Cap. 12: Sistemas de armazenamento em *massa*

Cap. 12: Sistemas de armazenamento em massa

- Visão geral da estrutura de armazenamento em massa
- Estrutura do disco
- Anexo do disco
- Escalonamento de disco
- Gerenciamento de disco
- Gerenciamento do espaço
- Estrutura RAID
- Anexo de disco
- Implementação de armazenamento estável
- Dispositivos de armazenamento terciário
- Problemas do sistema operacional
- Questões de desempenho

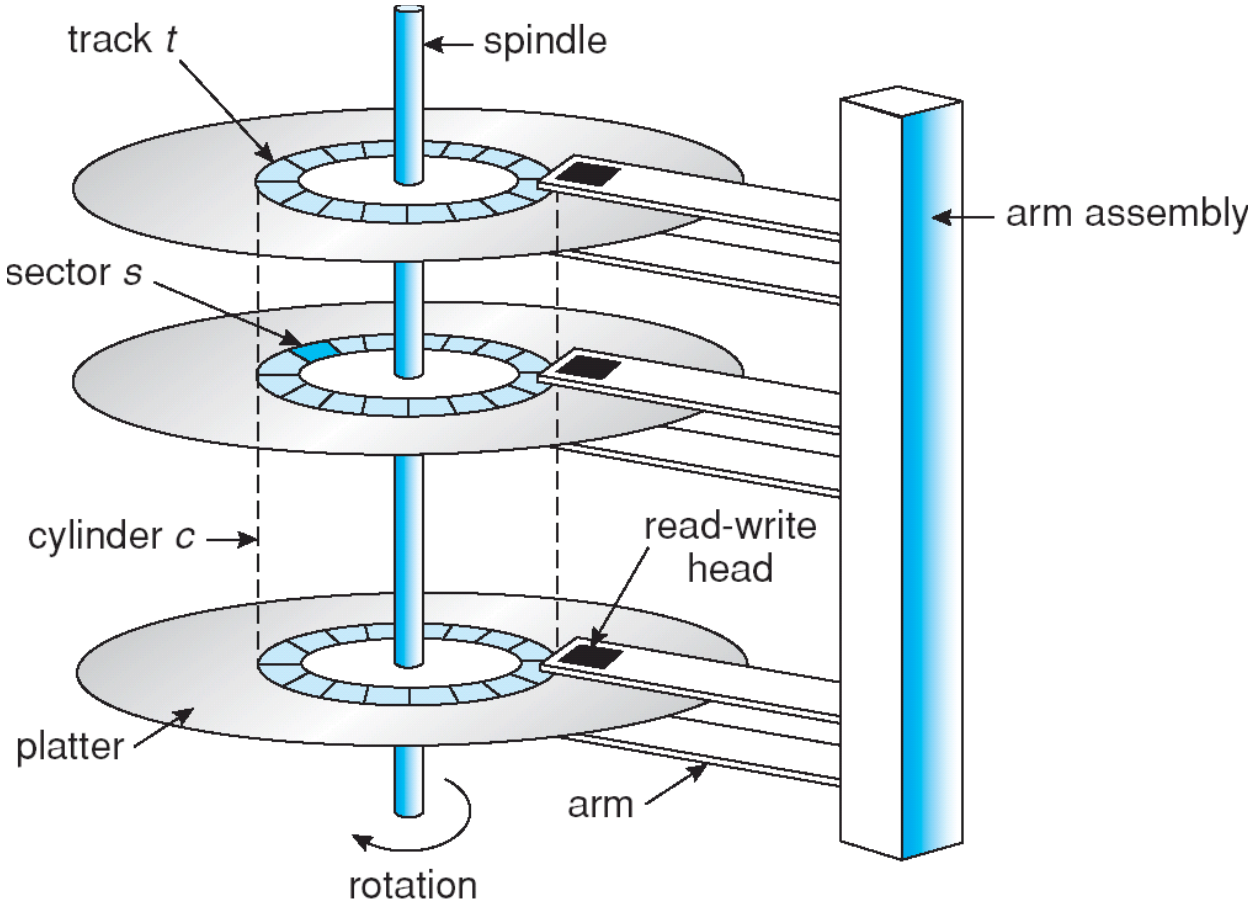
Objetivos

- Descrever a estrutura física dos dispositivos de armazenamento e os efeitos resultantes sobre o uso dos dispositivos
- Explicar as características de desempenho dos dispositivos de armazenamento em massa
- Discutir os serviços do sistema operacional fornecidos para o armazenamento em massa, incluindo RAID e HSM

Visão geral da estrutura de armazenamento em massa

- Os discos magnéticos fornecem a maior parte do armazenamento secundário dos computadores modernos
  - As unidades giram de 60 a 200 vezes por segundo
  - A **taxa de transferência** é a velocidade em que os dados fluem entre a unidade e o computador
  - **Tempo de posicionamento** (**tempo de acesso aleatório**) é o tempo para mover o braço do disco para o cilindro desejado (**tempo de busca**) e o tempo para o setor desejado aparecer sob a cabeça do disco (**latência de rotação**)
  - **Colisão da cabeça** resulta da cabeça do disco fazendo contato com a superfície do disco (isso é ruim)
- Os discos podem ser removíveis
- Unidade conectada ao computador via **barramento de E/S**
  - **Controlador de host** no computador usa barramento para "falar" com o **controlador de disco** embutido na unidade

Mecanismo de disco com cabeça móvel



Visão geral da estrutura de armazenamento em massa

- Fita magnética
  - Antigo meio de armazenamento secundário
  - Relativamente permanente e mantém grandes quantidades de dados
  - Tempo de acesso lento
  - Acesso aleatório ~1000 vezes mais lento que o disco
  - Usada principalmente para backup, armazenamento de dados usados com pouca freqüência, meio de transferência entre sistemas
  - mantida em uma bobina e avança e retrocede sob uma cabeça de leitura/escrita
  - Quando dados estão sob a cabeça, possui taxas de transferência comparáveis ao disco

