Sistemas de armazenamento em massa

Prof. Dr. João Carlos Lopes Fernandes
E-mail: joao.fernandes1@docente.unip.br


Marlboro
vodafone
©AFP/Arquivo Toru Yamanaka

Curiosidade: Criptografia

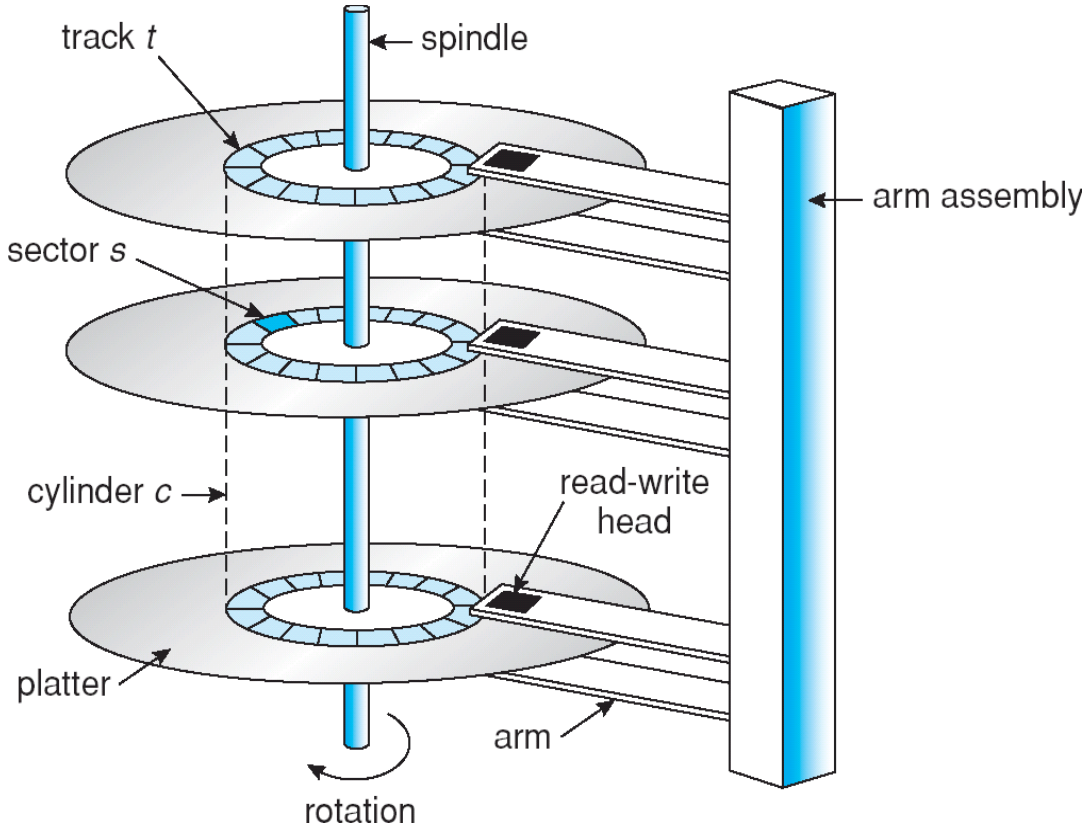
Sugestão de Leitura:

https://www.infowester.com/criptografia.php

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



Fita magnética

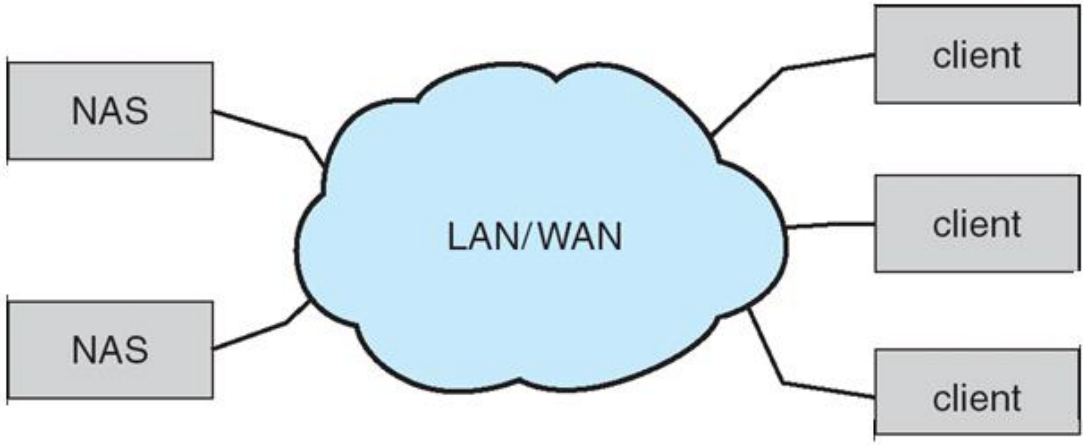
- Antigo meio de armazenamento secundário
- Relativamente permanente e mantém grandes quantidades de dados
- Tempo de acesso lento
- Acesso aleatório ~1000 vezes mais lento que o disco
- Usada principalmente para backup, armazenamento de dados usados com pouca frequência, meio de transferência entre 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 (como já visto)

