Estrutura de

Armazenamento e de

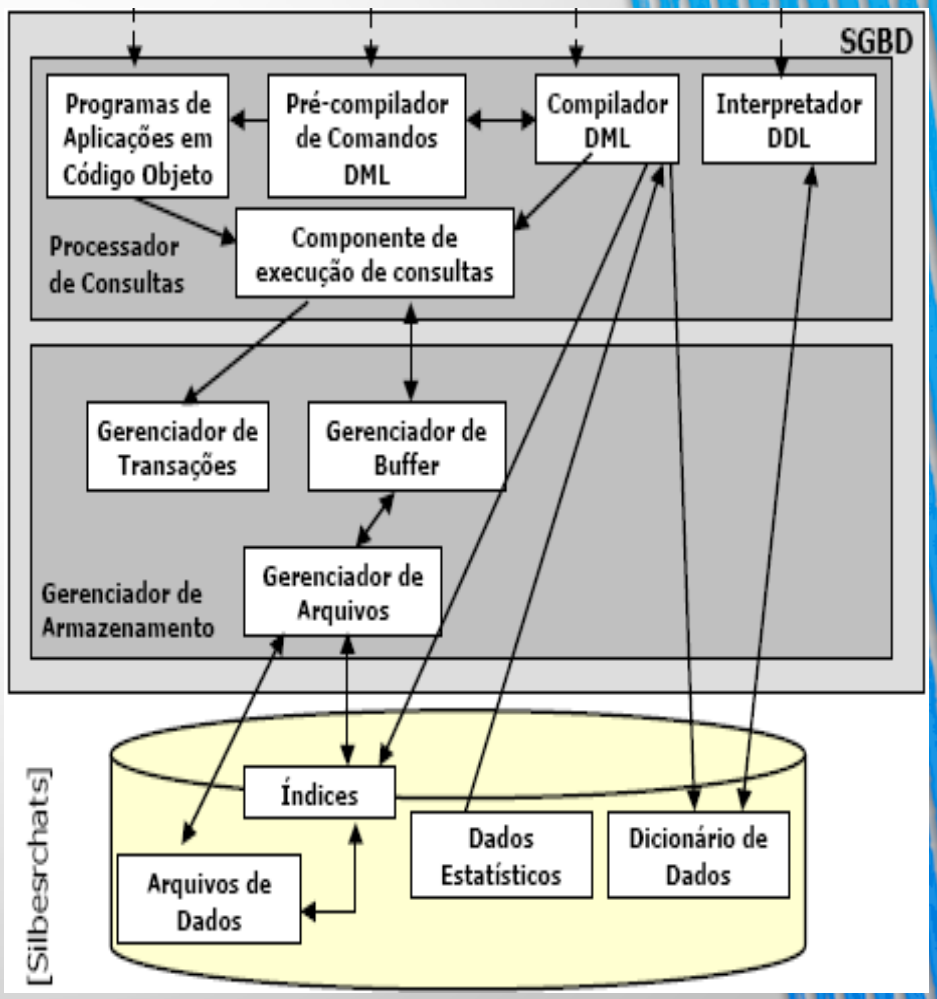
Arquivos

Objetivos

- Que tipos de memória existem num computador?
- Quais são as características físicas dos discos rígidos e das fitas e como que afetam o design de sistemas de bancos de dados?
- O que são os sistemas RAID de memória e quais são as suas vantagens?
- Como é que um SGBD registra o espaço em disco?
- Como é que um SGBD lê e modifica os dados em disco? Qual é o significado de um bloco enquanto unidade de armazenamento e transferência de dados?
- Como é que um SGBD cria e mantém arquivos de registros? Como é que os registros estão organizados em blocos, e como estão os blocos organizados dentro de um arquivo?

