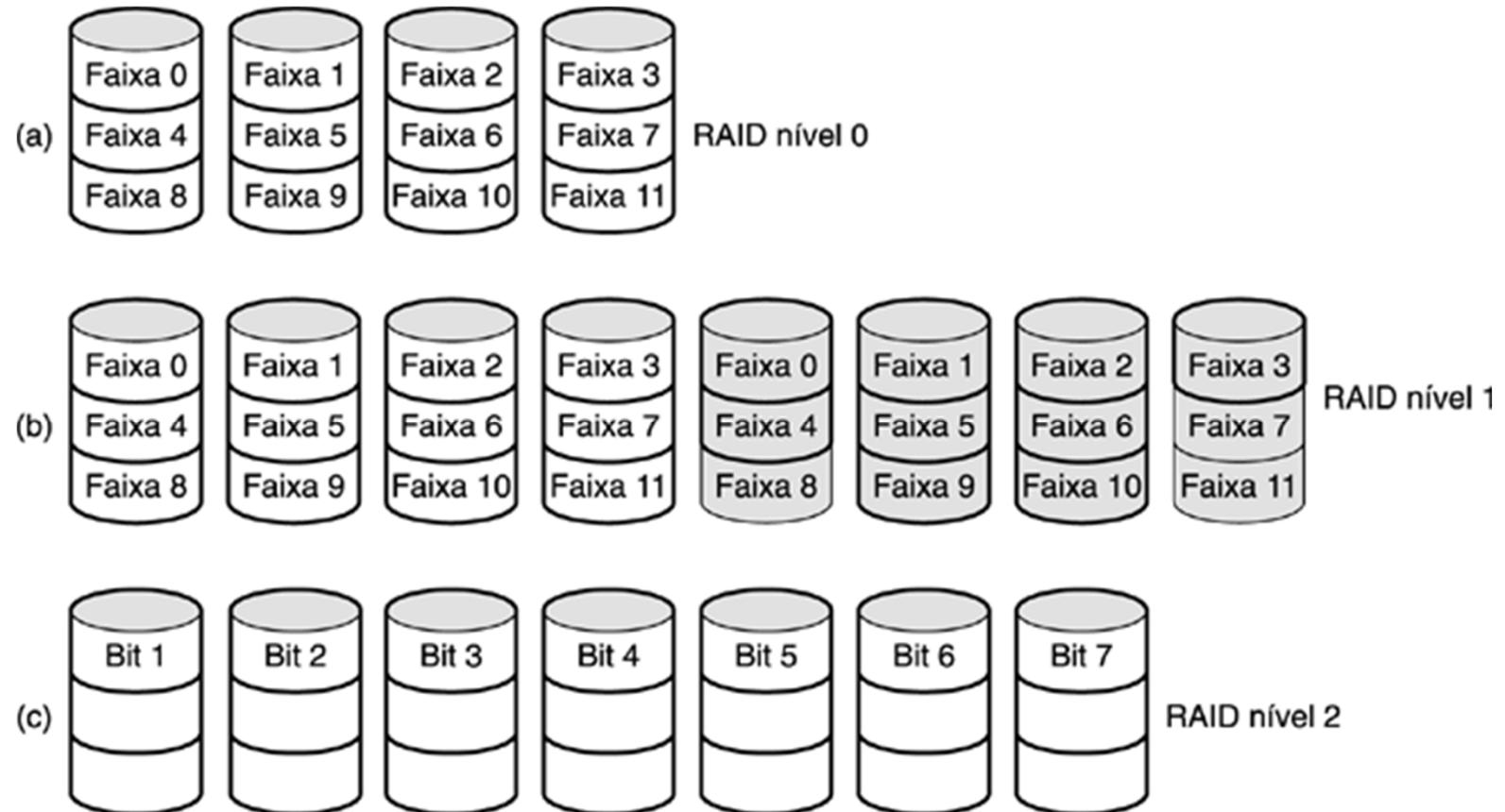
Hardware do Disco (3)