- 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 sequencialmente.
  - 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.

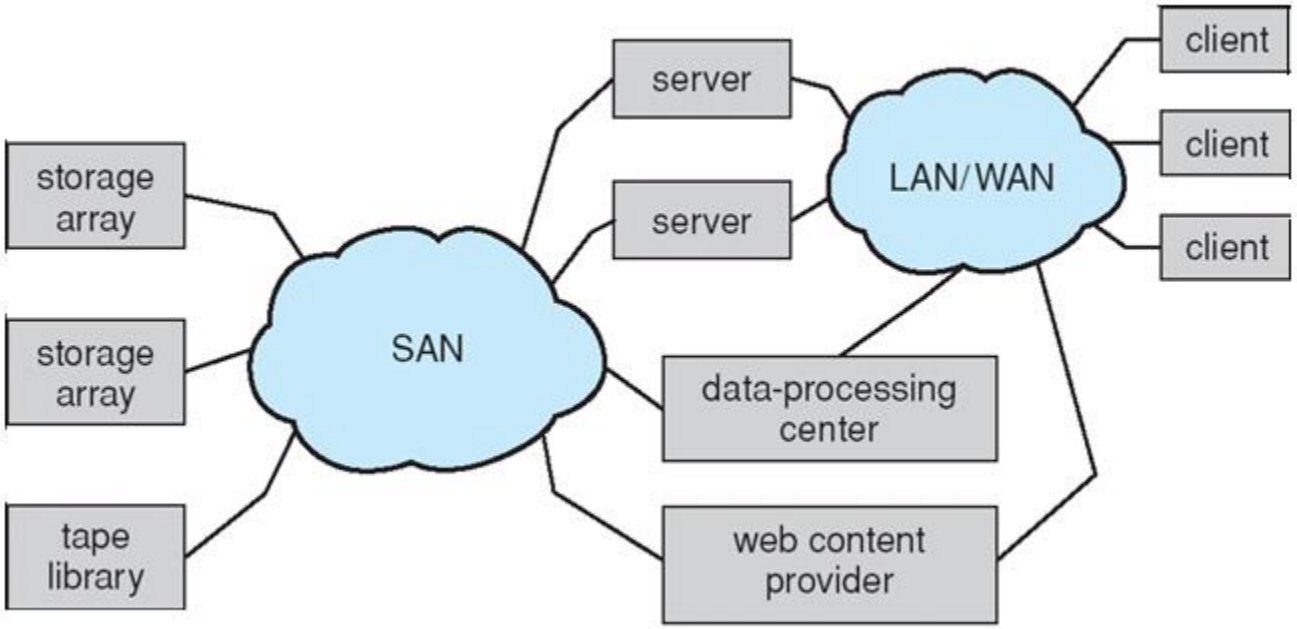
Armazenamento conectado à rede

- 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 - *Remote Procedure Call*) entre o host e o armazenamento



SAN (*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 (*First Come, First Served*)