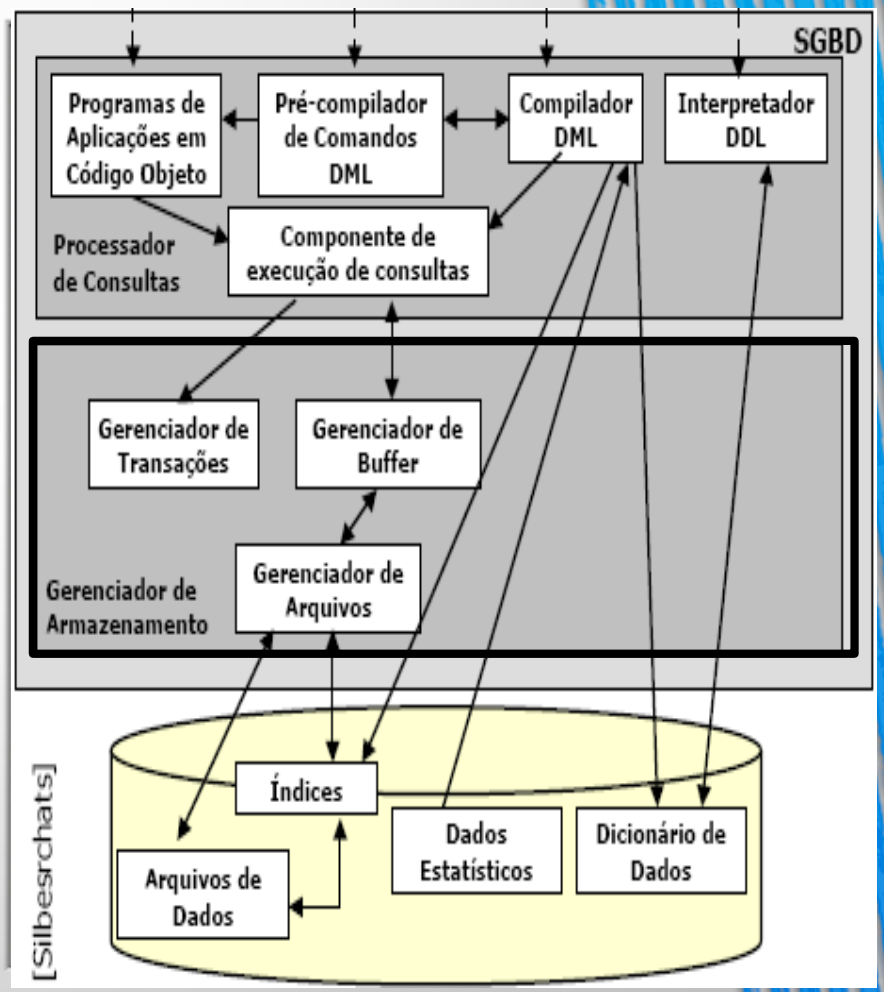
Estrutura Simplificada de um SGBD

- Bancos de Dados armazenam grandes quantidades de dados por períodos longos de tempo em meios de armazenamento secundário
- Um SGBD provê geralmente várias opções para organização física dos dados.
- Projeto Físico de Banco de Dados:
- Busca determinar o melhor tipo de organização dos dados, dentre todas as possíveis, para uma determinada aplicação;
- Especificar o Modelo Físico de Banco de Dados, levando em consideração o Modelo de Dados lógico e informações sobre volumes, acessos e necessidade de disponibilidade
- Visando garantir uma implementação com ótima performance
- Assegurando aspectos como padronização, portabilidade, disponibilidade e capacidade de recuperação tempestiva dos dados.
- Cada sistema tem as suas próprias particularidades.



Discos e Arquivos

- Um SGBD guarda informação em discos.
- Este fato tem grandes implicações no projeto de um SGBD
- **READ**: transferência de dados do disco para a memória principal (RAM).
- **WRITE**: transferência de dados da RAM para o disco.
- Ambas são operações de custo elevado em termos de tempo e espaço em memória, de modo que devem ser planejadas cuidadosamente!



Meios Físicos de Armazenamento

Aspectos a Serem Considerados

- Velocidade com a qual um dado pode ser acessado
- Custo para ler e armazenar cada unidade de dado
- Disponibilidade
- Perda de dados em caso de falha no sistema
- Falhas físicas nos dispositivos de armazenamento

Visão Geral dos Meios Físicos de Armazenamento

1. Voláteis

- Memória RAM (Principal)
- Memória Cache

2. Não-voláteis

1. Memória secundária

- Disco magnético
- Disco ótico
- Memória flash

2. Memória terceária

- Fita

Por que não armazenar tudo em memória principal?