RAID

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

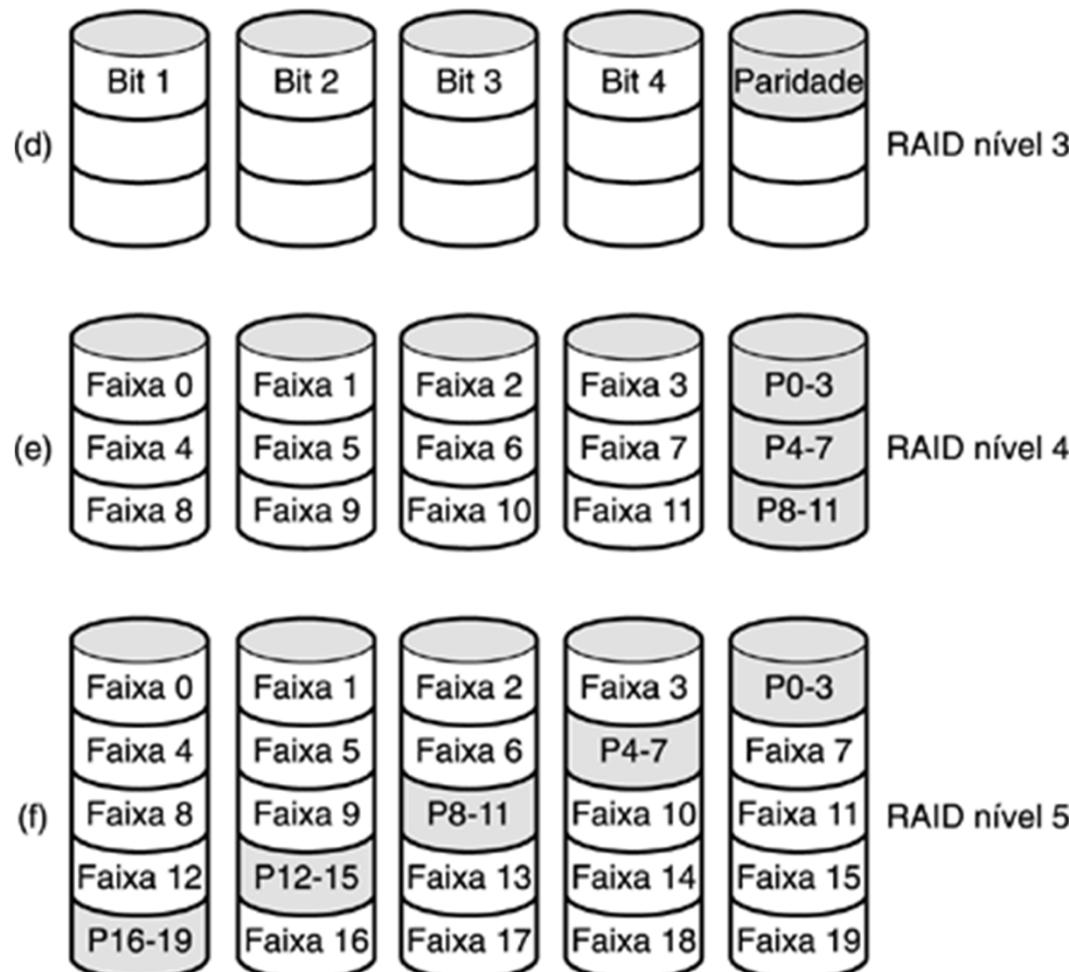
| RAID level | Min disks | Available storage capacity (%) | Read performance | Write performance | Write penalty | Protection |
|------------|-----------|--------------------------------|---|---|---------------|--|
| 1 | 2 | 50 | Better than single disk | Slower than single disk, because every write 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) |

Hardware do Disco (3)