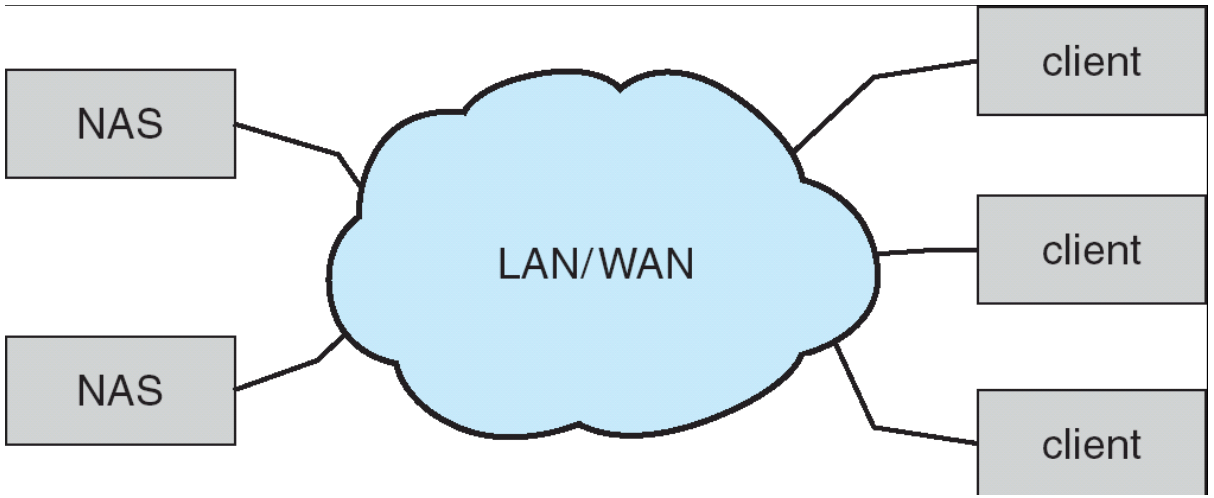
Estrutura de disco

- Unidades de disco são endereçadas como grandes arrays unidimensionais de *blocos lógicos*, onde o bloco lógico é a menor unidade de transferência.
- O array unidimensional de blocos lógicos é mapeado nos setores do disco seqüencialmente.
  - Setor 0 é o primeiro setor da primeira trilha no cilindro mais externo.
  - Mapeamento prossegue na ordem por essa trilha, depois o restante das trilhas nesse cilindro, e depois pelo restante dos cilindros de fora para dentro.

Armazenamento conectado à rede

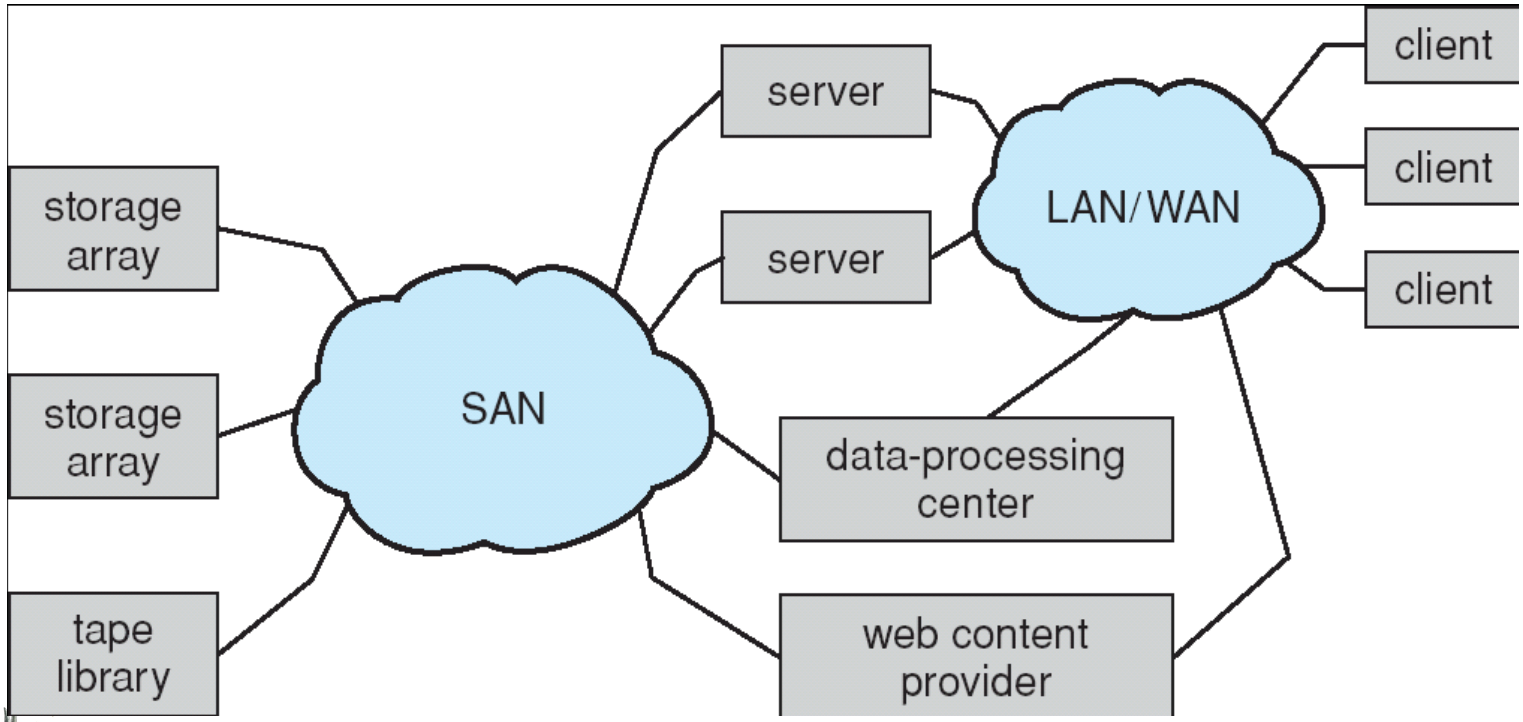
O armazenamento conectado à rede (**NAS - Network Attached Storage**) é o armazenamento disponível por uma rede, ao invés de uma conexão local (via barramento)

- Implementado via chamadas de procedimento remoto (RPCs) entre o host e o armazenamento



Storage Area Network

- Comum em ambientes de grande armazenamento
- Múltiplos hosts conectados a múltiplos arrays de armazenamento - flexível



Escalonamento de disco

- O sistema operacional é responsável por usar o hardware de forma eficiente – para as unidades de disco, isso significa ter um tempo de acesso rápido e largura de banda de disco.
- Tempo de acesso tem dois componentes principais
  - *Tempo de busca* é o tempo para o disco mover as cabeças até o cilindro contendo o setor desejado.
  - *Latência de rotação* é o tempo adicional aguardando o disco girar o setor desejado até a cabeça do disco.
- Largura de banda de disco é o número total de bytes transferidos, dividido pelo tempo total entre a primeira solicitação de serviço e o término da última transferência.