- Custo muito alto
- Memória volátil

Hierarquia de Armazenamento

1. Armazenamento **Primário**:

- Memória Principal + Caches de Memória:
- acessado diretamente pela CPU, acesso rápido, custo alto

2. Armazenamento **Secundário** ou On-Line:

- Discos magnéticos, óticos e memória flash
- armazena a base de dados em si, acesso lento
- dados são copiados nos meios de armazenamento primário para serem processados e depois reescritos novamente

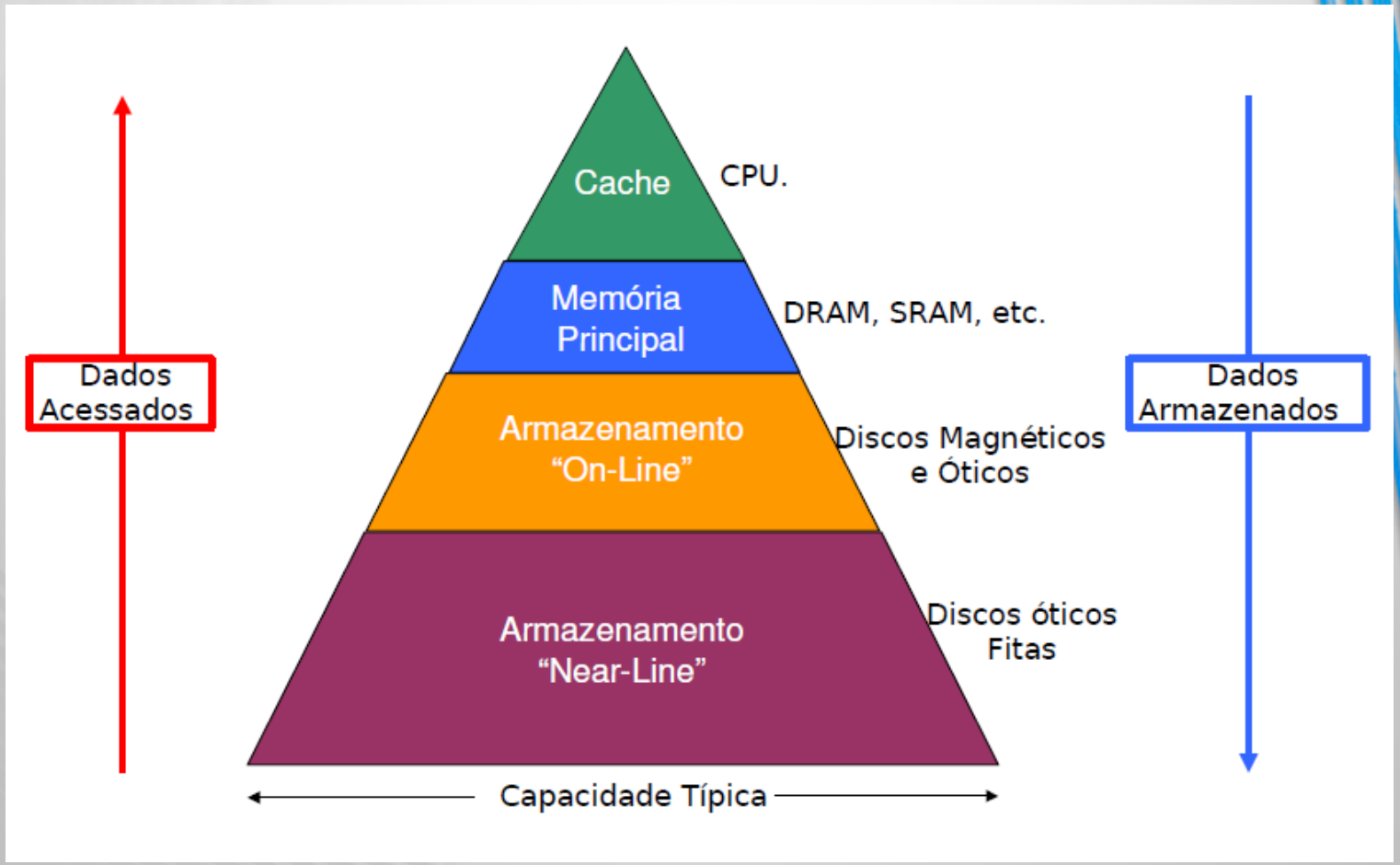
3. Armazenamento **Terceário** ou Off-Line:

- Fitas = para versões antigas da base de dados (ou backups)

Questão

Qual a diferença entre armazenamento primário e secundário?

