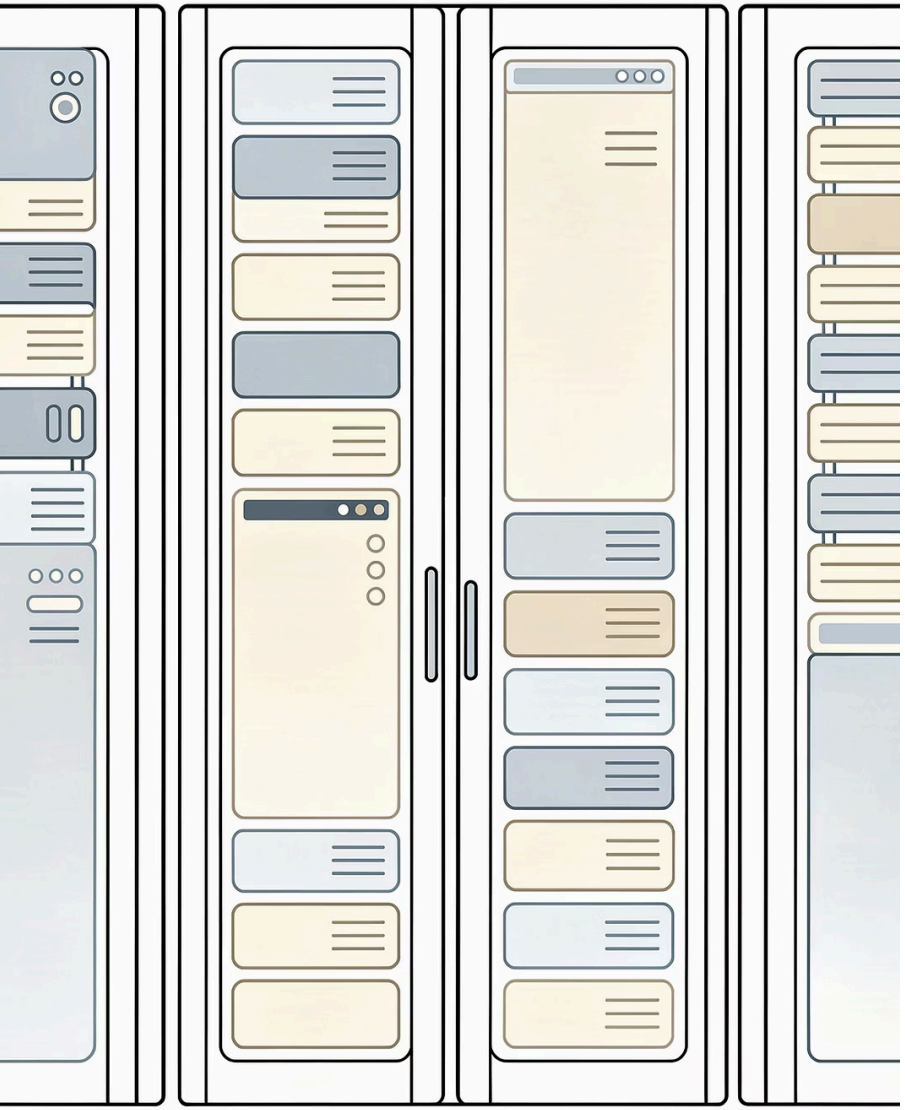


CEFET/RJ

BANCO DE DADOS

Sistemas de Armazenamento Físico

Explore os fundamentos dos sistemas de armazenamento físico em bancos de dados modernos. Desde a hierarquia de memória até tecnologias avançadas como RAID e SSDs, este capítulo oferece uma compreensão abrangente de como os dados são armazenados, acessados e protegidos em sistemas computacionais.

Eduardo Ogasawara

eduardo.ogasawara@cefet-rj.br
https://eic.cefet-rj.br/~eogasawara

Classificação de Mídias de Armazenamento Físico

As mídias de armazenamento podem ser diferenciadas em categorias fundamentais que determinam seu comportamento e adequação para diferentes aplicações. O armazenamento **volátil** perde seu conteúdo quando a energia é desligada, enquanto o armazenamento **não-volátil** mantém os dados mesmo sem alimentação elétrica, incluindo armazenamento secundário, terciário e memória principal com backup de bateria.

