Gerenciamento de Memória Secundária

Dados Armazenados de Maneira Persistente

Engenharia de Computação
Pontifícia Universidade Católica de Campinas

Prof. Dr. Denis M. L. Martins

Disclaimer

Parte do material apresentado a seguir foi adaptado de:

- IT Systems – Open Educational Resource, produzido por Jens~Lechtenböger; e
- Open Education Hub - Operating Systems

Imagens decorativas (à esquerda dos slides) retiradas de Unsplash

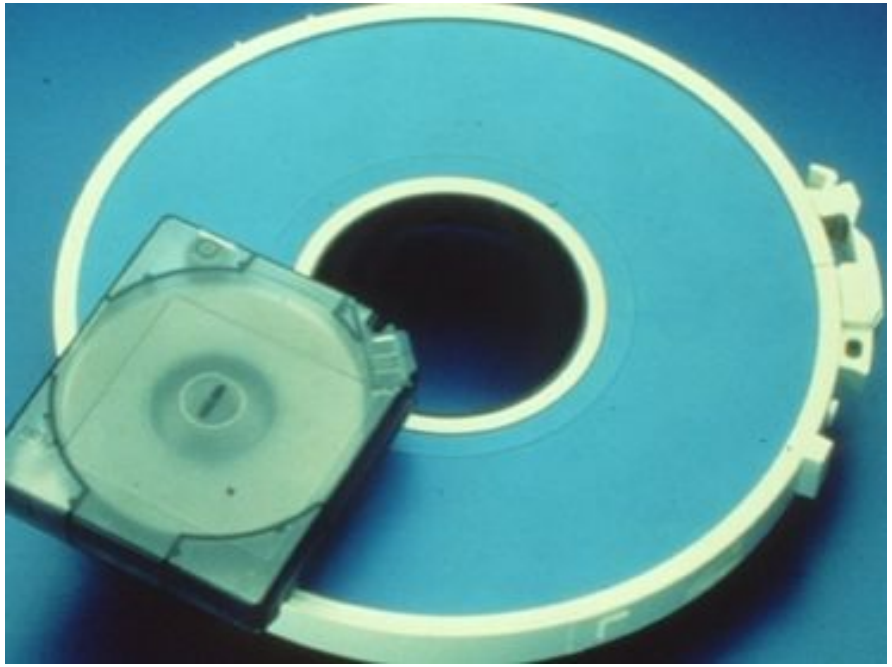
Objetivos

- Descrever a estrutura física dos dispositivos de armazenamento secundário e o efeito da estrutura do dispositivo em seus usos.
- Explicar as características de desempenho dos dispositivos de armazenamento em massa.
- Compreender o funcionamento de algoritmos de escalonamento de disco.

Estrutura de Armazenamento em Massa

- A maior parte do armazenamento secundário para computadores modernos é composta por unidades de disco rígido (HDDs) e dispositivos de memória não volátil (NVM).
- HDDs giram pratos revestidos magneticamente sob cabeçotes de leitura/escrita móveis.
  - As unidades giram a velocidades entre 60 e 250 rotações por segundo.
  - A taxa de transferência é a velocidade com que os dados fluem entre a unidade e o computador.
  - O tempo de posicionamento (tempo de acesso aleatório) é o tempo necessário para mover o braço do disco até o cilindro desejado (tempo de busca) e para que o setor desejado gire sob o cabeçote do disco (latência rotacional).
  - Uma falha no cabeçote de leitura/escrita ocorre quando o cabeçote entra em contato com a superfície do disco – um evento indesejável.
- Os discos podem ser removíveis.

Fita Magnética



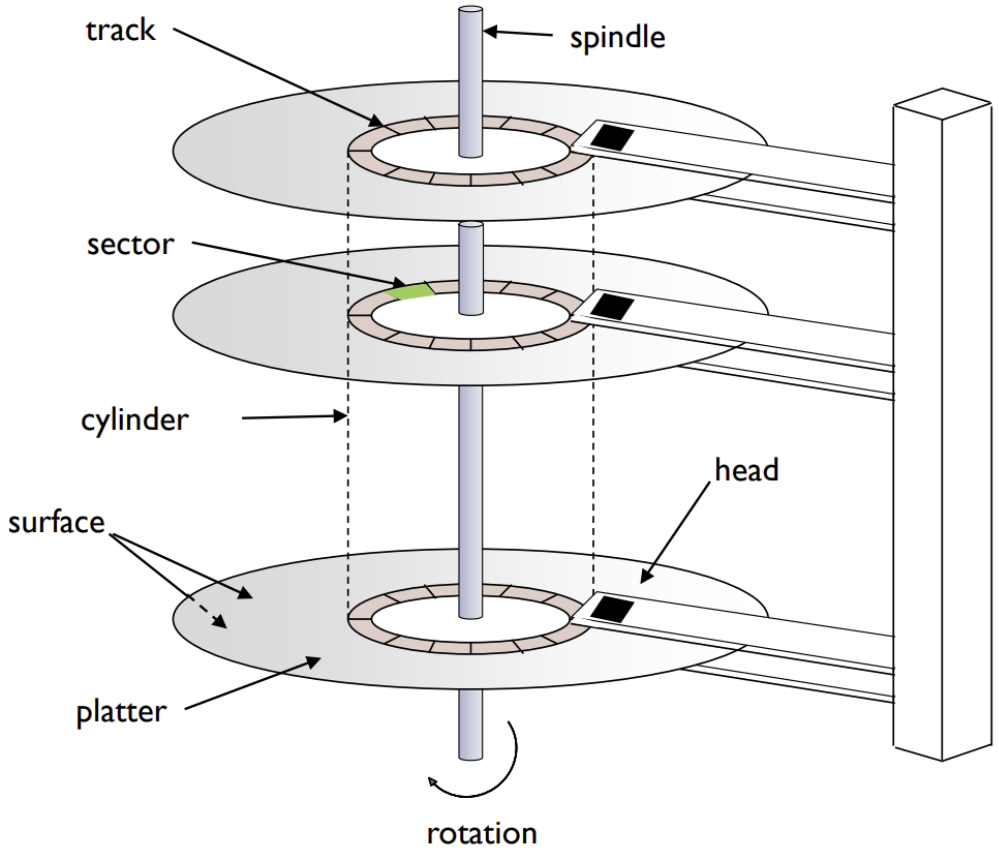
Fonte das Imagens: IBM

Por mais de 30 anos, a fita magnética dominou o armazenamento offline e a transferência de dados. Bobinas magnéticas grandes girando em movimentos rápidos e alternados tornaram-se um ícone cultural. Imagens de unidades de fita em movimento foram amplamente utilizadas para representar computadores em filmes e televisão.

Fita Magnética

- Meio de armazenamento secundário inicial.
  - Evoluiu de bobinas abertas para cartuchos.
  - Tempo de acesso lento.
  - Acesso aleatório é aproximadamente 1000 vezes mais lento que o acesso a discos.
- Principalmente utilizada para backup, armazenamento de dados de uso infrequente e como meio de transferência entre sistemas.
- Uma vez que os dados estejam sob o cabeçote de leitura/escrita, as taxas de transferência são comparáveis às dos discos (140+ MB/s).
- Capacidade de armazenamento típica varia entre 200 GB e 1,5 TB.