Escalonamento de disco (cont.)

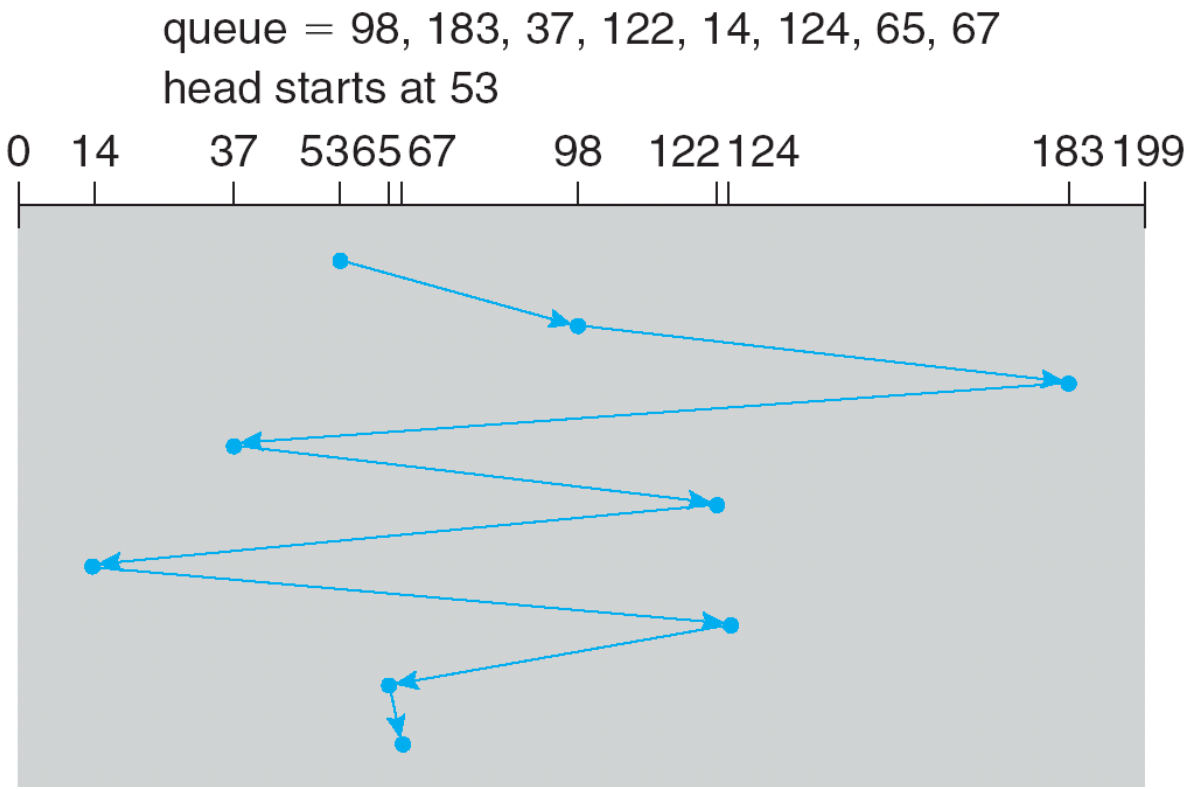
- Existem vários algoritmos para escalonar o atendimento das solicitações de E/S de disco.
- Ilustramos com uma fila de solicitação

98, 183, 37, 122, 14, 124, 65, 67

Ponteiro da cabeça inicialmente em 53

FCFS

Ilustração mostra movimento total da cabeça



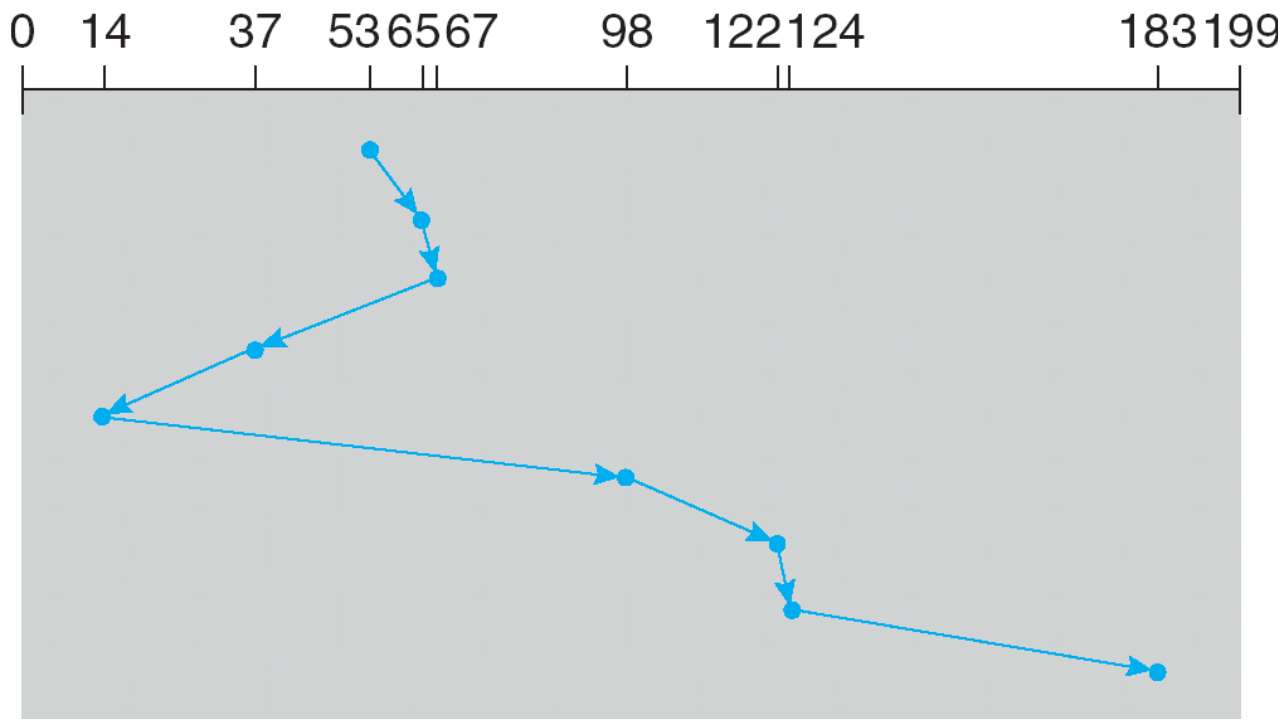
Shortest Seek-Time First (SSTF)

- Seleciona a solicitação com o tempo de busca mínimo a partir da posição atual da cabeça.
- Escalonamento SSTF é uma forma de escalonamento SJF; pode causar starvation de algumas solicitações.