Velocidade de Acesso

O tempo necessário para recuperar dados varia drasticamente entre diferentes tecnologias, desde nanossegundos em cache até segundos em fitas magnéticas. Esta característica determina a adequação para aplicações específicas.

Custo por Unidade

O custo por byte ou gigabyte influencia diretamente decisões arquiteturais. Tecnologias mais rápidas geralmente têm custos significativamente mais elevados por unidade de armazenamento.

Confiabilidade

A capacidade de manter dados intactos ao longo do tempo e resistir a falhas é crucial. Diferentes mídias oferecem níveis variados de durabilidade e resistência a condições adversas.

Hierarquia de Armazenamento e Tempo de Acesso

A hierarquia de armazenamento representa uma organização estratificada de tecnologias de memória, equilibrando velocidade, capacidade e custo.

A velocidade de acesso varia exponencialmente através da hierarquia, desde nanossegundos em cache até minutos em sistemas de fita offline. Compreender essas diferenças é fundamental para otimização de desempenho.

1. **Armazenamento Primário**

   Registradores e cache do processador

2. **Armazenamento Volátil**

   Memória principal (RAM)

3. **Armazenamento Não-Volátil**

   Flash e memória persistente

4. **Armazenamento Secundário**

   Discos magnéticos e SSDs

5. **Armazenamento Terciário**

   Fitas magnéticas e mídia óptica

1. **Cache: 1-5 nanossegundos**

   Acesso extremamente rápido, linha de cache por vez

2. **Memória Principal: 60-100 nanossegundos**

   Acesso por linha de cache, ainda muito rápido

3. **Flash/SSD: 20-100 microssegundos**

   Acesso por página, ordens de magnitude mais lento

4. **Disco Magnético: 5-10 milissegundos**

   Acesso por página, inclui tempo de busca mecânica

5. **Fita Magnética Online: 100+ milissegundos**

   Acesso sequencial, tempo de posicionamento significativo

Discos Magnéticos: Arquitetura e Funcionamento

Os discos magnéticos representam a espinha dorsal do armazenamento secundário em sistemas de banco de dados. Compreender sua arquitetura física é essencial para otimizar o desempenho de acesso aos dados.

Componentes Fundamentais

- **Cabeça de leitura/escrita:** Componente que magnetiza e detecta padrões magnéticos na superfície do disco
- **Trilhas (tracks):** Círculos concêntricos na superfície do prato, com 50K-100K trilhas por prato em discos modernos
- **Setores:** Menor unidade de dados que pode ser lida ou escrita, tipicamente 512 bytes
- **Cilindro:** Conjunto da i-ésima trilha de todos os pratos em um mesmo alinhamento vertical

Processo de Leitura/Escrita

1. O braço do disco posiciona a cabeça sobre a trilha correta através de movimento mecânico preciso
2. O prato gira continuamente em alta velocidade (5.400 a 15.000 RPM)
3. Os dados são lidos ou escritos conforme o setor passa sob a cabeça
4. Montagens de múltiplos pratos (1 a 5) compartilham um eixo comum com uma cabeça por superfície

Discos modernos podem ter de 500 a 1.000 setores nas trilhas internas e de 1.000 a 2.000 nas trilhas externas, refletindo a maior circunferência das trilhas periféricas.

Métricas de Desempenho de Discos

Tempo de Acesso

Intervalo desde a solicitação de leitura/escrita até o início da transferência de dados. Composto por tempo de busca e latência rotacional.

Tempo de Busca (Seek Time)

Tempo para reposicionar o braço sobre a trilha correta. O tempo médio de busca é metade do pior caso, variando de 4 a 10 milissegundos em discos típicos.

Latência Rotacional

Tempo para o setor acessado aparecer sob a cabeça. Varia de 4 a 11 milissegundos (5.400 a 15.000 RPM). A latência média é metade desse valor.

A **latência total** varia de 5 a 20 milissegundos dependendo do modelo do disco. A **taxa de transferência de dados** representa a velocidade com que os dados podem ser recuperados ou armazenados, variando de 25 a 200 MB por segundo como taxa máxima, sendo menor para trilhas internas devido à menor densidade linear.

Métricas de Desempenho: Padrões de Acesso

Blocos de Disco

Um **bloco de disco** é a unidade lógica para alocação e recuperação de armazenamento, tipicamente de 4 a 16 kilobytes. Blocos menores resultam em mais transferências do disco, enquanto blocos maiores desperdiçam espaço devido a blocos parcialmente preenchidos.