Hierarquia de Armazenamento

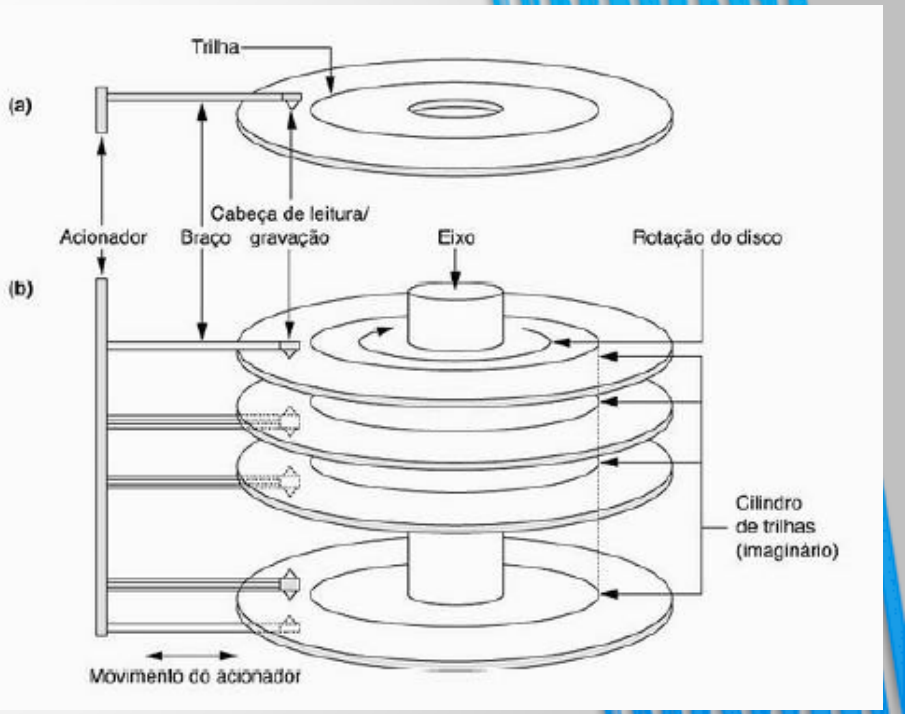


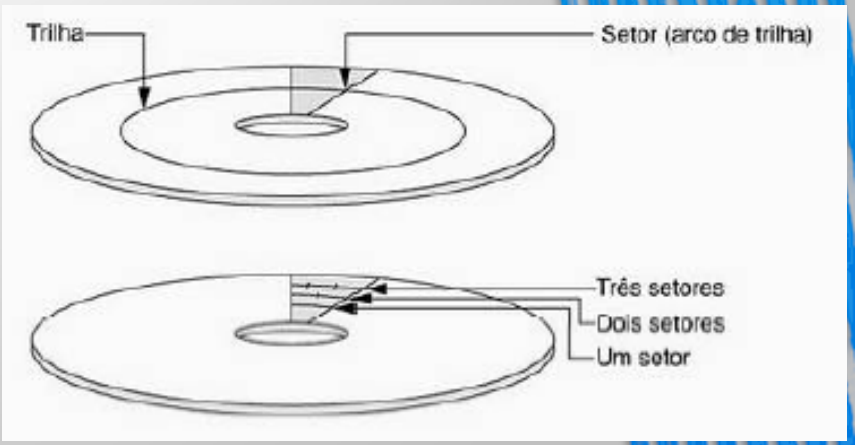
Discos Magnéticos

- Principal dispositivo de memória secundária utilizado.
- Principal vantagem com relação às fitas:
- **acesso aleatório (randômico) vs. sequencial.**
- Dados são armazenados e devolvidos em unidades chamadas blocos ou páginas.
- Ao contrário da RAM, o tempo para devolver um bloco de um disco varia com a sua localização em disco, o que tem grande impacto no desempenho de um SGBD.