Moving-head Disk Mechanism



Fonte da imagem: Shy @ Verlog.io - Ver também How Hard Drives Work @ Youtube

Hard Disk Drives (HDDs)

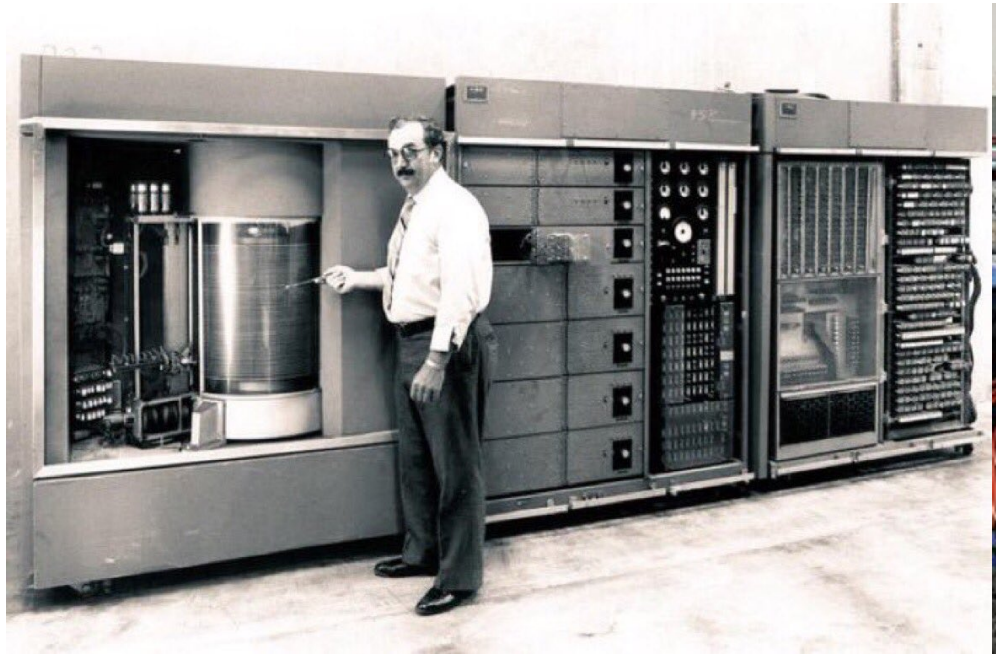
- O diâmetro dos pratos varia de 0.85" a 14" (historicamente).
  - Os tamanhos mais comuns são 3.5", 2.5", e 1.8".
- A capacidade varia de 30 GB a 3 TB por unidade.
- Desempenho:
  - Taxa de Transferência: 6 Gb/s (teórica), 1 Gb/s (real).
  - Tempo de Busca de 3 ms a 12 ms: 9 ms é comum para unidades de desktop.
  - O tempo médio de busca é medido ou calculado com base em 1/3 das trilhas.
  - Latência dependente da velocidade do spindle

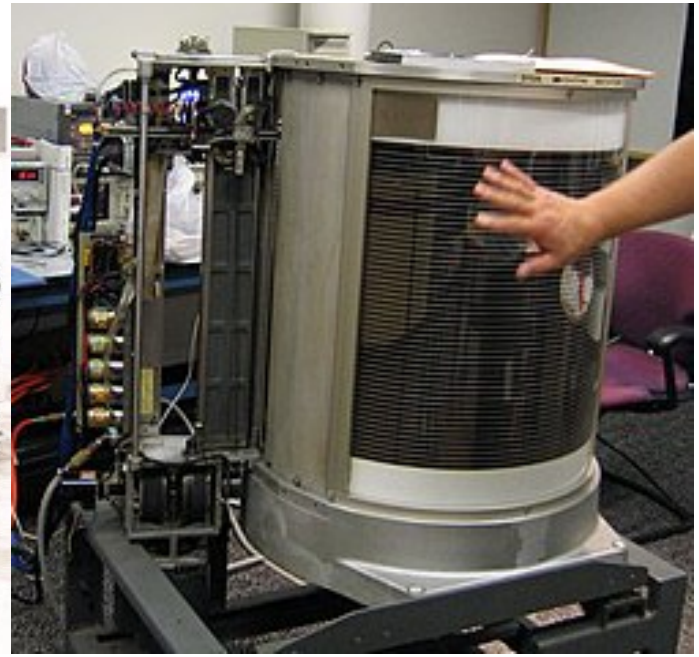
Desempenho de Unidades de Disco Rígido (HDDs)

- Tempo médio de acesso = tempo médio de busca + latência média.
  - Para um disco rápido: 3 ms + 2 ms = 5 ms.
  - Para um disco lento: 9 ms + 5,56 ms = 14,56 ms.
- Tempo médio de I/O = tempo médio de acesso + (quantidade a transferir / taxa de transferência) + sobrecarga do controlador.
- Exemplo: transferir um bloco de 4 KB em um disco de 7200 RPM com um tempo médio de busca de 5 ms e uma taxa de transferência de 1 Gb/s com uma sobrecarga do controlador de 0,1 ms = 5 ms + 4,17 ms + 0,1 ms + tempo de transferência.
  - Tempo de Transferência = *(4 KB / 1 Gb/s) * (8 Gb / GB) * (1 GB / 1024 KB) = 32 / 1024 = 0,031 ms*
  - Tempo médio de I/O para bloco de 4 KB = *9,27 ms + 0,031 ms = 9,301 ms*

O Primeiro Disco Rígido Comercial

- Em 1956, o computador IBM RAMDAC incluiu o sistema de armazenamento em disco IBM Modelo 350
- Capacidade de 5 milhões (7 bits) de caracteres, com 50 pratos de 50 x 24", e um tempo de acesso < 1s

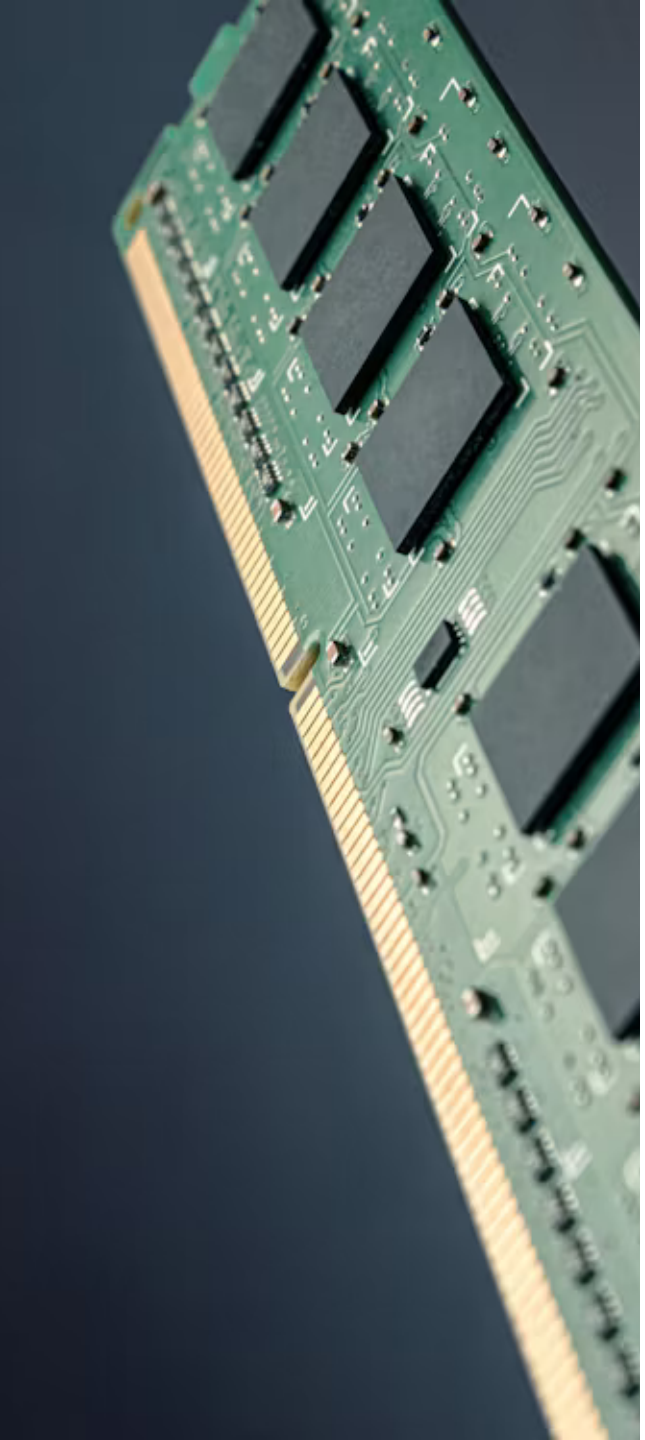




Fonte das imagens: IBM e Wikipedia

Dispositivos de Memória Não Volátil

- Quando semelhantes às unidades de disco rígido: **discos de estado sólido (SSDs)**
- Outras formas incluem dispositivos USB (pen/flash drives), substitutos de DRAM (memória dinâmica de acesso aleatório) montados na placa-mãe e armazenamento principal em dispositivos como smartphones
- Podem apresentar maior confiabilidade do que HDDs
- Apresentam menor capacidade de armazenamento, mas mais veloz
- São mais caros por MB (megabyte). Podem ter vida útil mais curta
- Ausência de partes móveis elimina o tempo de busca e a latência rotacional