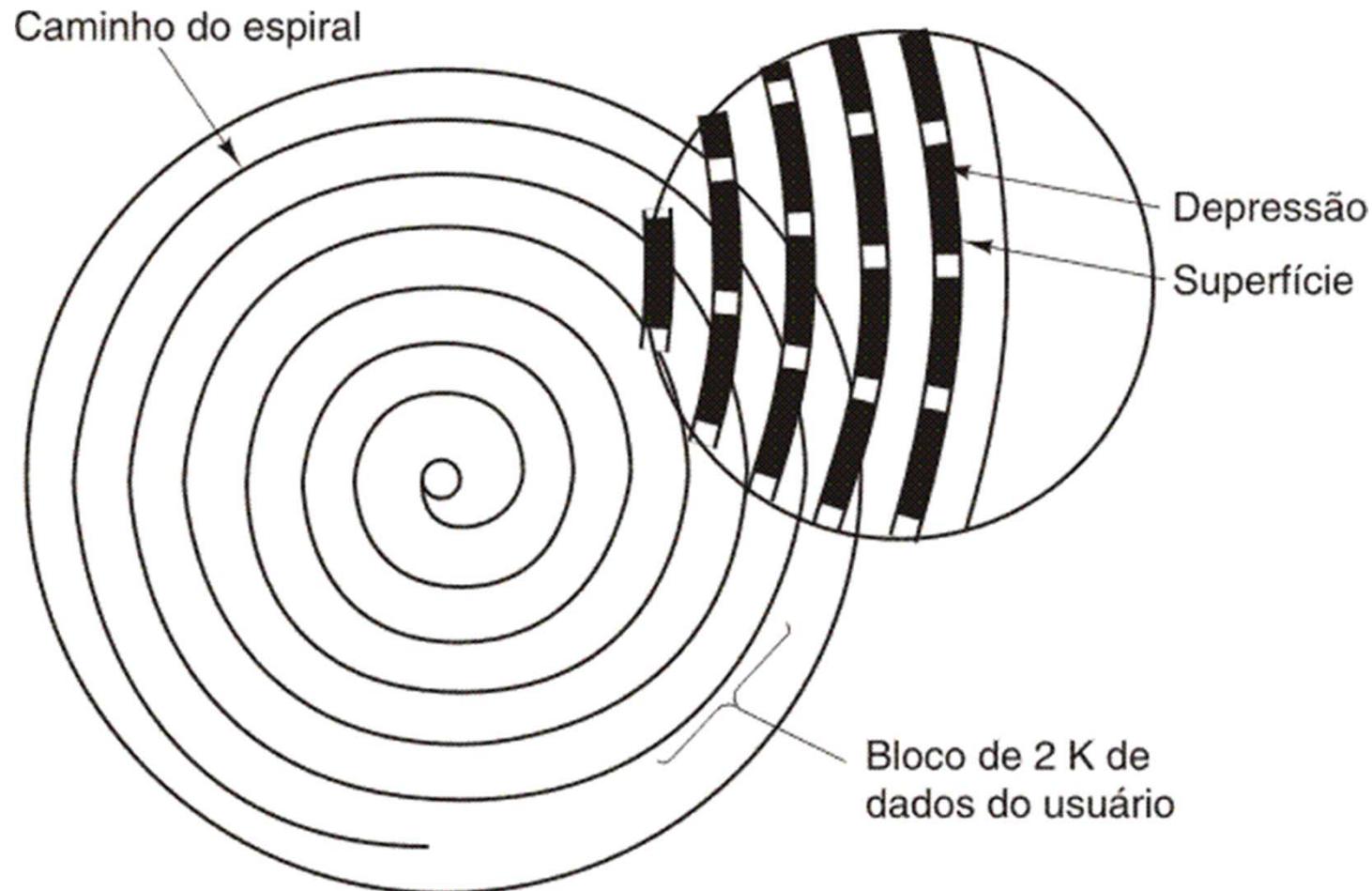
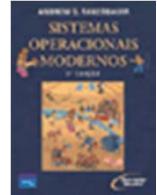
- RAIDs níveis 0 a 2
- Discos de segurança e de paridade são os sombreados

Hardware do Disco (4)



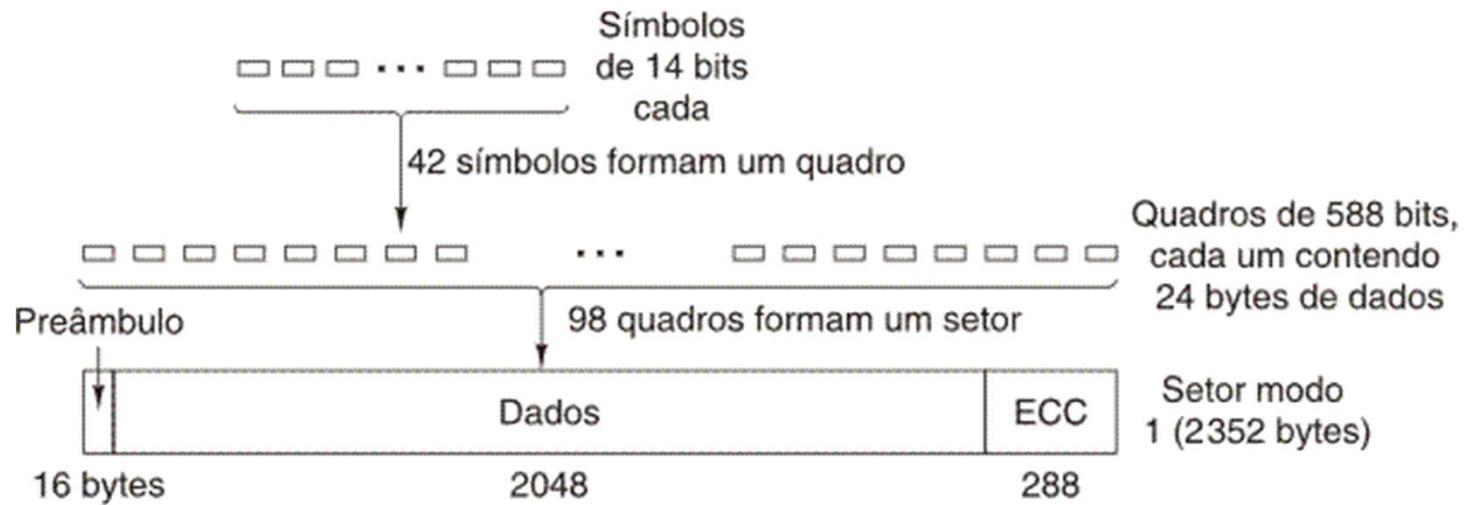
- RAIDs níveis 3 a 5
- Discos de segurança e de paridade são os sombreados