IOPS: Operações por Segundo

A métrica **IOPS** (I/O Operations Per Second) representa o número de leituras de blocos aleatórios que um disco pode suportar por segundo. Discos magnéticos de geração atual oferecem de 50 a 200 IOPS.

Acesso Sequencial

Solicitações sucessivas são para blocos consecutivos do disco. A busca é necessária apenas para o primeiro bloco, resultando em altas taxas de transferência.

Acesso Aleatório

Solicitações sucessivas são para blocos que podem estar em qualquer lugar do disco. Cada acesso requer uma busca, resultando em taxas de transferência baixas devido ao tempo desperdiçado em buscas.

Armazenamento Flash: Nova Geração

A tecnologia flash revolucionou o armazenamento de dados, oferecendo velocidades drasticamente superiores aos discos magnéticos. Existem dois tipos principais: NOR flash e NAND flash, sendo este último amplamente utilizado para armazenamento devido ao menor custo.

NAND Flash

Requer leitura página por página (512 bytes a 4 KB) com tempo de 20 a 100 microssegundos. Pouca diferença entre leitura sequencial e aleatória.

Limitações de Escrita

Cada página pode ser escrita apenas uma vez. Deve ser apagada para permitir reescrita, impactando a vida útil e desempenho.

Discos de Estado Sólido

Usam interfaces de disco padrão orientadas a blocos, mas armazenam dados em múltiplos dispositivos flash internamente.

Os **SSDs** (Solid State Disks) oferecem taxas de transferência de até 500 MB/seg usando SATA e até 3 GB/seg usando NVMe PCIe, representando um avanço significativo em relação aos discos magnéticos tradicionais.

Métricas de Desempenho de SSDs

Os SSDs oferecem desempenho superior em praticamente todas as métricas quando comparados aos discos magnéticos, especialmente em operações aleatórias e leituras paralelas.

10K	40K	100K	350K
IOPS de Leitura	**IOPS de Escrita**	**IOPS SATA Paralelo**	**IOPS NVMe**
Leituras aleatórias típicas de 4 KB por segundo	Escritas aleatórias típicas de 4 KB por segundo	Leituras de 4 KB com 32 requisições em paralelo (QD-32)	Leituras de 4 KB com QD-32 em NVMe PCIe

> Os SSDs suportam leituras paralelas massivas, permitindo que múltiplas operações de I/O sejam processadas simultaneamente. Escritas paralelas podem atingir 100.000 IOPS com QD-32, com alguns modelos alcançando valores ainda maiores.

Taxa de transferência de dados para leituras/escritas sequenciais: 400 MB/seg para SATA3 e de 2 a 3 GB/seg usando NVMe PCIe. **Discos híbridos** combinam pequena quantidade de cache flash com disco magnético maior, oferecendo compromisso entre desempenho e custo.

RAID: Arrays Redundantes de Discos Independentes

RAID é uma tecnologia fundamental que gerencia grandes números de discos, oferecendo uma visão unificada de um único disco com características superiores. A tecnologia aborda três objetivos principais através do uso coordenado de múltiplos discos físicos.

Alta Capacidade e Velocidade

Usando múltiplos discos em paralelo, sistemas RAID oferecem capacidades que excedem discos individuais e velocidades multiplicadas através de acesso simultâneo.

Alta Confiabilidade

Armazenamento redundante de dados permite recuperação mesmo quando discos individuais falham, protegendo contra perda de dados críticos.

Desafios de Probabilidade

A chance de que algum disco em um conjunto de N discos falhe é muito maior que a chance de um disco específico falhar individualmente.

Exemplo Prático: Um sistema com 100 discos, cada um com MTTF (Mean Time To Failure) de 100.000 horas (aproximadamente 11 anos), terá um MTTF de sistema de apenas 1.000 horas (aproximadamente 41 dias). Isso demonstra a importância crítica de técnicas de redundância.

Melhoria da Confiabilidade via Redundância

A **redundância** é a estratégia fundamental para evitar perda de dados, armazenando informações extras que podem ser usadas para reconstruir dados perdidos em caso de falha de disco.

Espelhamento (Mirroring)

Duplica cada disco físico. Disco lógico consiste em dois discos físicos. Cada escrita é realizada em ambos os discos, enquanto leituras podem ocorrer de qualquer um.