Memória Volátil

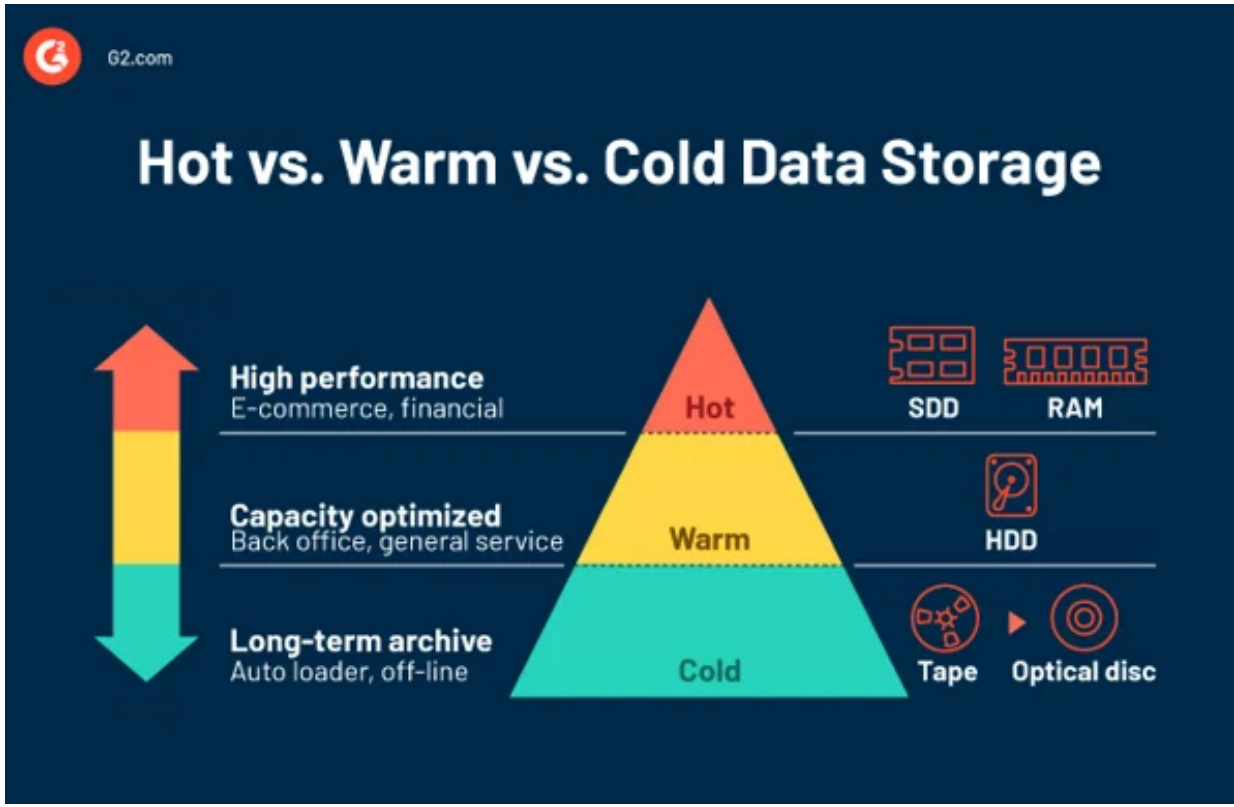
- A DRAM (Memória Dinâmica de Acesso Aleatório) é frequentemente utilizada como dispositivo de armazenamento em massa.
  - Tecnicamente não se qualifica como armazenamento secundário devido à sua volatilidade, mas pode apresentar sistemas de arquivos e ser utilizada como um armazenamento secundário extremamente rápido.
- Unidades RAM (conhecidas por diversos nomes, incluindo discos RAM) são apresentadas como dispositivos de bloco brutos, geralmente formatados com um sistema de arquivos.

Memória Volátil (cont.)

- Os computadores já possuem buffers e caches implementados em RAM, então qual a razão para RAM drives?
  - Caches/buffers alocados/gerenciados por programadores, sistemas operacionais ou hardware.
  - Unidades RAM estão sob controle do usuário.
  - Encontradas em todos os principais sistemas operacionais: Linux (`/dev/ram`), macOS (`diskutil` para criação), Linux (`/tmp` com sistema de arquivos `tmpfs`).
- Utilizadas como armazenamento temporário de alta velocidade.
  - Programas podem compartilhar grandes volumes de dados rapidamente, através da leitura/escrita em uma unidade RAM.

Data Tiering

- **Quente**: para dados estruturados e acessados com muita frequência.
  - Utilizados por funcionários ou clientes
- **Morno**: para dados estruturados e acessados com frequência moderada.
  - Utilizados para relatórios ou análises
- **Frio**: para dados estruturados ou não estruturados que são acessados com pouca frequência.
  - Motivos de conformidade legal
- Fonte da imagem: G2
- Leia o paper 📄: Data Tiering in Heterogeneous Memory Systems



Armazenamento nos Videogames

- **Final Fantasy VI (1994)**: O jogo apresenta elementos notáveis, incluindo gráficos de *16 bits* de ótima qualidade (que inspira muitos jogos ainda hoje), personagens cativantes e memoráveis que acompanham o jogador por cerca de 40 horas.
- **Tamanho**: *2,14 MB*.
- Fonte: The Gamer
- Fonte da Imagem: Flicker

Armazenamento nos Videogames

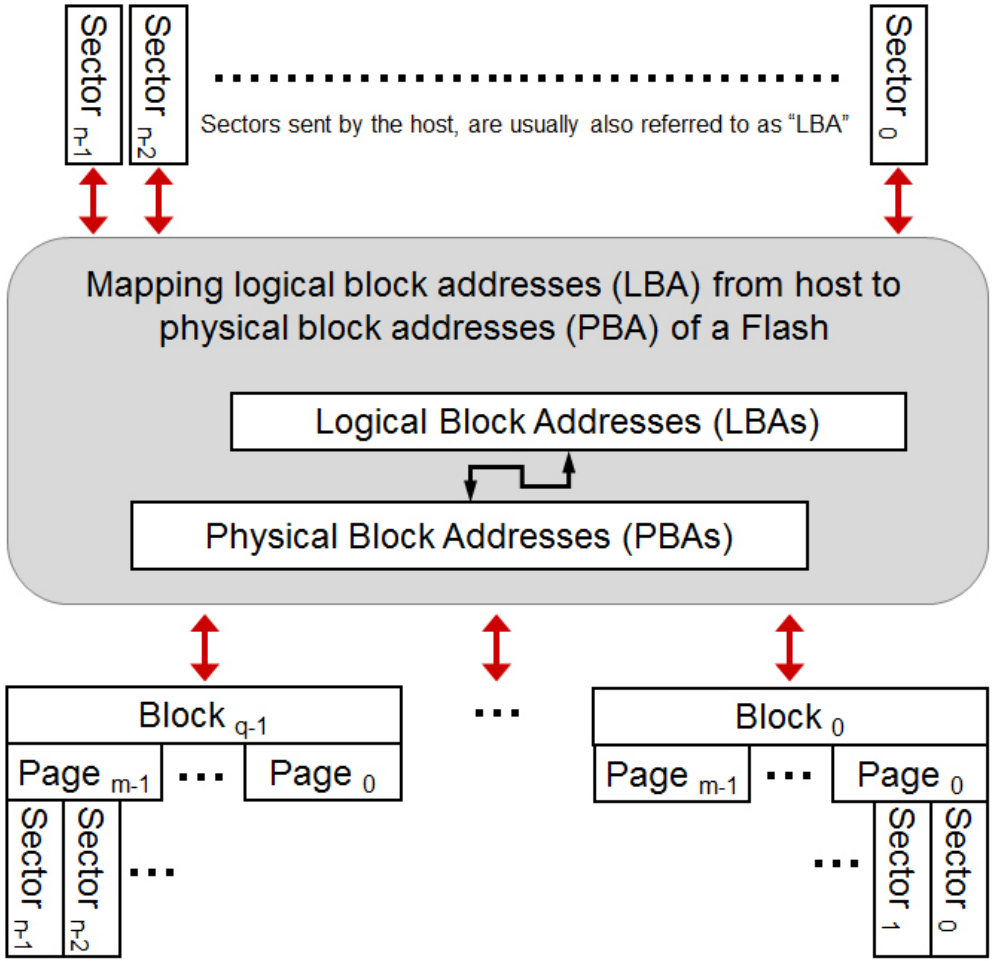
- **Final Fantasy VII (1997)**: O tamanho do jogo é tão grande que ele se estende por três discos diferentes.
- **Tamanho**: Cada disco possui cerca de *450 MB*, totalizando *1317 MB*.
- Fonte: The Gamer.
- Fonte da imagem: Arcade Art Work