Hardware do Disco (5)



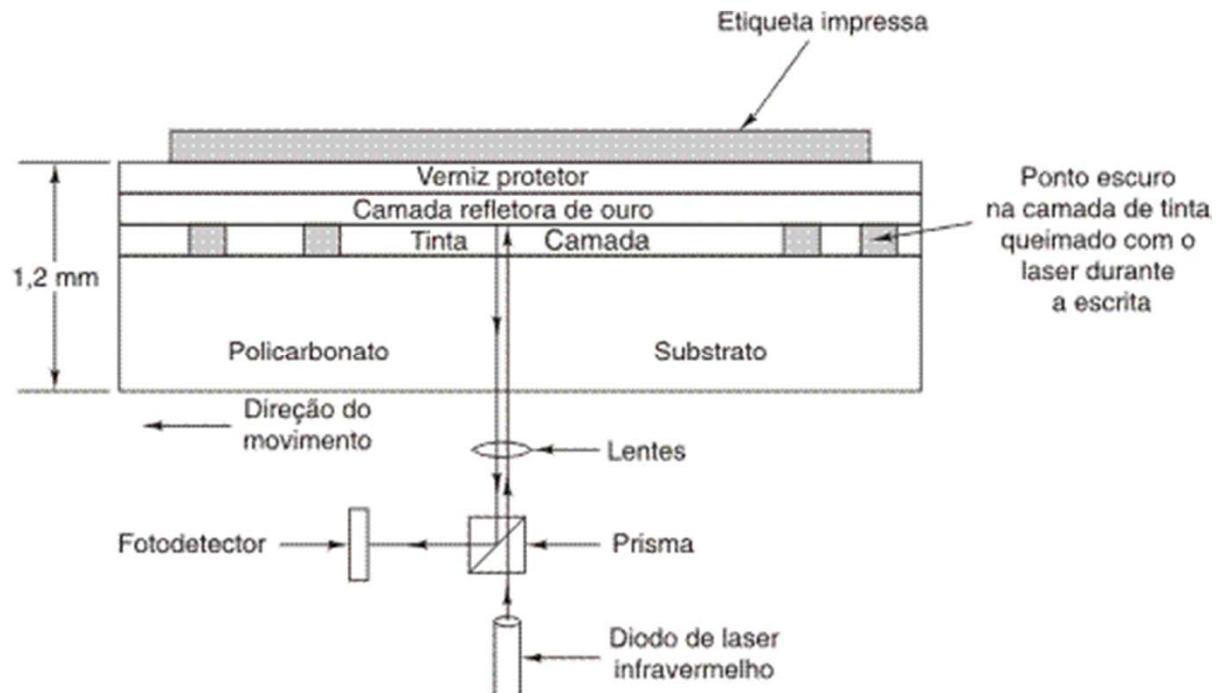
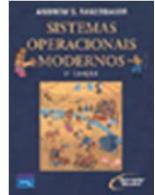
Estrutura de gravação de um CD ou CD-ROM

Hardware do Disco (6)