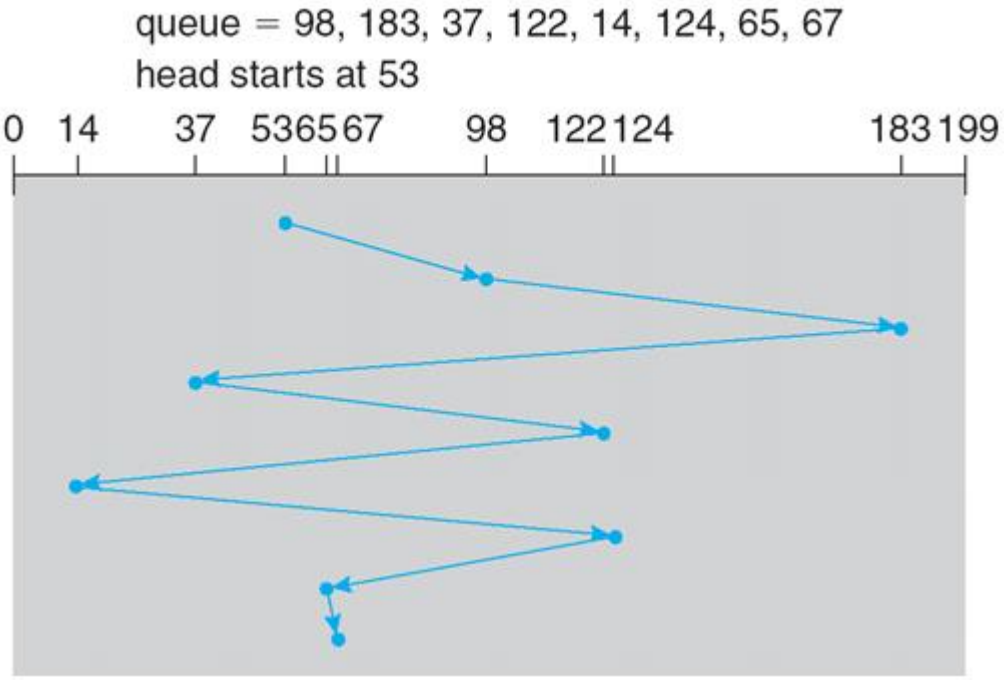
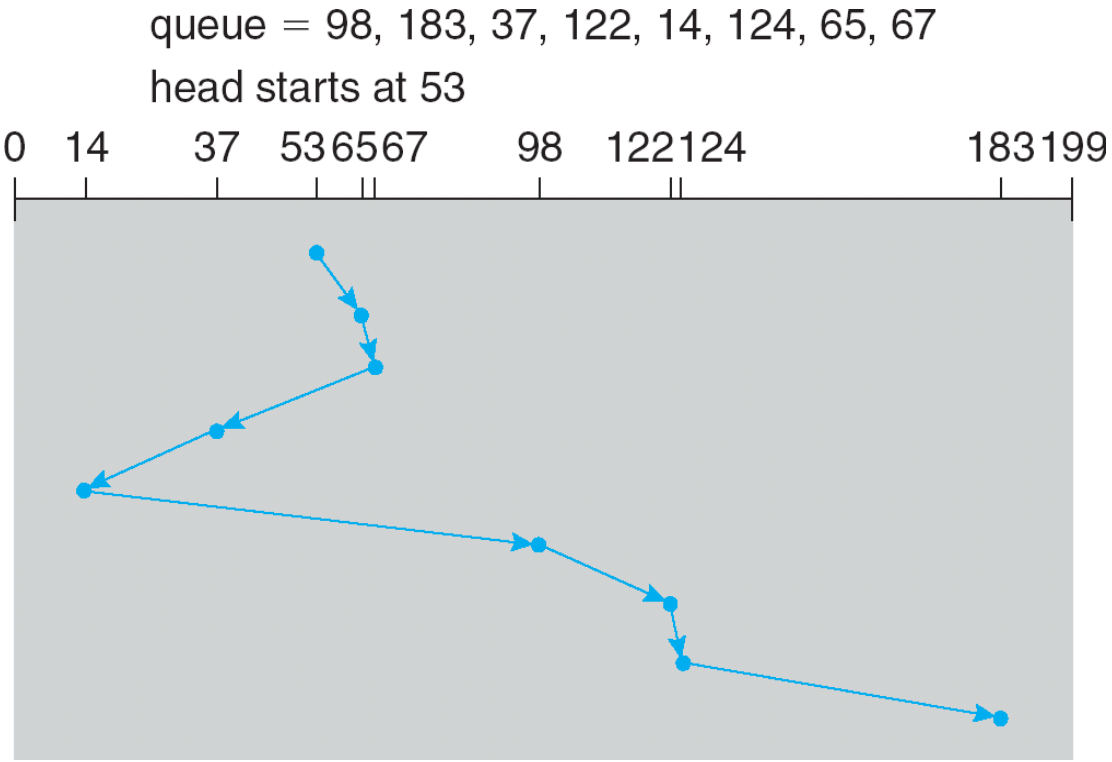