Estrutura do Disco

- Um disco pode ser subdividido em **partições**.
- Discos ou partições podem ser protegidos contra falhas através de **RAID**.
  - Partições: minidiscos ou fatias.
  - Disco ou partição pode ser utilizado no estado **bruto** – sem um sistema de arquivos – ou **formatado** com um sistema de arquivos.
- A entidade contendo o sistema de arquivos é conhecida como **volume**. Cada volume rastreia informações sobre esse sistema em um **diretório do dispositivo** ou uma **tabela de volumes**, indicando seu conteúdo.
- Fonte da imagem: diskinternals.com

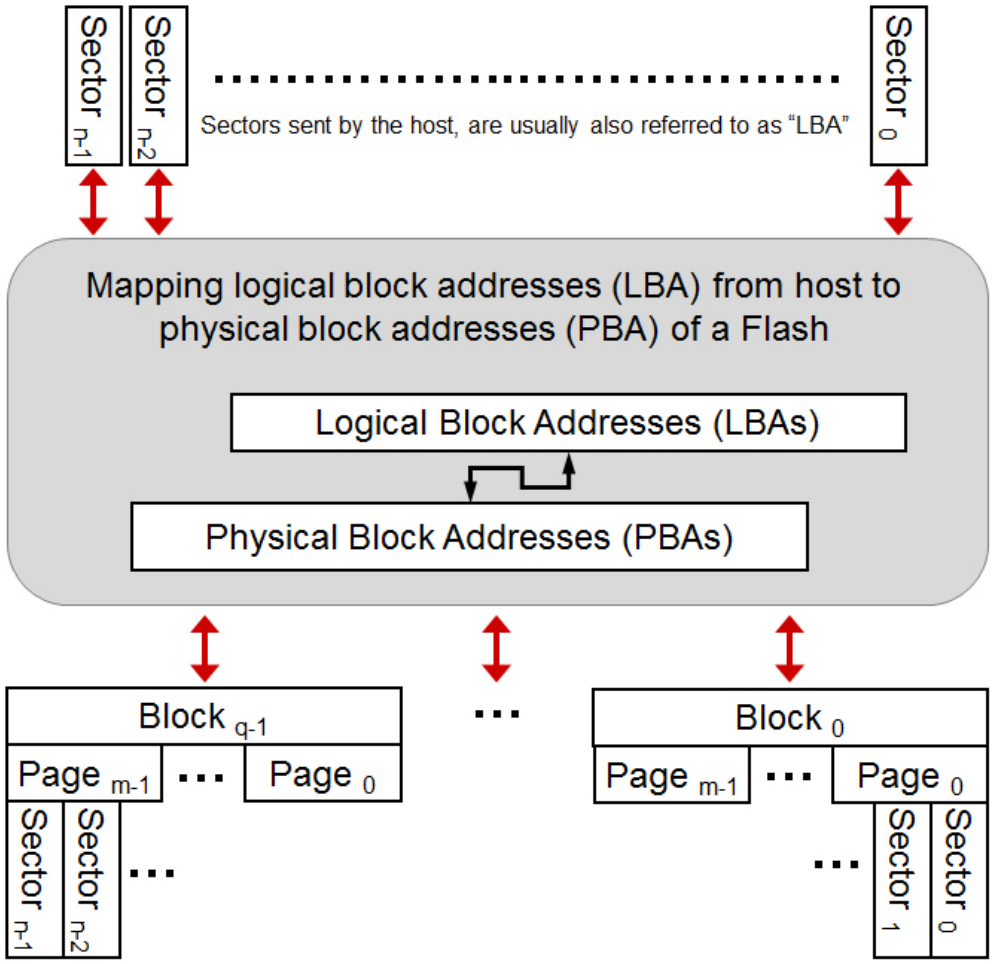
Mapeamento de Endereços

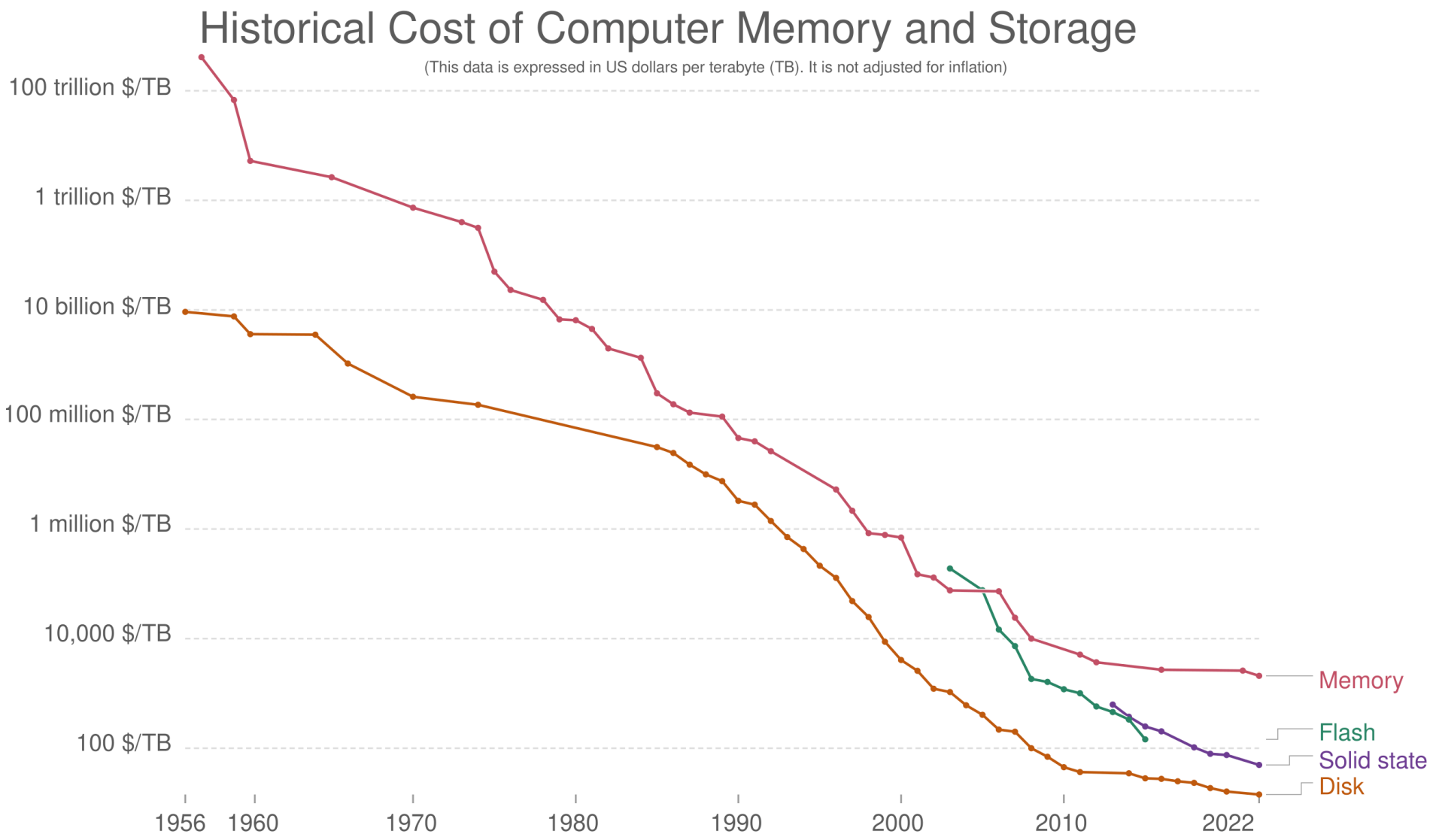
- As unidades de disco rígido são endereçadas como grandes matrizes unidimensionais de blocos lógicos, onde o bloco lógico é a menor unidade de transferência.
- A formatação de baixo nível cria blocos lógicos no meio físico.
- Essa matriz unidimensional de blocos lógicos é mapeada sequencialmente para os setores do disco.
- Fonte da imagem: electronicspecifier