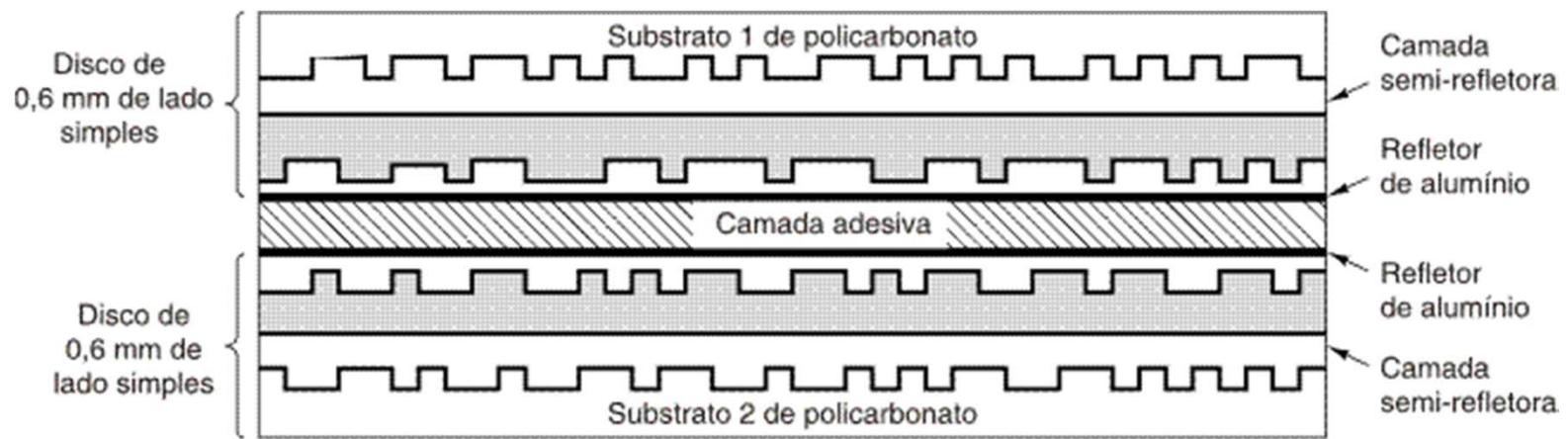
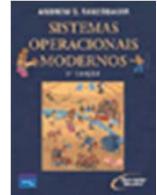
Esquema lógico dos dados em um CD-ROM

Hardware do Disco (7)



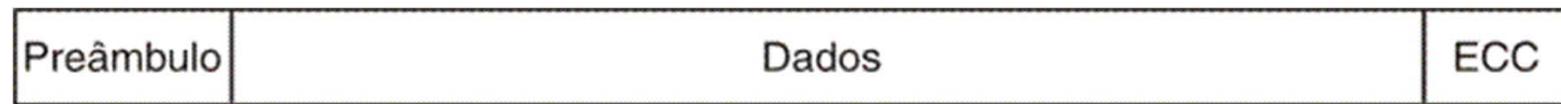
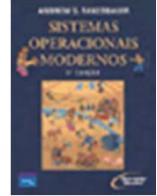
- Secção transversal de um disco CD-R e laser
 - sem escala
- CD-ROM prateado tem estrutura similar
 - sem camada de tinta
 - com camada de alumínio em vez de ouro

Hardware do Disco (8)