Ilustração mostra movimento total da cabeça

SSTF (*Shortest Seek-Time First*)

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

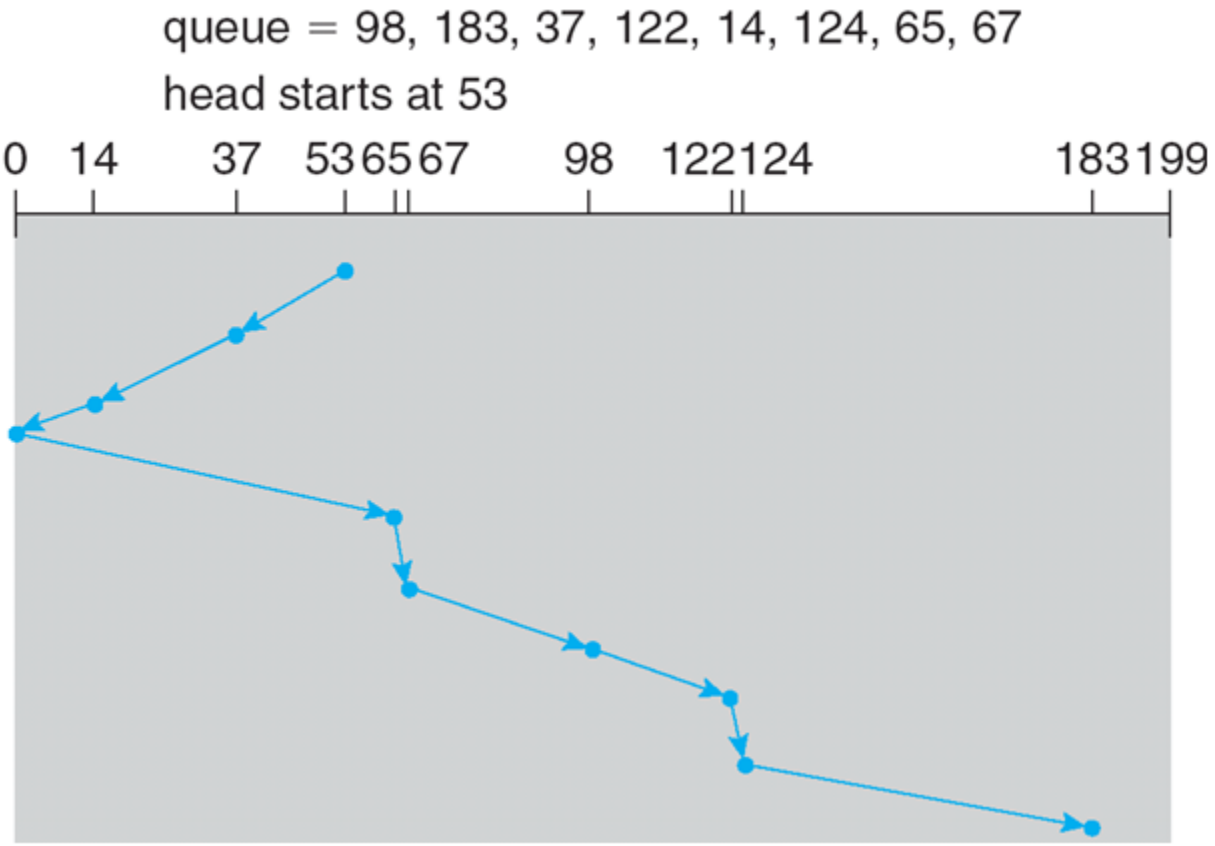
SSTF (cont.)


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

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.