Mapeamento de Endereços

- O setor 0 corresponde ao primeiro setor da primeira trilha no cilindro mais externo.
- O mapeamento prossegue em ordem através dessa trilha, depois pelas demais trilhas nesse mesmo cilindro e, finalmente, pelos demais cilindros, do mais externo para o mais interno.
- A conversão de endereços lógicos para físicos deveria ser direta.
- Exceto pela presença de setores defeituosos.
- Devido à velocidade angular constante (CAV), o número de setores por trilha não é constante.
- Fonte da imagem: electronicspecifier





Fonte da imagem: History of hard disk drives @ Wikipedia

Escalonamento de Disco

- O sistema operacional é responsável por utilizar o hardware de forma eficiente – no caso das unidades de disco rígido, isso implica em minimizar o tempo de acesso e maximizar a largura de banda do disco.
- Minimizar o tempo de busca.
- Tempo de busca é aproximadamente proporcional à distância de busca.
- A largura de banda (*bandwidth*) do disco é definida como o número total de bytes transferidos dividido pelo tempo total entre o primeiro pedido de serviço e a conclusão da última transferência.

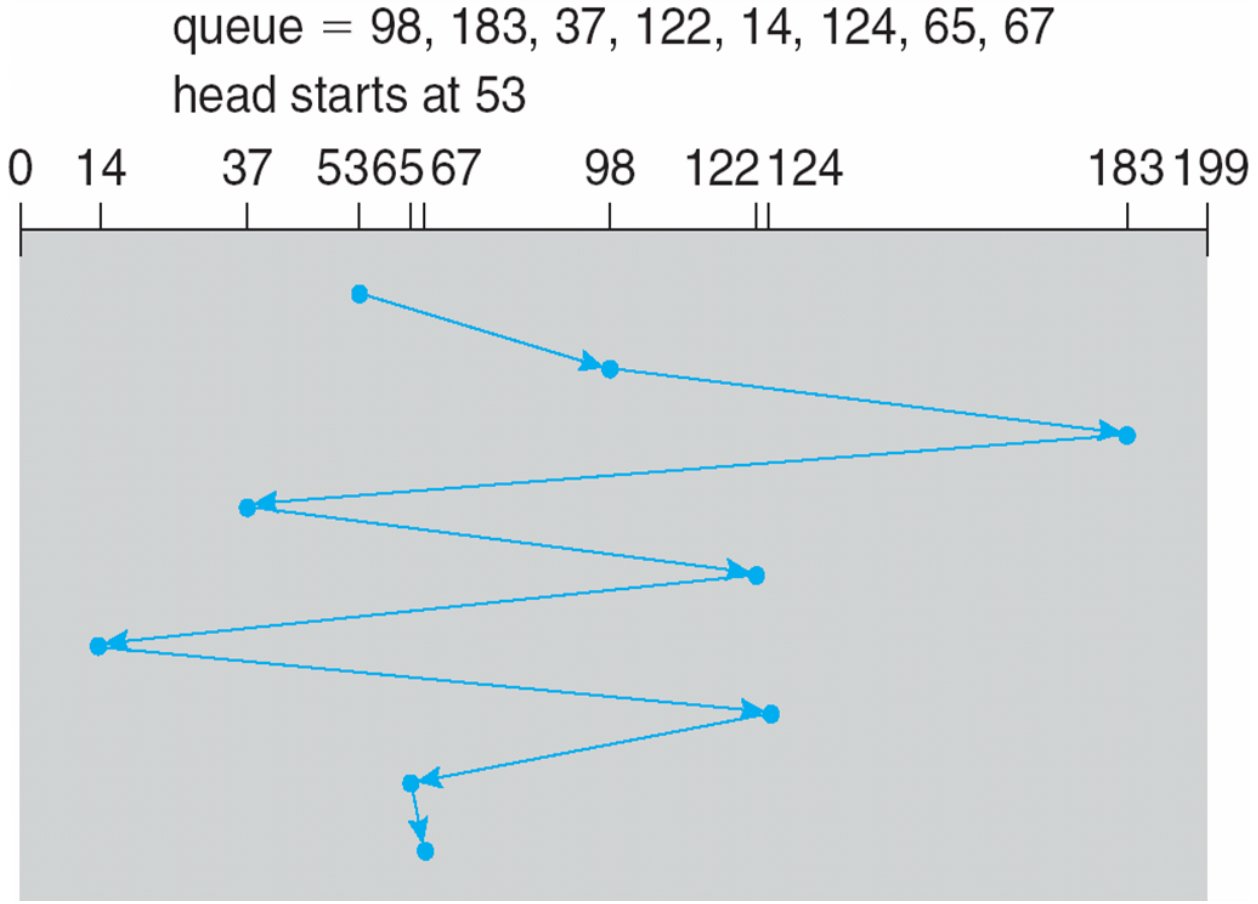
Escalonamento de Disco (cont.)

- Existem diversas fontes de requisições de E/S do disco:
  - Processos do sistema
  - Processos do usuário
- Uma requisição de E/S inclui o modo de entrada ou saída, endereço do disco, endereço da memória e número de setores a serem transferidos.
- O SO mantém uma fila de requisições, por disco ou dispositivo.
  - Um disco inativo pode trabalhar imediatamente em uma requisição de E/S; um disco ocupado significa que o trabalho deve ser enfileirado.
- Note que os controladores de disco possuem buffers pequenos e podem gerenciar uma fila de requisições de E/S (de "profundidade" variável).

Escalonamento de Disco (cont.)

- No passado, o sistema operacional era responsável pelo gerenciamento da fila e pelo agendamento dos cabeçotes do disco rígido. Atualmente, isso está integrado aos dispositivos de armazenamento, controladores.
  - Atualmente, integrado aos dispositivos de armazenamento, controladores.
  - Apenas fornecem endereços lógicos de bloco (LBAs) e lidam com a ordenação das requisições.
- Diversos algoritmos existem para agendar a execução das requisições de E/S do disco.