SSTF (cont.)

queue = 98, 183, 37, 122, 14, 124, 65, 67

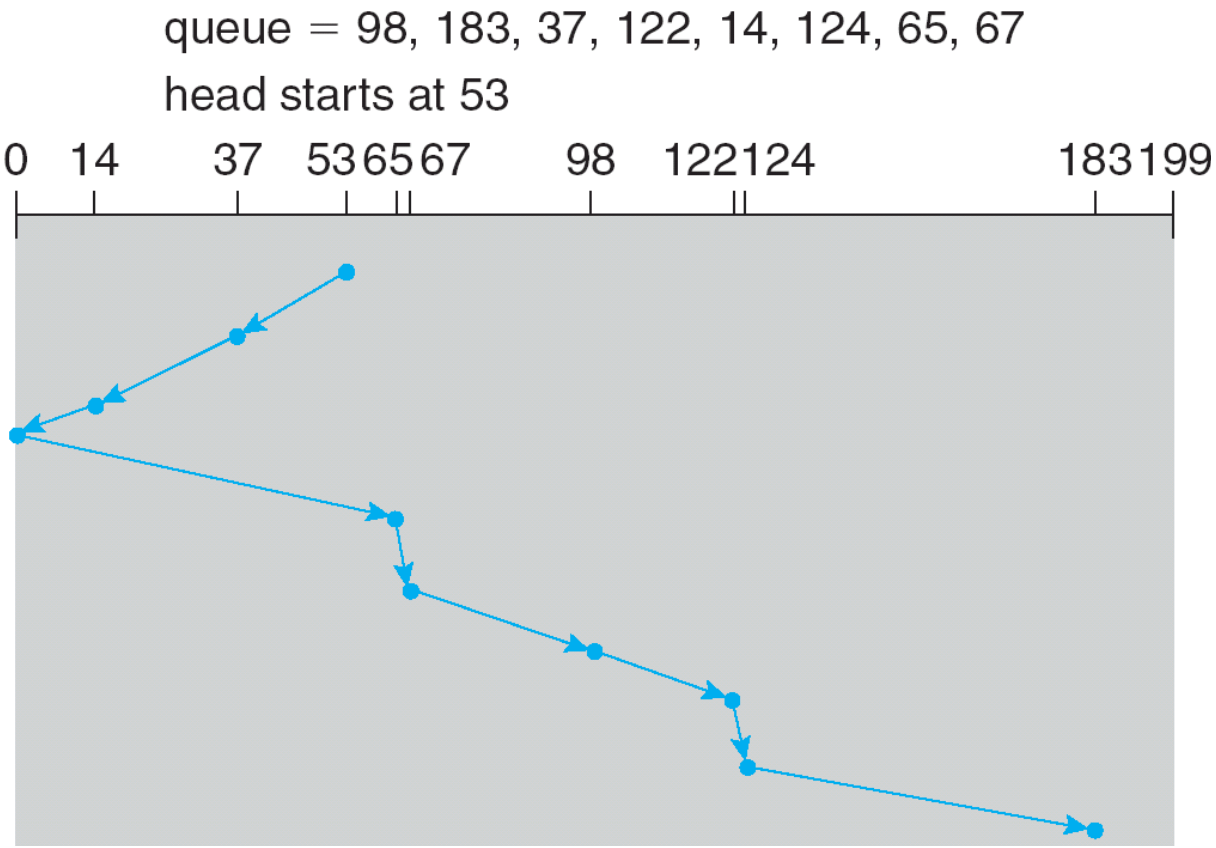
head starts at 53



SCAN

- O braço do disco começa em uma extremidade do disco e se move para a outra extremidade, atendendo solicitações até que chegue à outra extremidade, onde o movimento da cabeça é revertido e o atendimento continua.
- Às vezes chamado de *algoritmo do elevador.*

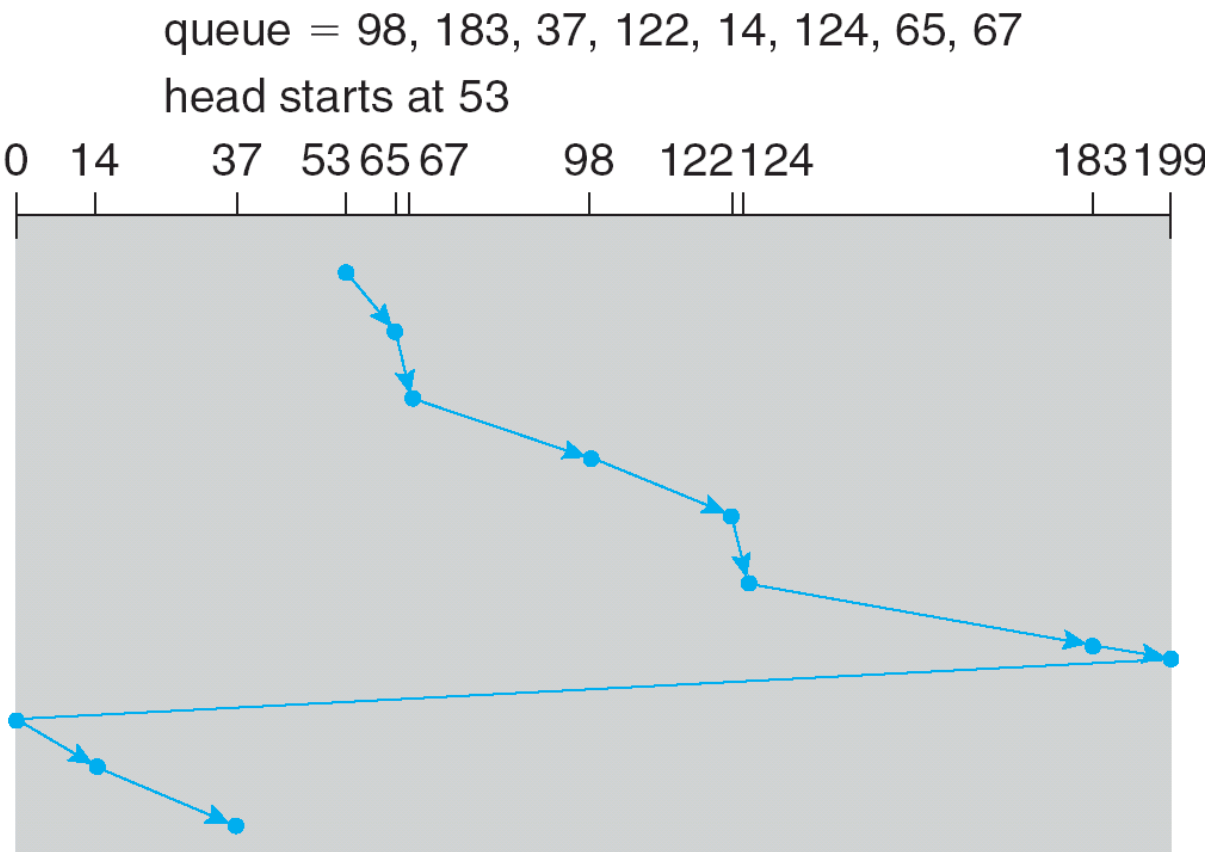
SCAN (cont.)


queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53
0
14
37
53
65
67
98
122
124
183
199

C-SCAN

- Fornece um tempo de espera mais uniforme que SCAN.
- A cabeça se move de uma extremidade do disco para a outra, atendendo solicitações enquanto prossegue. Quando atinge o outro extremo, imediatamente retorna ao início do disco, sem atender quaisquer solicitações no retorno.
- Trata os cilindros como uma lista circular que contorna o último cilindro e volta ao primeiro.

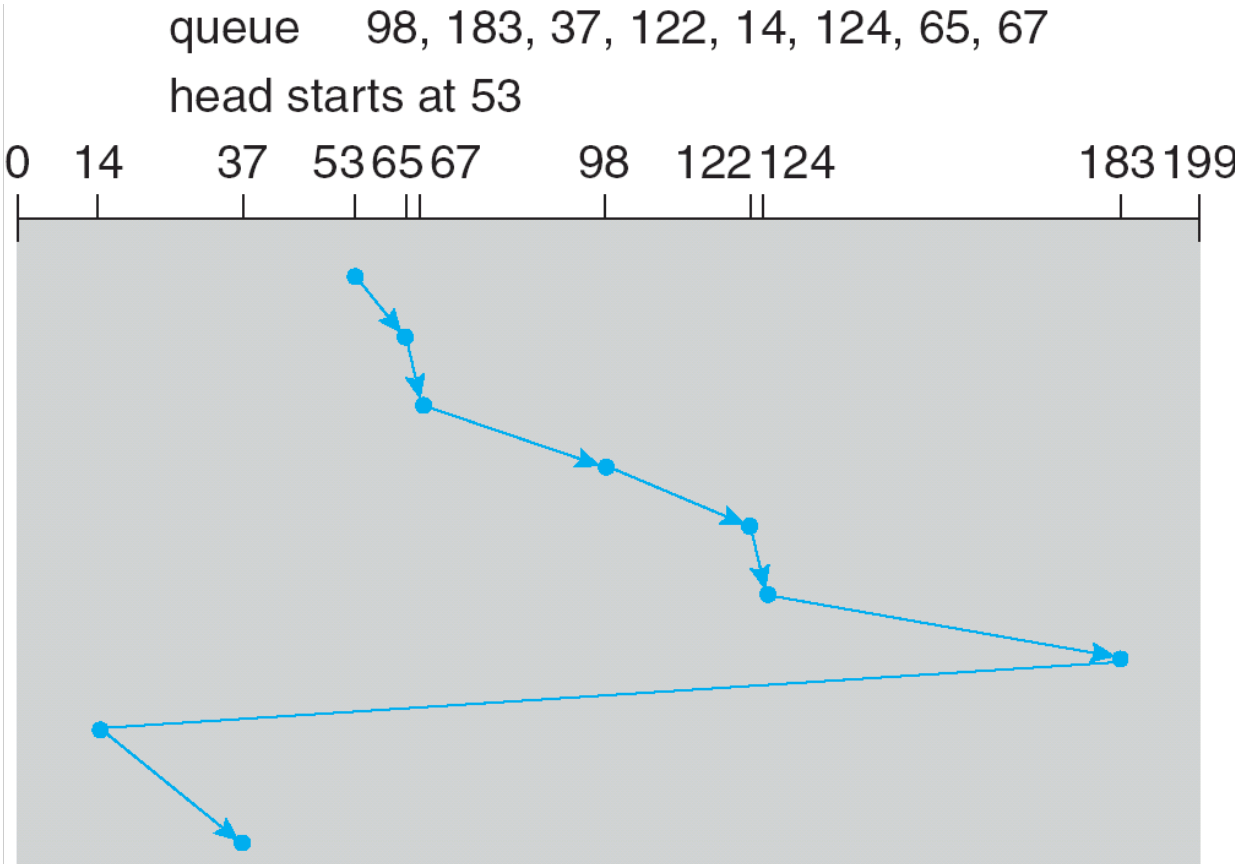
C-SCAN (cont.)


queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53
0
14
37
53
65
67
98
122
124
183
199

C-LOOK

- Versão de C-SCAN
- O braço só vai até a distância da última solicitação em cada direção, depois reverte a direção imediatamente, sem primeiro ir até o final do disco.