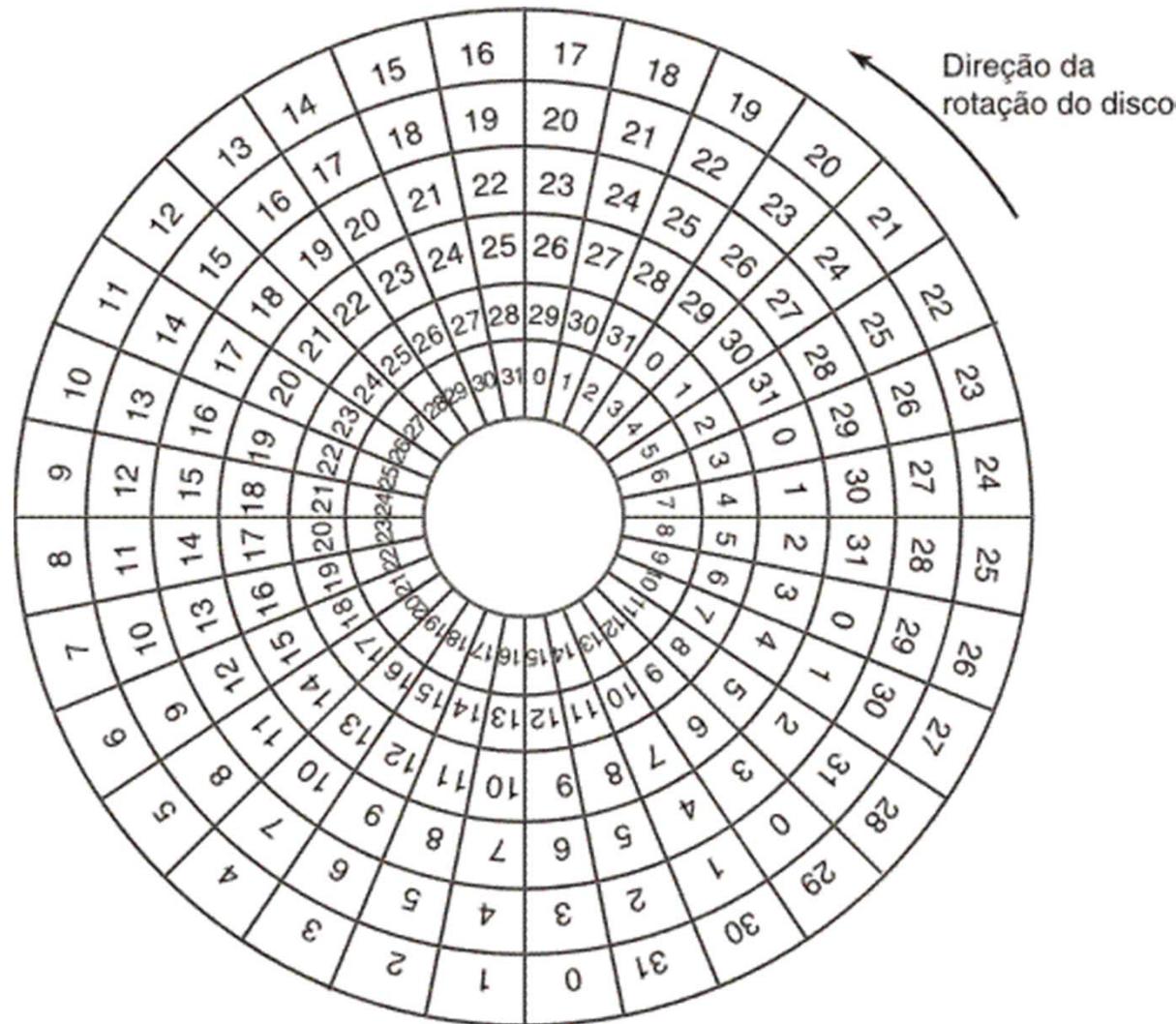
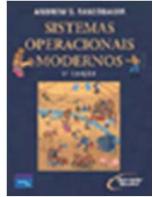
Disco DVD com lado duplo e camada dupla

Formatação de Disco (1)



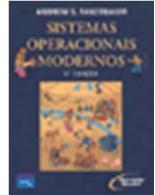
Um setor do disco

Formatação de Disco (2)