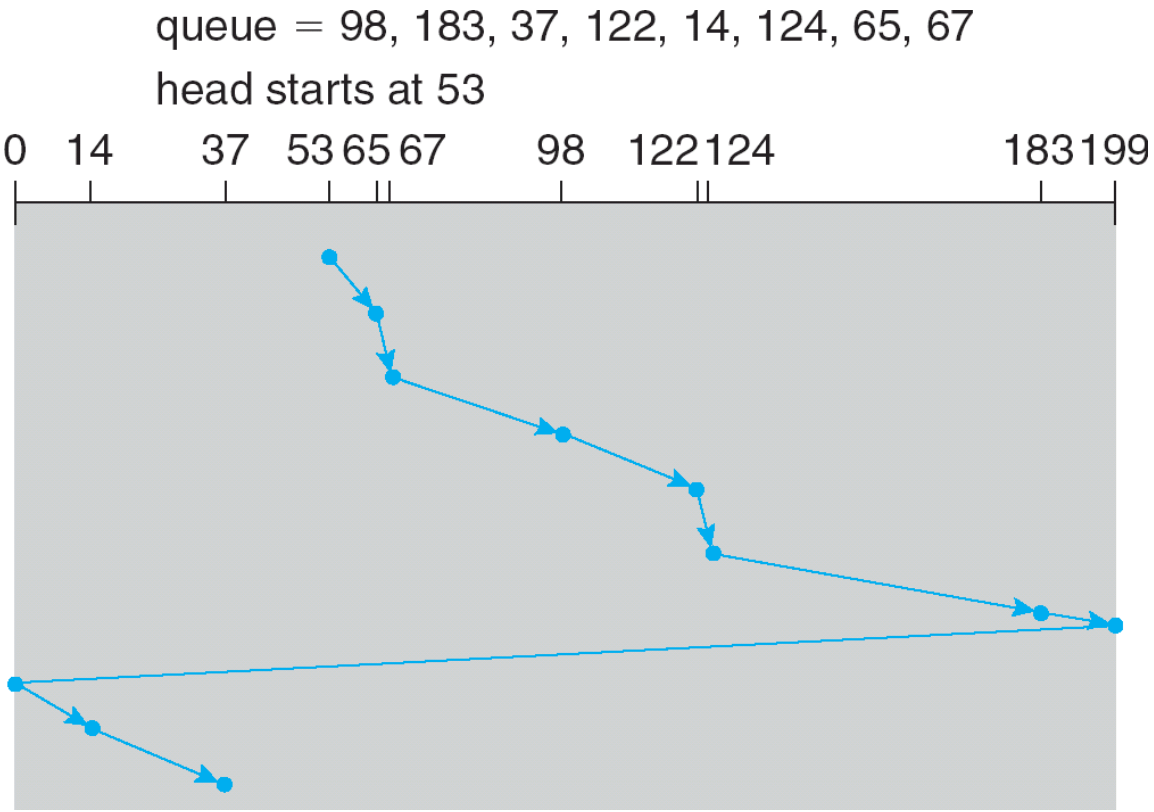
SCAN (cont.)


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.

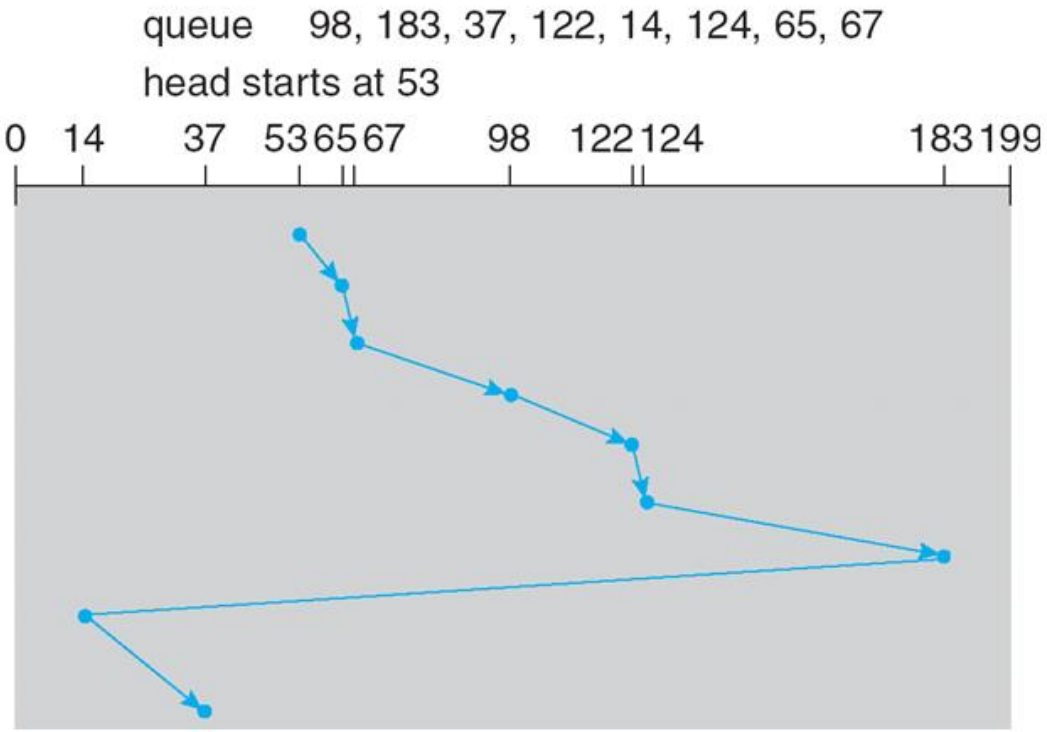
C-SCAN (cont.)