C-LOOK (cont.)


queue 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53
0
14
37
53
65
67
98
122
124
183
199

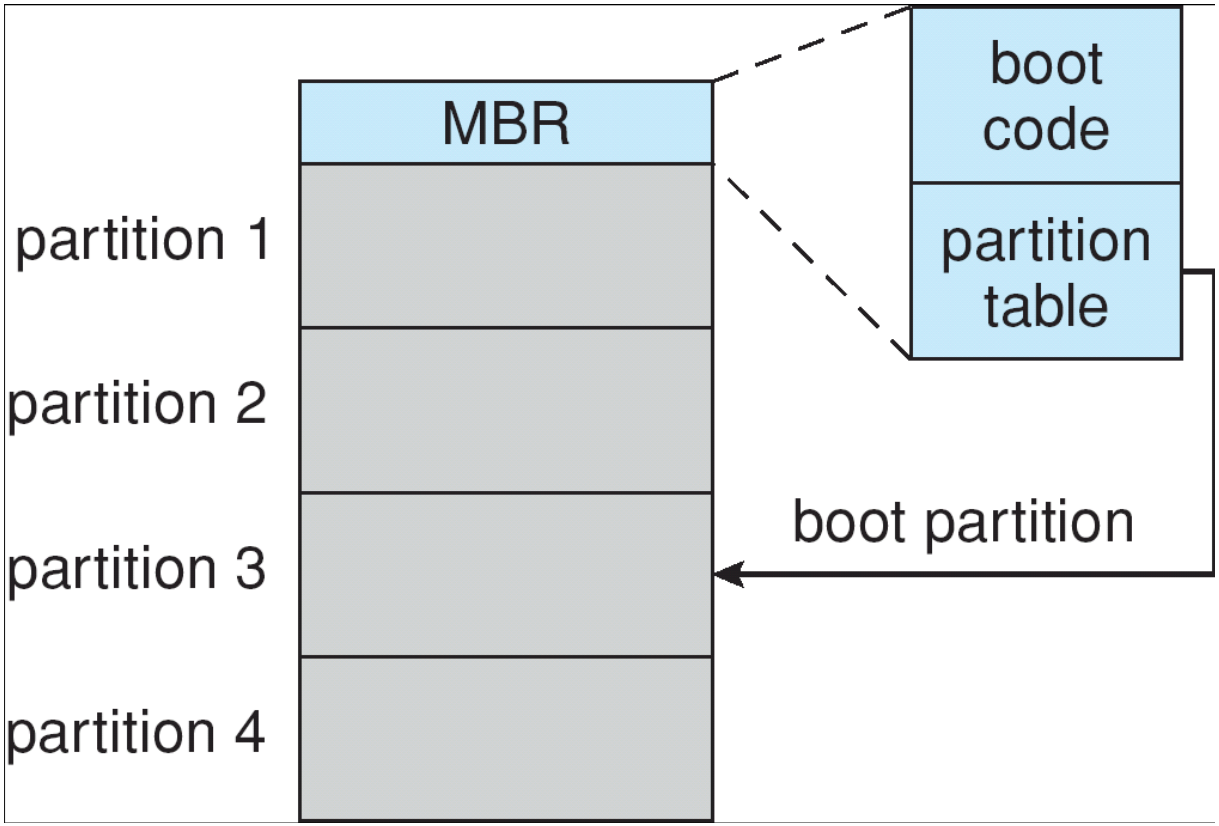
Selecionando um algoritmo

- SSTF é comum e tem um apelo natural
- SCAN e C-SCAN funcionam melhor para sistemas que têm cargas pesadas sobre o disco
- O desempenho depende do número e tipo de solicitações.
- Requisições para serviço de disco podem ser influenciadas pelo método de alocação de arquivo.
- O algoritmo de escalonamento de disco deve ser escrito como um módulo separado do sistema operacional, permitindo que seja substituído por um algoritmo diferente, se necessário.
- SSTF ou LOOK é uma escolha razoável para o algoritmo padrão.

Gerenciamento de disco

- *Formatação de baixo nível* ou *formatação física* – Dividindo um disco em setores que o controlador pode ler e gravar.
- Para usar um disco para manter arquivos, o sistema operacional ainda precisa registrar suas próprias estruturas de dados no disco.
  - *Partição* do disco em um ou mais grupos de cilindros.
  - *Formatação lógica* ou “criação do sistema de arquivos”.
- Bloco de boot inicializa sistema.
  - O bootstrap é armazenado na ROM.
  - Programa *carregador de bootstrap*.
- Métodos usados para tratar de blocos defeituosos.

Boot por um disco no Windows 2000



Gerenciamento do swap space

- Swap space — Memória virtual usa espaço do disco como extensão da memória principal.
- Swap space pode estar junto do sistema de arquivos normal ou, como geralmente acontece, pode estar em uma partição de disco separada.
- Gerenciamento do swap-space
  - aloca swap space quando processo inicia

Estrutura RAID

RAID - múltiplas unidades de disco

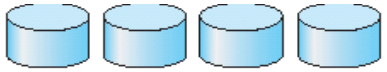
- gera **confiabilidade** via **redundância**

- RAID é organizado em seis diferentes níveis.

RAID (cont.)

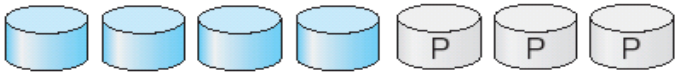
- Várias melhorias nas técnicas de uso de disco envolvem o uso de múltiplos discos funcionando cooperativamente.
- Espalhamento de disco usa um grupo de discos como uma unidade de armazenamento.
- Esquemas RAID melhoram o desempenho e melhoram a confiabilidade do sistema de armazenamento armazenando dados redundantes.

Níveis de RAID



(a) RAID 0: non-redundant striping.

(b) RAID 1: mirrored disks.



(c) RAID 2: memory-style error-correcting codes.

(d) RAID 3: bit-interleaved parity.

(e) RAID 4: block-interleaved parity.

(f) RAID 5: block-interleaved distributed parity.

(g) RAID 6: P + Q redundancy.

Implementação de armazenamento estável

- Para implementar armazenamento estável:
  - Replicar informações em mais de um meio de armazenamento não volátil com modos de falha independentes.
  - Atualizar informações de maneira controlada para garantir que possamos recuperar os dados estáveis após qualquer falha durante a transferência ou recuperação de dados.

Dispositivos de armazenamento terciário

- Baixo custo é a característica principal do armazenamento terciário.
- Geralmente, o armazenamento terciário é montado usando *mídia removível*
- Exemplos comuns de mídia removível são disquetes e CD-ROMs; outros tipos estão disponíveis (pendrive, HD externa).