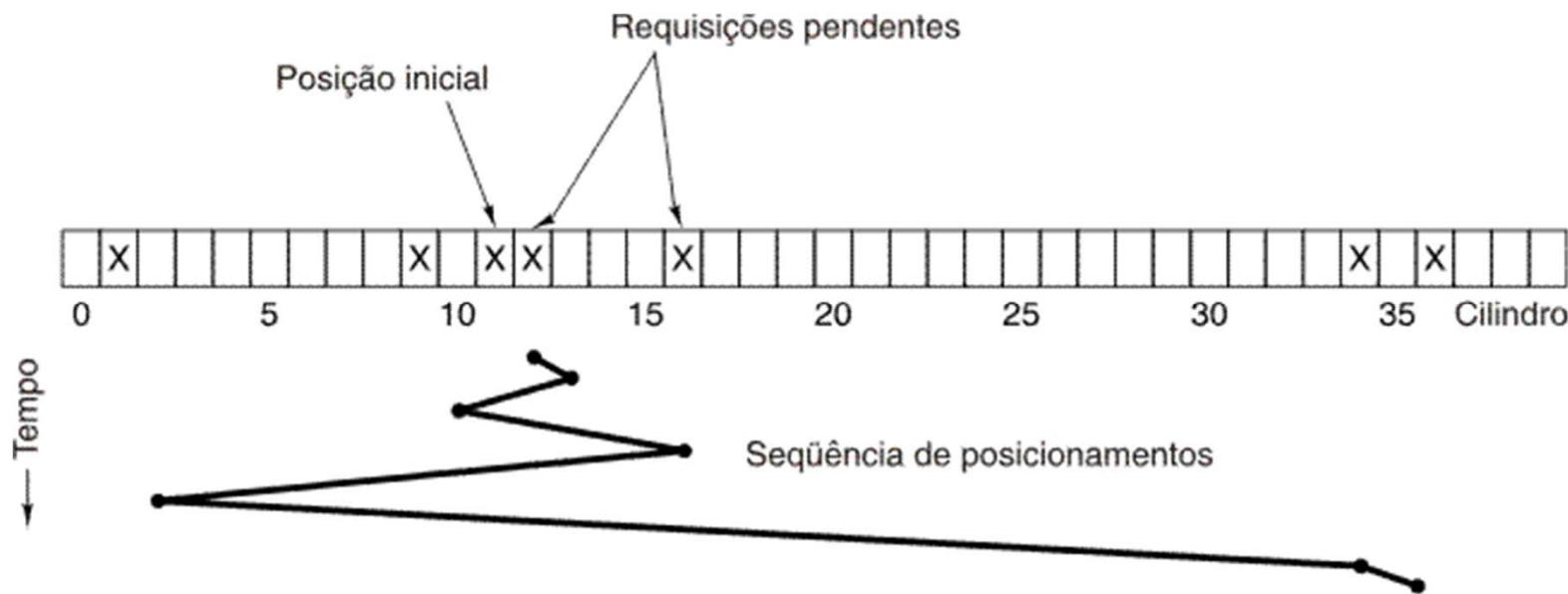
Uma ilustração da torção cilíndrica

Algoritmos de Escalonamento de Braço de Disco (1)



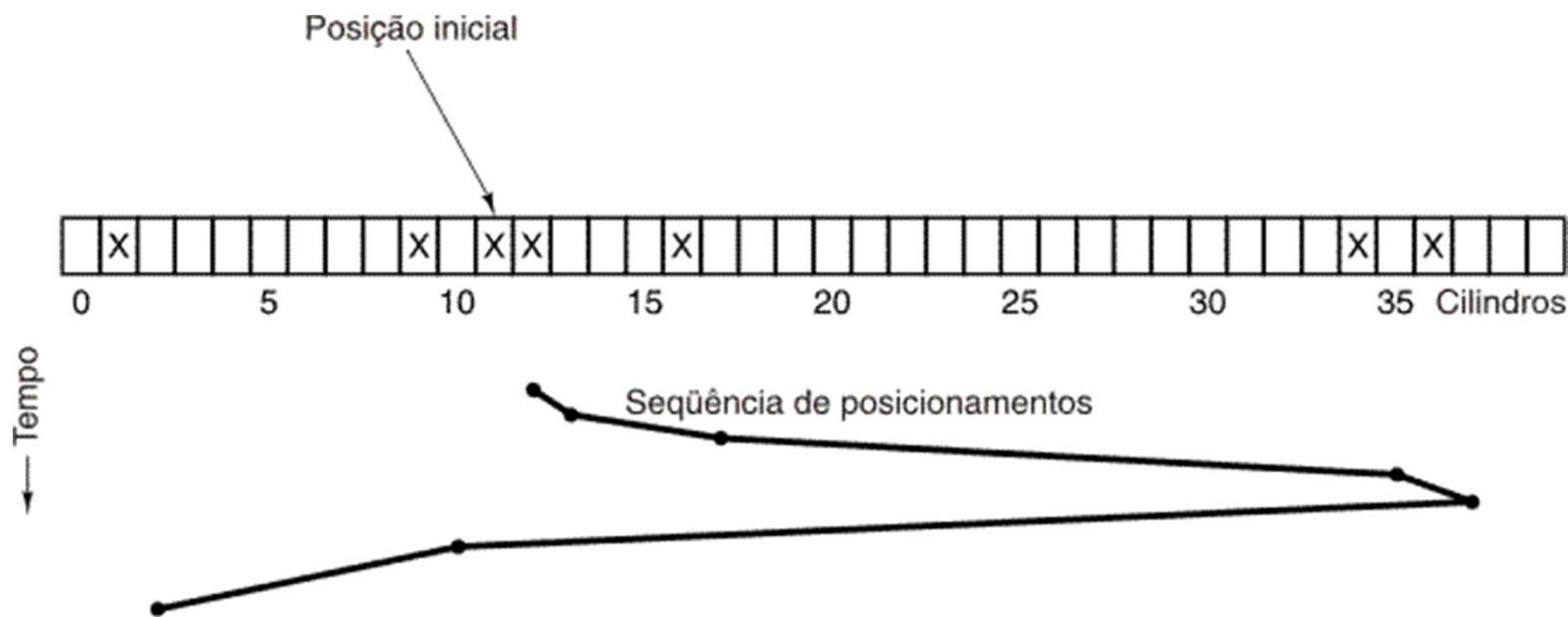
- Tempo necessário para ler ou escrever um bloco de disco é determinado por 3 fatores
 1. tempo de posicionamento
 2. atraso de rotação
 3. tempo de transferência do dado
- Tempo de posicionamento domina
- Checagem de erro é feita por controladores

Algoritmos de Escalonamento de Braço de Disco (2)



Algoritmo de escalonamento de disco *Posicionamento Mais Curto Primeiro* (SSF)

Algoritmos de Escalonamento de Braço de Disco (3)