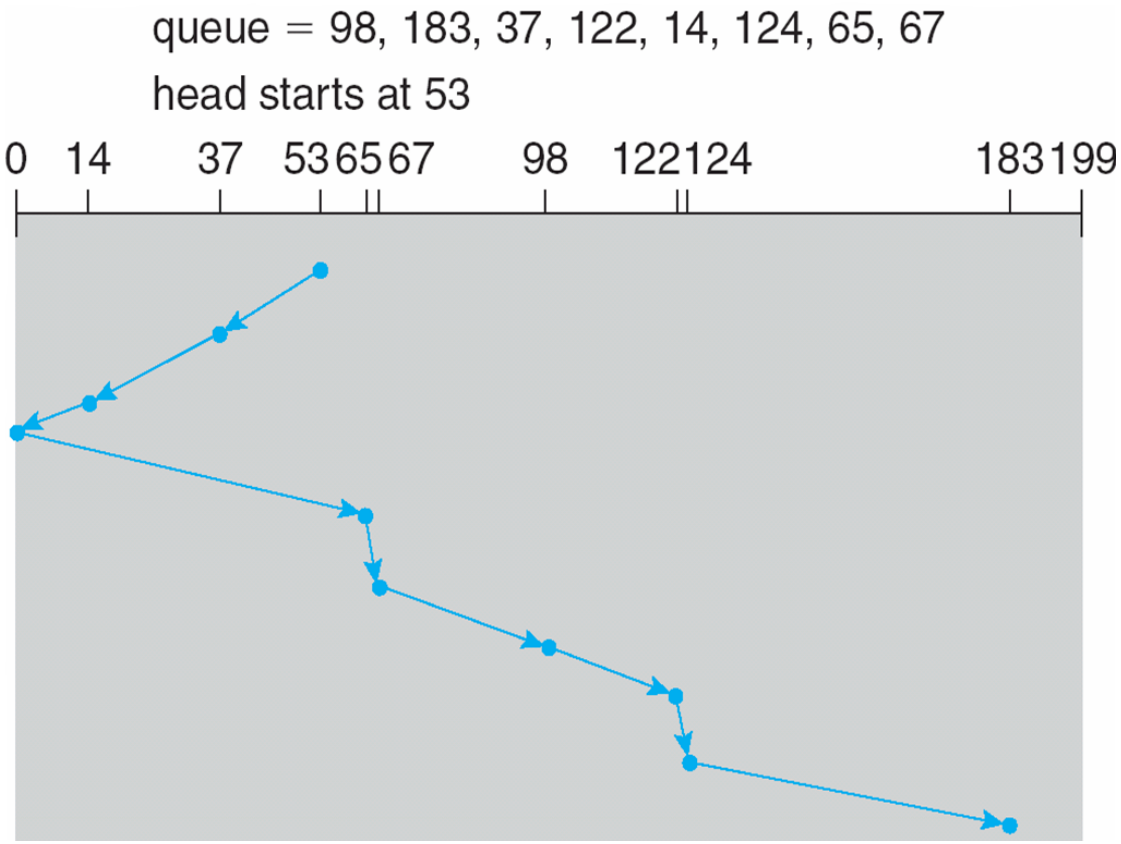
FCFS



Movimento total do cabeçote de 640 cilindros. Fonte: A. Silberschatz *et. al*, *Operating Systems Concepts*.

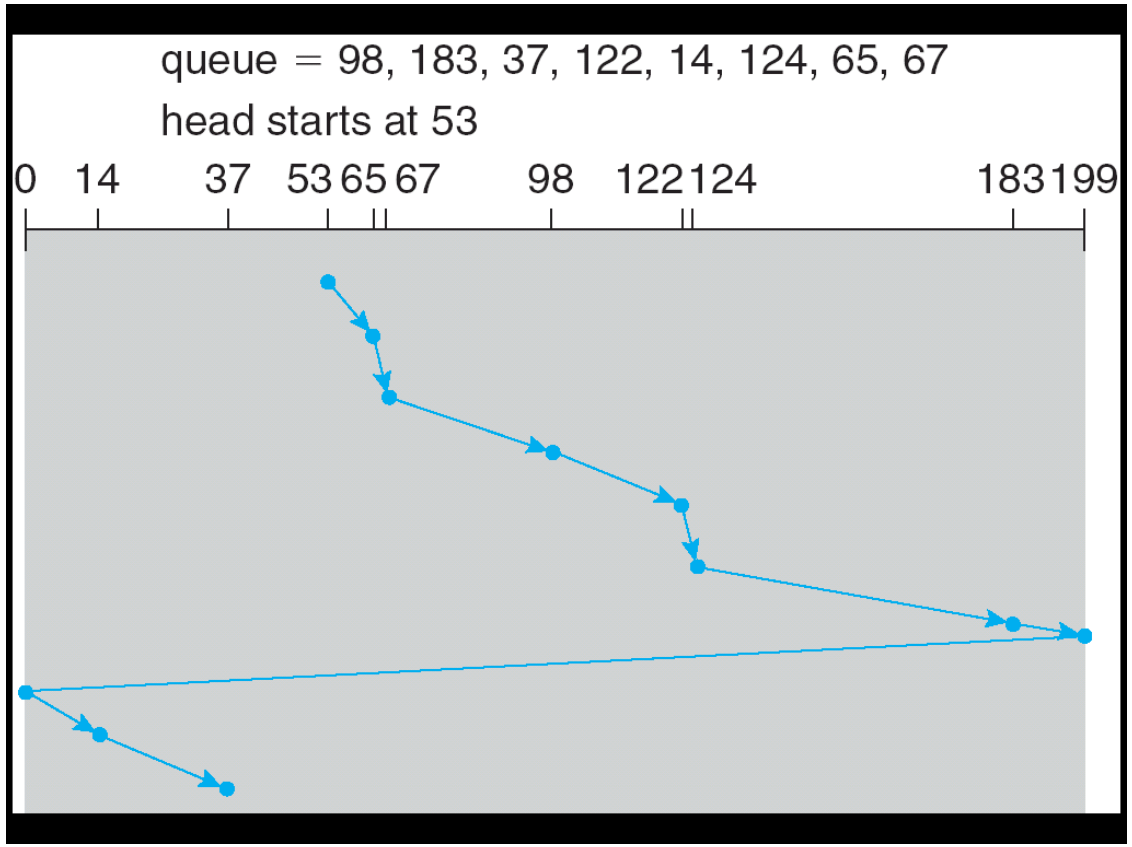
SCAN (algoritmo do elevador)

- O braço do disco inicia em uma extremidade do disco e move-se em direção à outra extremidade, atendendo às requisições até chegar ao final oposto, onde o movimento do cabeçote é invertido (repete a operação).
- As requisições forem densamente distribuídas uniformemente, a maior densidade estará na outra extremidade do disco e essas requisições aguardarão o tempo mais longo.
- Na figura: movimento total do cabeçote de 208 cilindros. Fonte: A. Silberschatz *et. al*, *Operating Systems Concepts*.



C-SCAN

- O cabeçote move-se de uma extremidade do disco à outra, atendendo às requisições durante o percurso.
- Ao atingir a outra extremidade, ele retorna imediatamente ao início do disco, sem atender nenhuma requisição na viagem de volta.
- Trata os cilindros como uma lista circular que se envolve do último cilindro para o primeiro.
- Fonte da imagem: A. Silberschatz *et. al*, *Operating Systems Concepts*.