Discos WORM

- Os dados nos discos de leitura-escrita podem ser modificados indefinidamente.
- Discos WORM (“Write Once, Read Many Times”) só podem ser gravados uma vez.
- Fina camada de alumínio entre duas placas de vidro ou plástico.
- Para gravar um bit, a unidade usa uma luz de laser para queimar um pequeno furo pelo alumínio; as informações podem ser destruídas, mas não alteradas.
- Muito duráveis e confiáveis.
- Discos *somente de leitura*, como CD-ROM e DVD, vêm de fábrica com os dados pré-gravados.

Aspectos do sistema operacional

- As principais tarefas do SO são gerenciar dispositivos físicos e apresentar uma abstração de máquina virtual às aplicações
- Para discos rígidos, o SO oferece duas abstrações:
  - Dispositivo bruto – um array de blocos de dados.
  - Sistema de arquivos – o SO enfileira e escalona as requisições intercaladas de várias aplicações.

Interface de aplicação

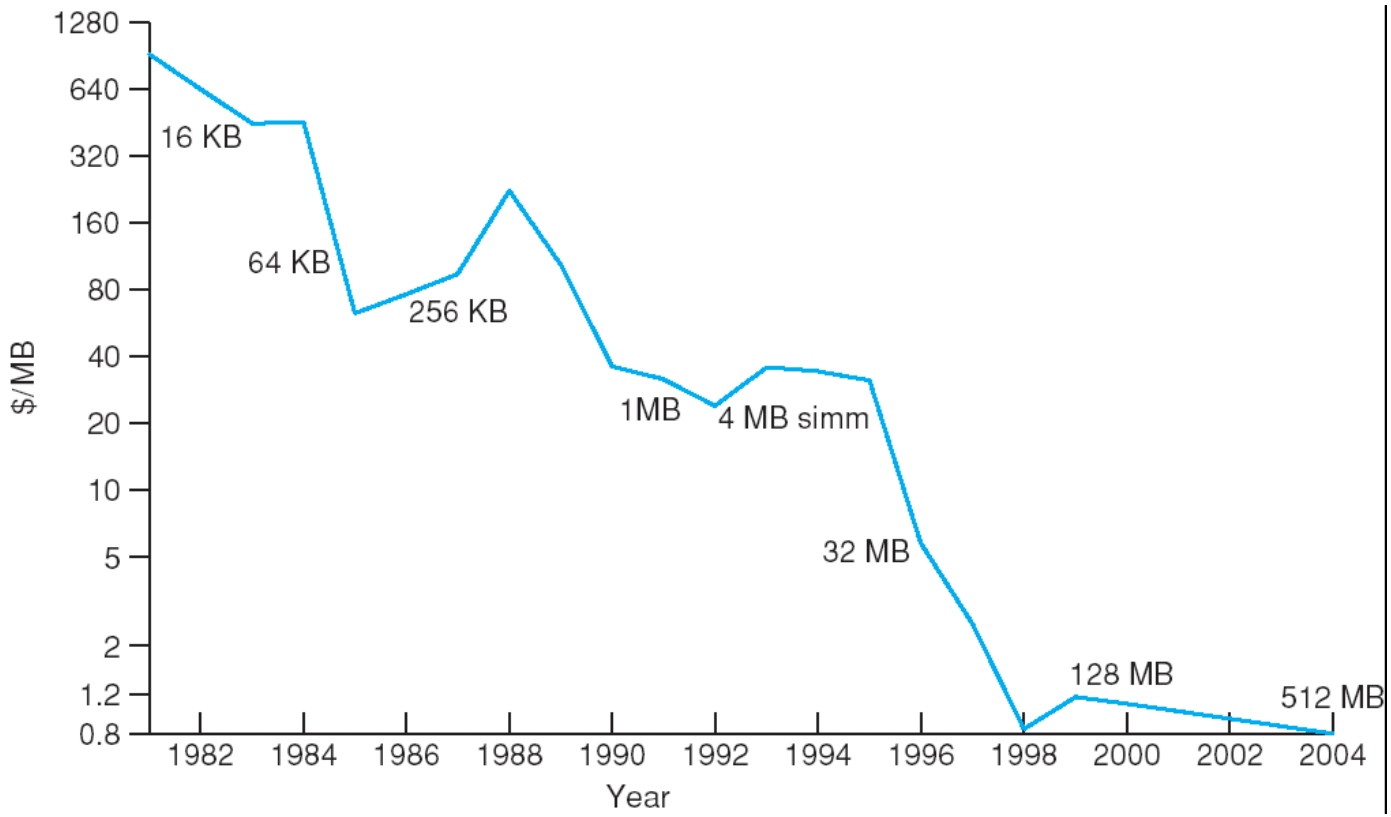
- A maioria dos SOs trata de discos removíveis quase exatamente como os discos fixos
- As fitas são apresentadas como um meio de armazenamento bruto, ou seja, uma aplicação não abre um arquivo na fita, ela abre a unidade de fita inteira como um dispositivo bruto.

Como o SO não oferece serviços do sistema de arquivos, a aplicação precisa decidir como usar o array de blocos.Como cada aplicação cria suas próprias regras de como organizar uma fita, uma fita cheia de dados geralmente só pode ser usada pelo programa que a criou.