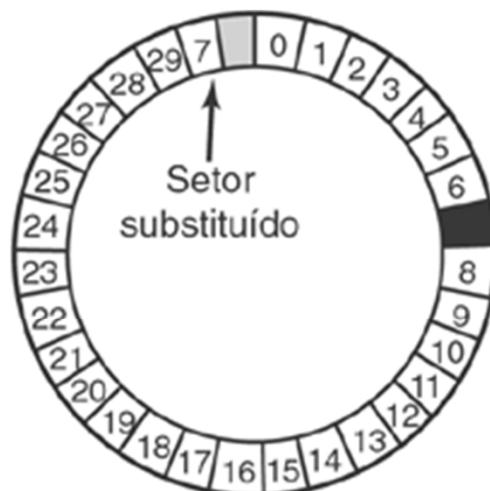


O algoritmo do elevador para o escalonamento das requisições do disco

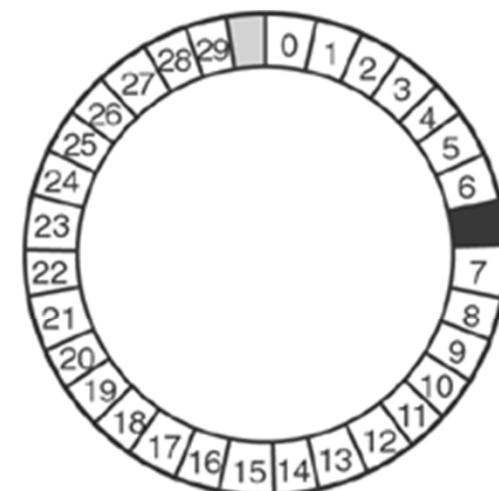
Tratamento de Erro



(a)



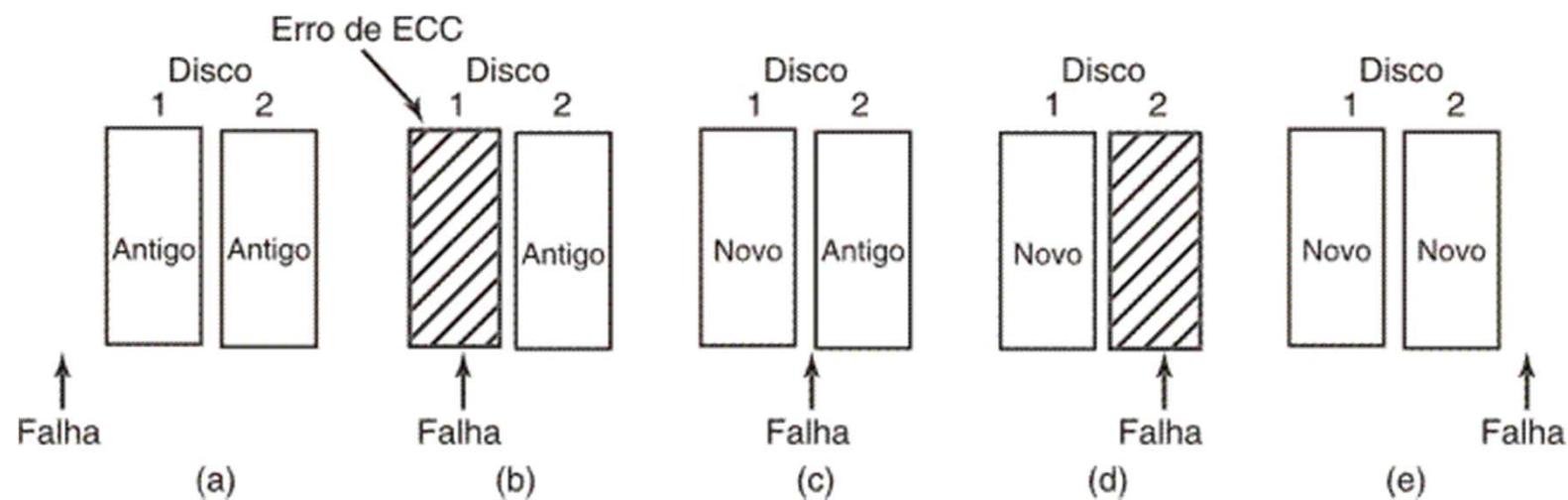
(b)



(c)

- a) Uma trilha de disco com um setor defeituoso
- b) Substituindo um setor reserva por um setor defeituoso
- c) Deslocando todos os setores para pular o setor defeituoso

Armazenamento estável



Análise da influência das falhas nas escritas estáveis