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

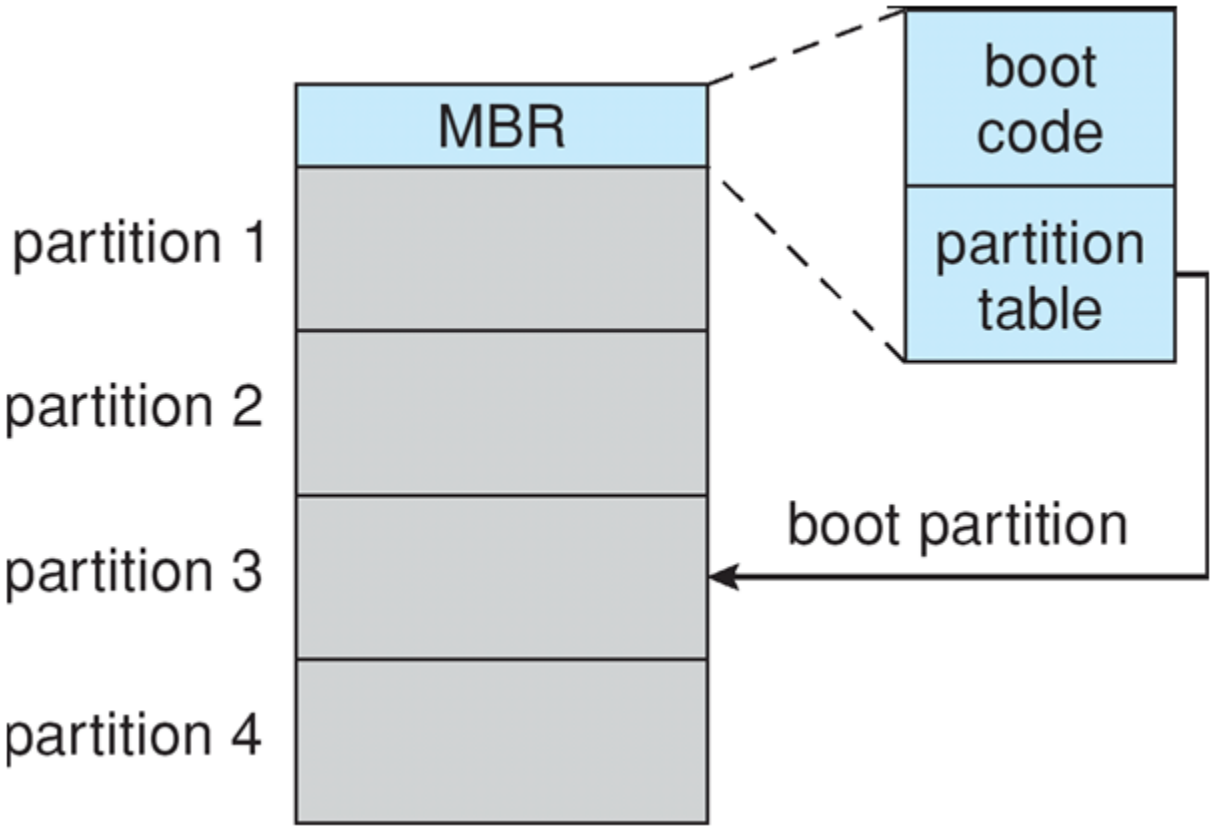
Selecionando um algoritmo

- 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



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

RAID (*Redundant Array of Independent Disks*)

- RAID - múltiplas unidades de disco
  - gera confiabilidade via redundância
- RAID é organizado em seis diferentes níveis.