Seleção de Algoritmo

- Como qualquer algoritmo de escalonamento, o desempenho depende muito da quantidade dos dados e dos tipos de solicitação
- SCAN e C-SCAN são menos prováveis de gerar problema de inanição, quando uma dada requisição nunca é atendida
- Métricas relevantes:
  - Tempo de busca (seek time): Tempo necessário para posicionar o cabeçote na trilha desejada
  - Tempo de latência rotacional (Latency time): Tempo necessário para atingir o início do setor a ser lido/escrito
  - Tempo para escrita/leitura efetiva dos dados

Gerenciamento de Dispositivos de Armazenamento

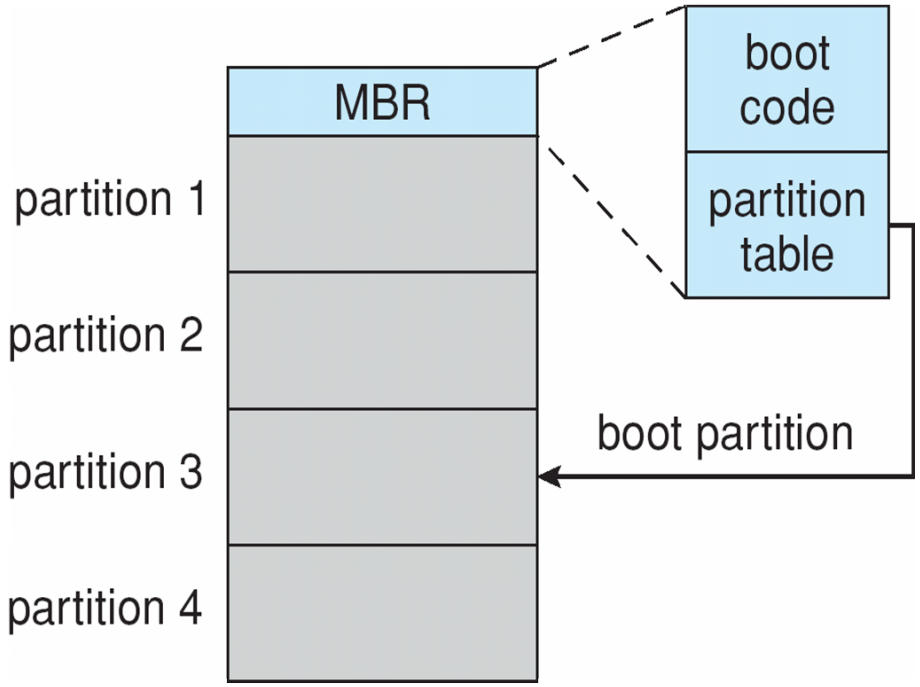
- **Formatação de baixo nível** (**formatação física**): Divide um disco em setores que o controlador do disco pode ler e escrever.
  - Setor pode conter informações de cabeçalho, dados e código de correção de erros (ECC).
  - Geralmente possui 512 bytes de dados, mas pode ser selecionável.
- Para utilizar um disco para armazenar arquivos, o sistema operacional ainda precisa registrar suas próprias estruturas de dados no disco.
  - **Particiona** o disco em um ou mais grupos de cilindros, cada um tratado como um disco lógico.
  - **Formatação lógica**, ou "criação de um sistema de arquivos".
  - Para aumentar a eficiência, a maioria dos sistemas de arquivos agrupa blocos em **clusters**.
    - Operações de E/S do disco são realizadas em blocos.
    - Operações de E/S de arquivo são realizadas em clusters.

Gerenciamento de Dispositivos de Armazenamento (cont.)

- A **partição raiz** contém o sistema operacional. Outras partições podem conter outros sistemas operacionais, outros sistemas de arquivos ou estar em estado bruto.
  - São montadas durante a inicialização do sistema.
  - Outras partições podem ser montadas automaticamente ou manualmente.
- Durante a montagem, a consistência do sistema de arquivos é verificada.
  - Verifica-se se todos os metadados estão corretos.
    - Se não estiverem, tenta-se corrigi-los e repetir o processo.
    - Caso contrário, adiciona-se à tabela de montagem e permite-se o acesso.
- O bloco de inicialização pode apontar para o volume de inicialização ou um conjunto de blocos que contenham código suficiente para saber como carregar o kernel a partir do sistema de arquivos.
  - Ou um programa de gerenciamento de boot para inicialização multi-OS.

Bootloader

- Acesso direto ao disco bruto para aplicativos que desejam realizar seu próprio gerenciamento de blocos, evitando a intervenção do sistema operacional (por exemplo, bancos de dados).
- O bloco de inicialização inicializa o sistema.
  - O bootstrap é armazenado em ROM ou firmware.
  - O programa **bootstrap loader** é armazenado nos blocos de inicialização da partição de boot.
- Métodos como o **sector sparing** são utilizados para lidar com blocos defeituosos.
- Na imagem: Inicialização a partir de armazenamento secundário no Windows. Fonte: A. Silberschatz *et. al*, *Operating Systems Concepts*.



Anexação de Disco

- Armazenamento anexado ao host acessado através de portas de E/S comunicando-se com barramentos de E/S.
- Diversos barramentos estão disponíveis, incluindo Ata Attachment Avançado (ATA), Serial ATA (SATA), eSATA, Serial Attached SCSI (SAS), Universal Serial Bus (USB) e Fibre Channel (FC). Mais comum: SATA.
- **NVM Express (NVMe)**: Nova interface rápida para NVM, conectando-se diretamente ao barramento PCI. Velocidade significativamente maior.

Anexação de Disco (cont.)

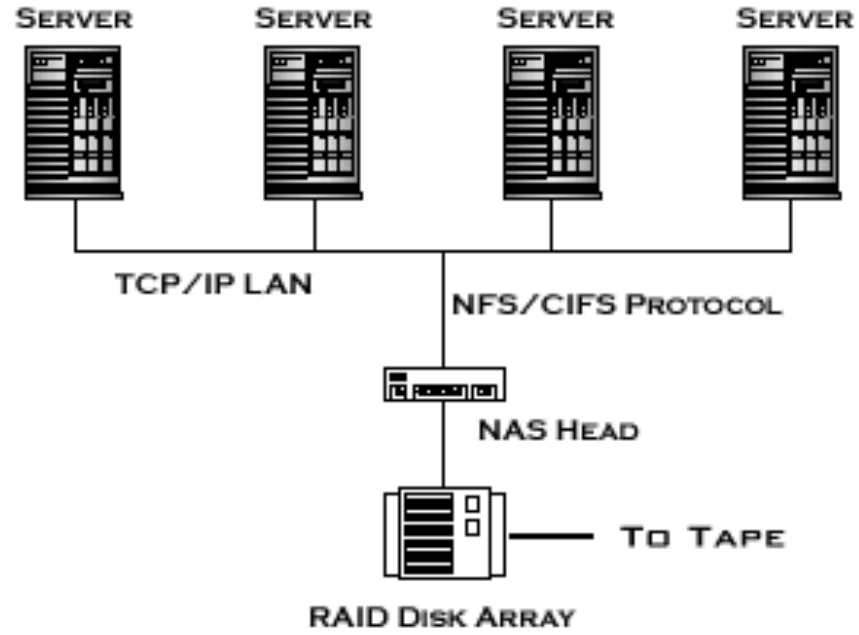
- As transferências de dados em um barramento são realizadas por processadores eletrônicos especializados chamados controladores (ou adaptadores host-barramento, HBAs).
  - Um controlador host está na extremidade do computador do barramento e um controlador de dispositivo está na extremidade do dispositivo.
  - O computador coloca um comando no controlador host, utilizando portas de E/S mapeadas na memória.
  - O controlador host envia mensagens para o controlador do dispositivo.
  - Os dados são transferidos via DMA entre o dispositivo e a DRAM do computador.

Anexação de Armazenamento

- Os computadores acessam o armazenamento de três maneiras:
  - Anexo ao host (local)
  - Anexo à rede
  - Nuvem (cloud)
- O acesso anexo ao host ocorre através de portas de E/S locais, utilizando uma das diversas tecnologias disponíveis.
- Para conectar muitos dispositivos, utilizam-se barramentos de armazenamento como USB, FireWire e Thunderbolt.
- Sistemas de alta performance utilizam Fibre Channel (FC).
  - Arquitetura serial de alta velocidade que utiliza cabos de fibra óptica ou cobre.