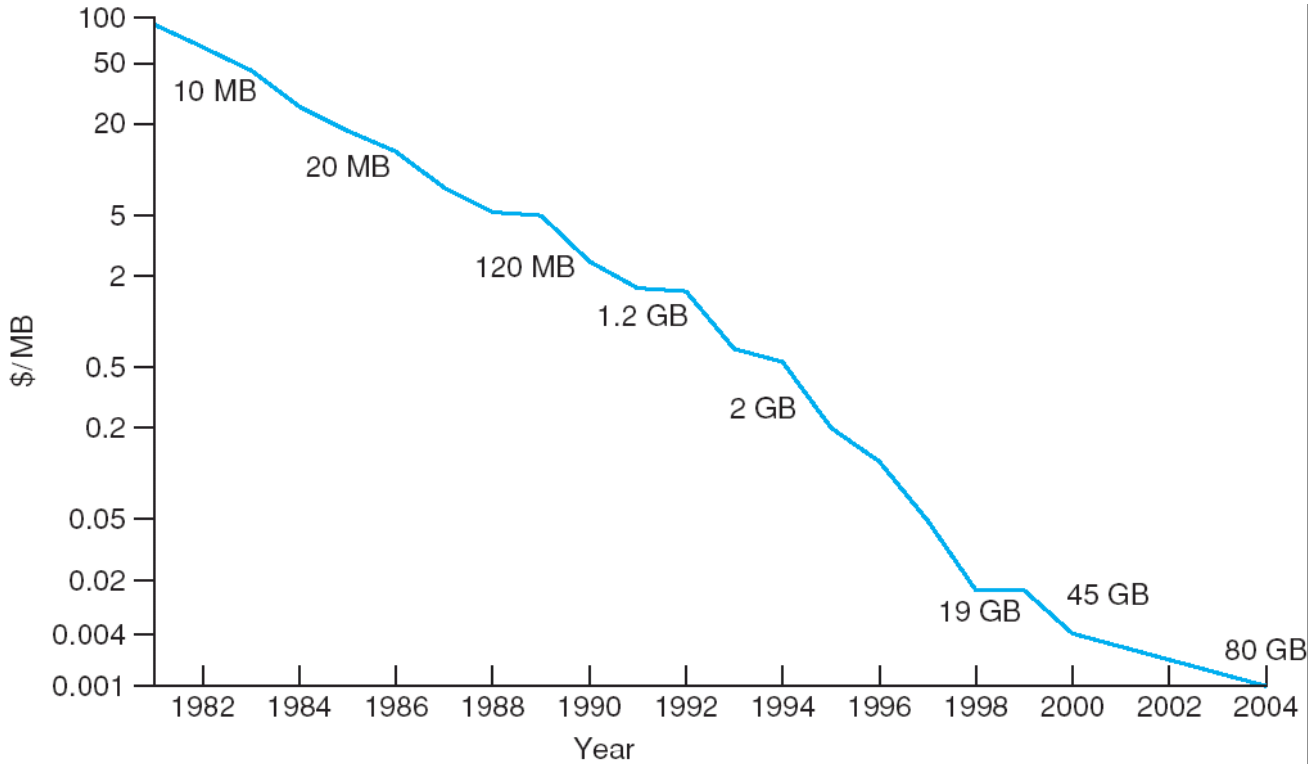
Hierarchical Storage Management (HSM)

- Um sistema de armazenamento hierárquico estende a hierarquia de armazenamento além da memória principal e armazenamento secundário, para incorporar o armazenamento terciário – normalmente implementado como um *jukebox* de fitas ou discos removíveis.
- Normalmente, incorpora armazenamento terciário estendendo o sistema de arquivos.
  - Arquivos pequenos e usados freqüentemente permanecem no disco.
  - Arquivos inativos, grandes e antigos, são arquivados no *jukebox*.
- HSM normalmente é encontrado em centros de supercomputação e outras grandes instalações, que possuem enormes volumes de dados.

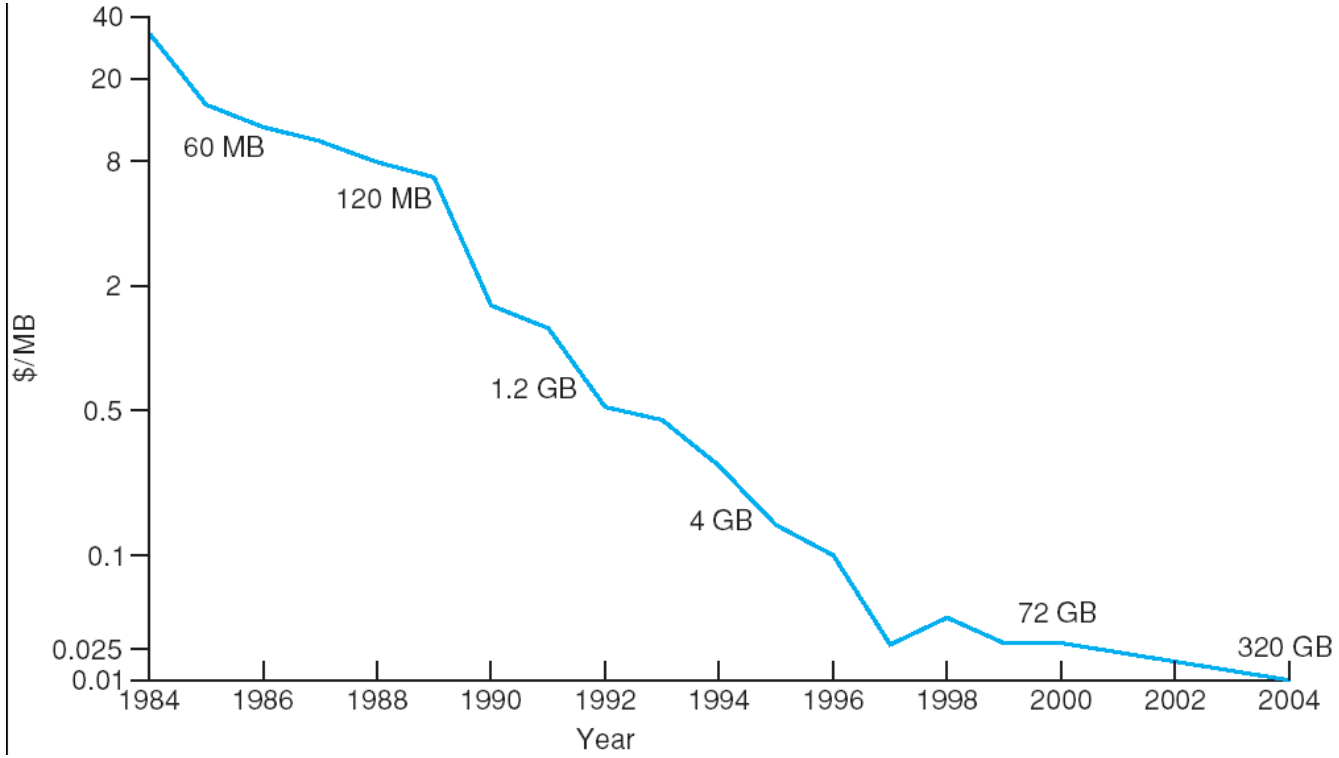
Preço por megabyte de RAM, de 1981 a 2004



Preço por megabyte do disco rígido magnético, de 1981 a 2004



Preço por megabyte de uma unidade de fita, de 1984 a 2000



Final do Capítulo 12