Originalmente denominado de "*Redundant Array of Inexpensive Drives*"

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 dos dados redundantes.

(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. (Nuvem)
- Geralmente, o armazenamento terciário é montado usando mídia removível
- Exemplos comuns de mídia removível eram 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. (Depende do armazenamento)
- 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.

HSM (*Hierarchical Storage Management*)

- 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 frequentemente 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.

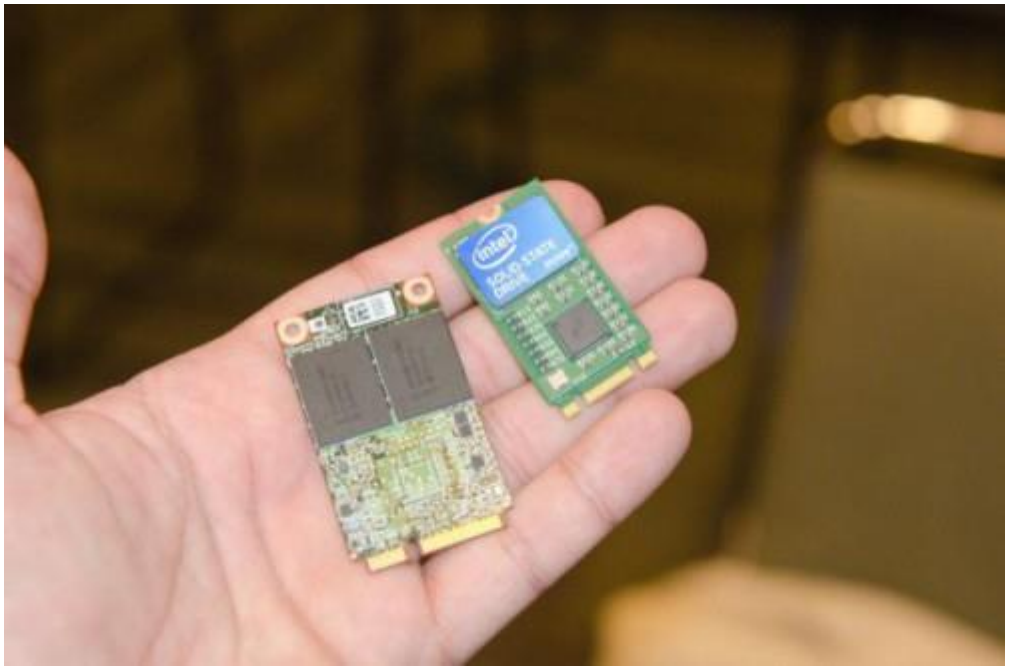
SSD - Solid-State Drive

Prof. Dr. João Carlos Lopes Fernandes
E-mail: joao.fernandes1@docente.unip.br

SSD - Solid-State Drive

- Os SSDs podem utilizar duas tecnologias: Mini-SATA ou mSATA (a mais antiga) e M.2, considerada a tecnologia mais recente.
- O tipo de unidade interfere na capacidade de armazenamento e desempenho do SSD.
- O termo M.2, também conhecido como Next Generation Form Factor (NGFF), se refere a uma conexão interna que se aplica a diferentes tipos de placas adicionais, como Wi-Fi, Bluetooth, navegação por satélites, entre outros.

M.2



mSATA a esquerda, M.2 a direita

NVMe

Assim como os SSDs tradicionais, os discos NVMe (*Non-Volatile Memory Express*) são baseados na memória Flash NAND. Por estarem diretamente atachados no barramento PCI Express (PCIe), usado anteriormente por outros dispositivos como placas gráficas, os discos NVMe não são afetados pelos gargalos da interface SATA. Os seus benefícios são:

Velocidade: A conexão PCIe x8 oferece desempenho de leitura / gravação e largura de banda maiores.

Capacidade: A capacidade de armazenamento dos discos NVMe disponíveis na OPEN inicia em 450 GB ate 9.6 TB.