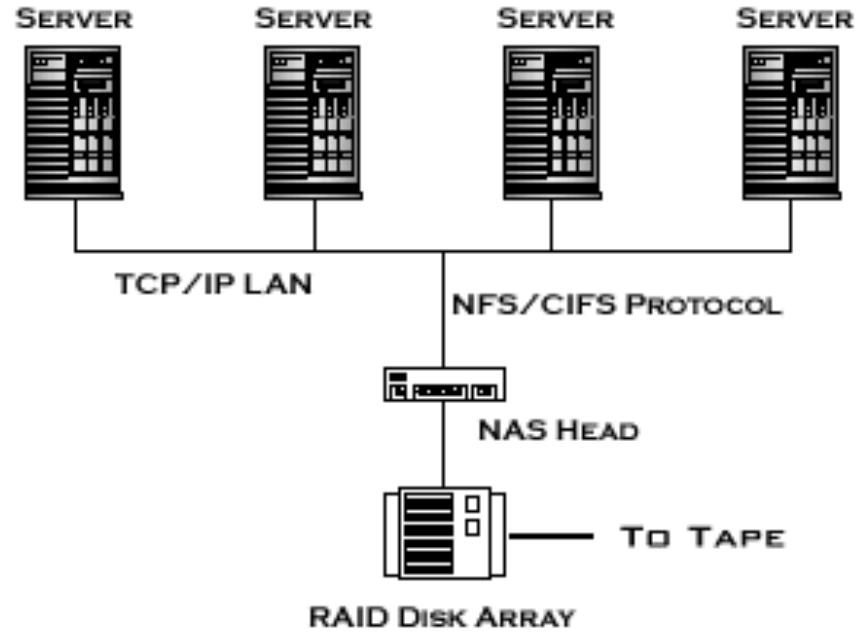
Armazenamento Anexo à Rede (NAS)

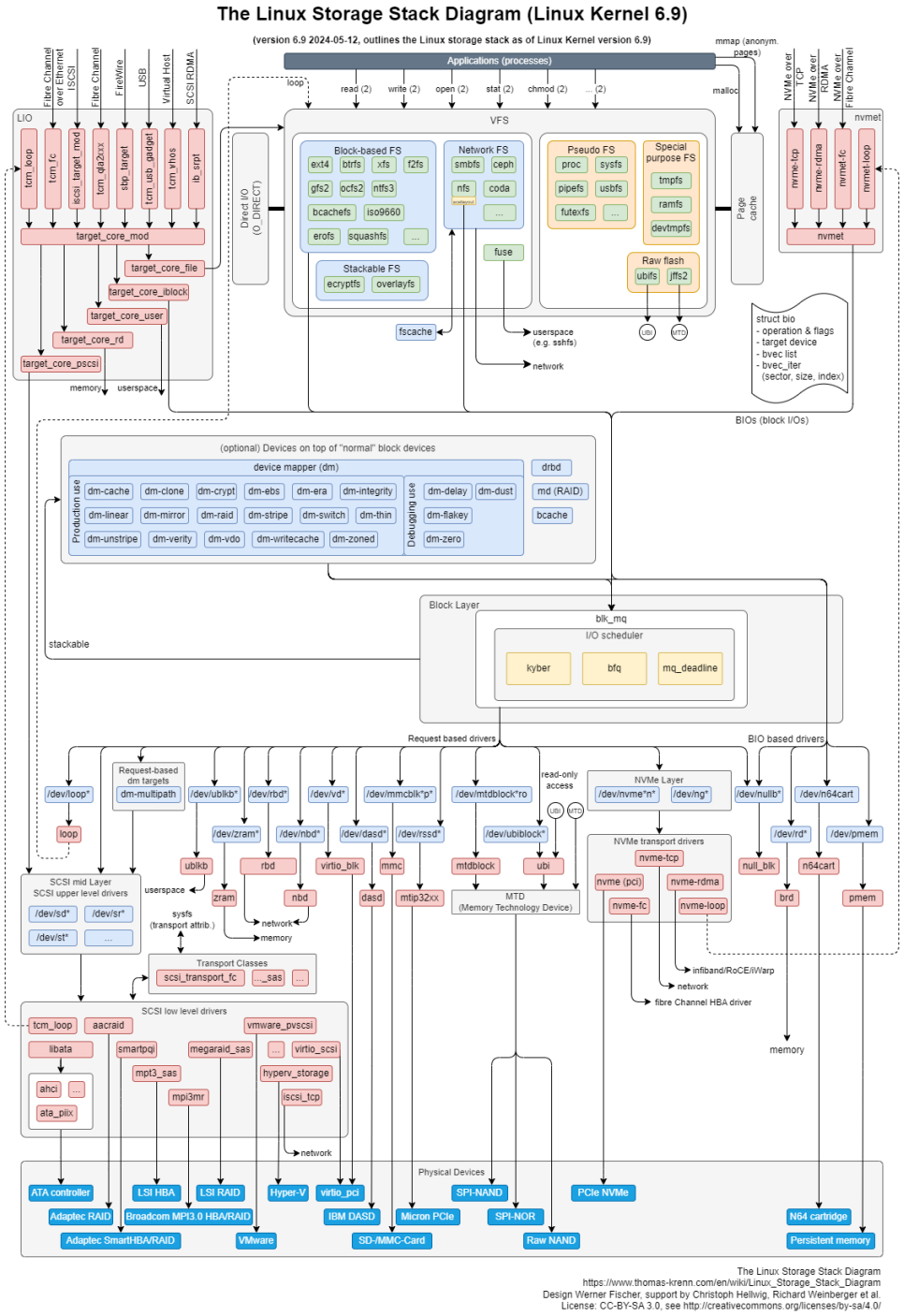
- *Network-Attached Storage (NAS)* é um armazenamento disponibilizado através de uma rede, em vez de por meio de uma conexão local (como um barramento).
  - Permite a anexação remota a sistemas de arquivos.
  - NFS e CIFS são protocolos comuns.
- Tipicamente implementado via chamadas de procedimento remoto (RPCs) entre o host e o armazenamento através de TCP ou UDP em uma rede IP.
- Fonte da imagem: Yitzhak Birk @ ResearchGate



Considerações na Seleção de NAS

- **Capacidade de Armazenamento:** Determine a quantidade de espaço de armazenamento necessária atualmente e no futuro.
- **Escalabilidade:** Procure dispositivos que suportem opções de expansão.
- **Desempenho:** Avalie velocidades de transferência de dados e suporte para configurações RAID.
    - Redundância de dados, snapshots para recuperação, backup





The Linux Storage Stack Diagram
https://www.thomas-krenn.com/en/wiki/Linux_Storage_Stack_Diagram
Design Werner Fischer, support by Christoph Hellwig, Richard Weinberger et al.
License: CC-BY-SA 3.0, see http://creativecommons.org/licenses/by-sa/4.0/

Linux Storage Stack Diagram

- Na imagem: A posição dos caminhos de dados NVMe e múltiplas filas internas dentro das diversas camadas da pilha de armazenamento do kernel Linux.
- Fonte: I/O Scheduling @ Wikipedia

Conclusão

- **Considerações para Sistemas Operacionais:**
  - **Abstração do Hardware:** Como o SO fornece uma interface consistente para diferentes tipos de dispositivos de armazenamento.
  - **Gerenciamento de Espaço Livre:** Alocação e desalocação eficiente de espaço em disco.
  - **Particionamento:** Dividir discos em partições lógicas para organizar dados e permitir múltiplos sistemas operacionais.
  - **Interfaces de Discos:** Visão geral das tecnologias de conexão (SATA, NVMe, USB, Fibre Channel) e suas características
- **Gerenciamento de E/S e Desempenho:**: Como o sistema operacional organiza as requisições para minimizar o tempo de acesso ao disco (abordagem teórica).

Conclusão (cont.)

- **Conceitos Chave:**
  - **Latência:** Tempo de resposta do dispositivo de armazenamento.
  - **Throughput:** Taxa de transferência de dados.
  - **Redundância:** Duplicação de dados para proteção contra falhas.
  - **Escalabilidade:** Capacidade de aumentar a capacidade de armazenamento conforme necessário.

Material Adicional

- Disk Scheduling Algorithms @ Geeks For Geeks 
- Paying for Cloud Storage is Stupid YouTube
- How do hard drives work? YouTube
- What is LTO TAPE?/FUJIFILM YouTube
- Data Tiering in Heterogeneous Memory Systems 

Dúvidas e Discussão