Proteção contra Falhas

Se um disco falhar, os dados permanecem disponíveis no disco espelho. Perda de dados só ocorre se ambos os discos falharem antes do reparo do sistema.

Probabilidade de Perda

A probabilidade do evento combinado é muito pequena, exceto para modos de falha dependentes como incêndio, colapso de edifício ou surtos elétricos.

> **Tempo Médio para Perda de Dados:** Depende do tempo médio de falha e do **tempo médio de reparo**. Por exemplo, MTTF de 100.000 horas e tempo médio de reparo de 10 horas resultam em tempo médio para perda de dados de 500×10^6 horas (ou 57.000 anos) para um par de discos espelhados, ignorando modos de falha dependentes.

Melhoria de Desempenho via Paralelismo

O paralelismo em sistemas de disco busca três objetivos principais que transformam o desempenho de acesso aos dados:

01

Balanceamento de Carga

Distribuir múltiplos acessos pequenos para aumentar o throughput geral do sistema

02

Paralelização de Acessos

Processar grandes acessos em paralelo para reduzir o tempo de resposta

03

Melhoria de Taxa de Transferência

Distribuir dados entre múltiplos discos através de striping para aumentar velocidade

Striping em Nível de Bit

Divide os bits de cada byte entre múltiplos discos. Em um array de oito discos, o bit *i* de cada byte é escrito no disco *i*. Cada acesso pode ler dados a oito vezes a taxa de um único disco, mas o tempo de busca/acesso é pior que para um único disco. Esta técnica não é mais muito utilizada.

Striping em Nível de Bloco

Com n discos, o bloco i de um arquivo vai para o disco $(i \bmod n) + 1$. Requisições para blocos diferentes podem ser executadas em paralelo se os blocos residem em discos diferentes. Uma requisição para uma longa sequência de blocos pode utilizar todos os discos em paralelo.

Níveis RAID: Arquiteturas Fundamentais

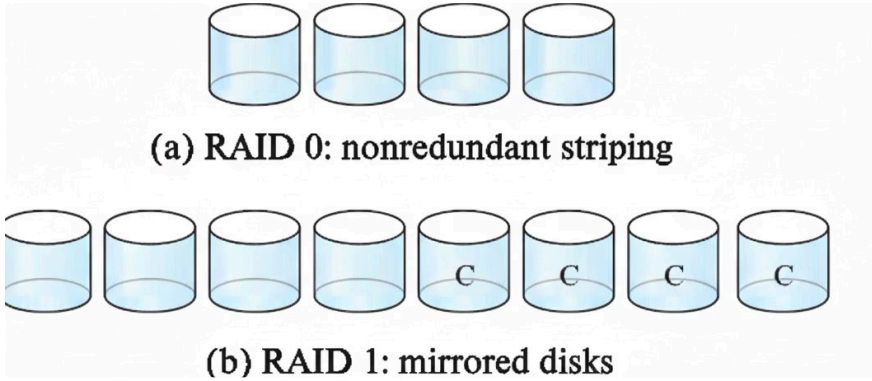
Os níveis RAID definem diferentes estratégias de organização de dados, cada uma com características específicas de desempenho, redundância e custo.

RAID Nível 0

Striping de blocos sem redundância. Utilizado em aplicações de alto desempenho onde a perda de dados não é crítica. Oferece máximo desempenho mas nenhuma proteção contra falhas.

RAID Nível 1

Discos espelhados com striping de blocos. Oferece o melhor desempenho de escrita e alta confiabilidade. Popular para aplicações como armazenamento de arquivos de log em sistemas de banco de dados.



A escolha entre RAID 0 e RAID 1 representa um trade-off fundamental: máximo desempenho sem proteção versus desempenho excelente com redundância completa.

Existem outros níveis RAID (2-6) que oferecem diferentes combinações de striping, espelhamento e paridade, cada um otimizado para casos de uso específicos em termos de desempenho, custo e confiabilidade.

Escolha do Nível RAID Apropriado

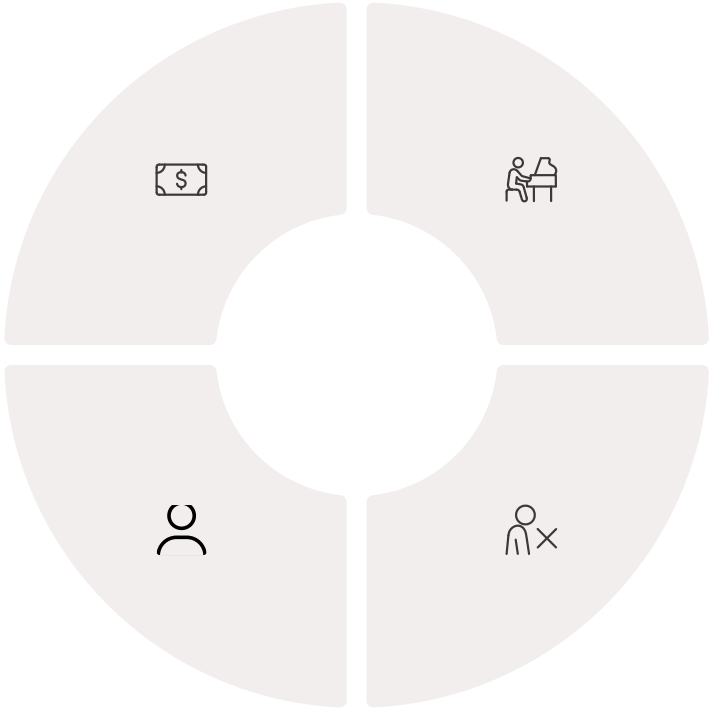
A seleção do nível RAID adequado requer análise cuidadosa de múltiplos fatores que impactam tanto a operação normal quanto cenários de falha e recuperação.

Custo Monetário

Investimento inicial em hardware e custos operacionais contínuos

Desempenho Normal

Número de operações de I/O por segundo e largura de banda durante operação regular



Desempenho em Reconstrução

Tempo e impacto no sistema durante a reconstrução de disco falhado

Desempenho em Falha

Capacidade de manter operações durante falha de disco

> **RAID 0** é usado apenas quando a segurança dos dados não é importante, por exemplo, quando os dados podem ser rapidamente recuperados de outras fontes ou quando são temporários por natureza. Em ambientes de produção críticos, níveis RAID com redundância são essenciais.

Otimização de Acesso a Blocos de Disco

Diversas técnicas avançadas podem ser empregadas para otimizar o desempenho de acesso a discos, minimizando latências e maximizando throughput em sistemas de banco de dados.

Buffering

Uso de buffer em memória para armazenar em cache blocos de disco frequentemente acessados, reduzindo drasticamente acessos físicos ao disco.

Read-ahead

Leitura antecipada de blocos extras de uma trilha, prevendo que serão solicitados em breve baseado em padrões de acesso sequencial.

Disk-arm-scheduling

Algoritmos que reordenam requisições de blocos para minimizar movimento do braço do disco, como o algoritmo do elevador que otimiza o caminho de busca.

Requisições Espalhadas

Múltiplas requisições (R1-R6) chegam para diferentes trilhas do disco, em ordem aleatória.

Direção Definida

O braço do disco inicia o movimento em uma direção (ex: para cima), atendendo às requisições que encontra pelo caminho.

Atendimento Sequencial

O braço atende às requisições (ex: R1, R3, R5) na ordem física de suas trilhas, minimizando movimentos bruscos.

Inversão de Sentido

Ao chegar ao limite do disco ou à última requisição em sua direção atual, o braço inverte o sentido.

Movimento Otimizado

A varredura contínua e ordenada das trilhas resulta em um deslocamento total do braço significativamente menor, otimizando o acesso.

A combinação dessas técnicas pode resultar em melhorias de desempenho de ordem de magnitude, especialmente em workloads com alta localidade de acesso ou padrões previsíveis de leitura/escrita.

Referências

Elmasri & Navathe

Fundamentals of Database Systems

Pearson, 2016

Referência abrangente sobre fundamentos de sistemas de bancos de dados, cobrindo aspectos teóricos e práticos.

Korth, Sudarshan & Silberschatz

Database System Concepts

McGraw-Hill, 2019

Texto fundamental que serviu como base para a maioria dos exemplos apresentados nesta apresentação.

Özsu & Valduriez

Principles of Distributed Database Systems

Springer Nature, 2019

Obra especializada em sistemas de bancos de dados distribuídos, essencial para compreensão avançada.

Nota: Os conceitos e exemplos apresentados baseiam-se principalmente na literatura clássica de sistemas de bancos de dados, em especial *Database System Concepts* e *Fundamentals of Database Systems*.