Componentes de um disco

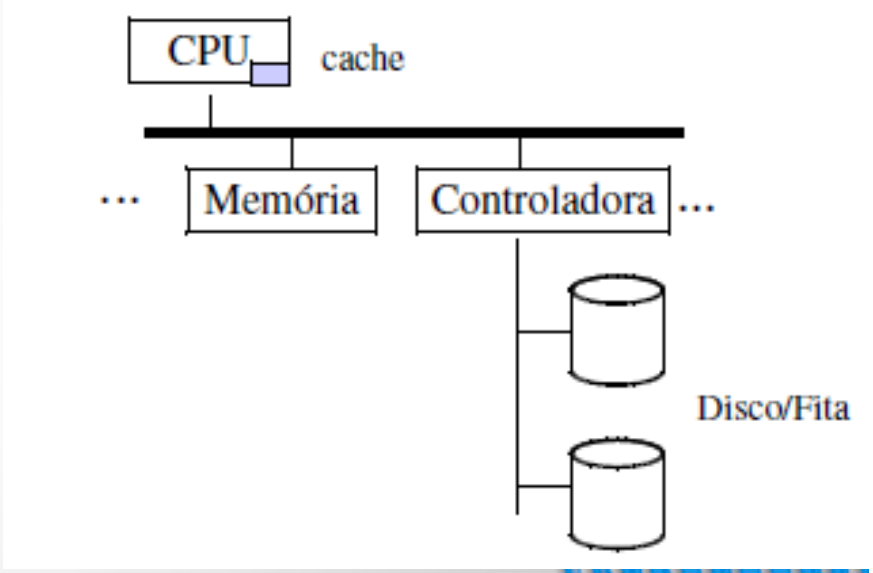
- Os blocos são armazenados em trilhas **trilhas** que formam cada prato (disco)
- 500 a 200 trilhas por superficie
- Cada trilha é dividida em vários **setores** (menor unidade de informação lida/escrita)
- 32 a 128 setores por trilha
- 1 setor = 32 até 4096 bytes
- Os **pratos** giram. Por exemplo,5400 rpm.
- A agulha move-se para dentro ou para fora de modo a posicionar a cabeça sobre a trilha desejada.
- As trilhas de mesmo diâmetro formam **<u>cilindro</u>** (imaginário!).
- *<u>Só uma cabeça lê/escreve em cada instante.</u>*





Subsistema de discos

- Um controlador de discos, comumente embutido na unidade de disco, controla o disco e o interliga aos sistema de computação, fazendo interface entre o disco e a memória RAM
- Recebe comandos de leitura/escrita de dados
- Remapeamento de setores ruins
- ATA, SATA, SCSI (mais usada para conectar discos a computadores pessoais e estações de trabalho)



Medidas de Desempenho de Disco

1. **Capacidade**

2. **Tempo de Acesso** – acesso de leitura ou escrita requer três passos:

- Tempo de procura (seek): posicionamento do braço na trilha correta. De 4 a 10 ms.
- Tempo de atraso (latência rotacional): espera até o setor desejado seja rotacionado até a cabeça de leitura/escrita. 5400 to 15000 r.p.m.
- Tempo de transferência: transferência dos bits armazenados no setor que está ao alcance da cabeça. 25 a 100 Mb por segundo
- Bloco: unidade de transferência

3. **Confiabilidade**

- Tempo médio para a ocorrência de falhas: média de tempo que se pode esperar que o disco trabalhe sem que ocorra falhas. De 3 a 5 anos.

Exemplo - Seagate

Especificações	4 TB[1]	3 TB[1]	2 TB[1]	1 TB[1]
Número do modelo	ST4000DM000	ST3000DM001	ST2000DM001	ST1000DM003
Nome do modelo	Desktop HDD	anteriormente Barracuda®	anteriormente Barracuda	anteriormente Barracuda
Opções de interface	SATA de 6 Gb/s com NCQ	SATA de 6 Gb/s com NCQ	SATA de 6 Gb/s com NCQ	SATA de 6 Gb/s com NCQ
Desempenho				
Cache, multisegmentado (MB)	64	64	64	64
Taxas de transferência aceitas por SATA (Gb/s)	6,0/3,0/1,5	6,0/3,0/1,5	6,0/3,0/1,5	6,0/3,0/1,5
Média de busca, leitura (ms)	<8,5	<8,5	<8,5	<8,5
Média de busca, gravação (ms)	<9,5	<9,5	<9,5	<9,5
Taxa média de dados, leitura/gravação (MB/s)	160	156	156	156
Taxa de dados sustentada máx., leitura DE (MB/s)	180	210	210	210
Configuração/organização				
Cabeças/discos	8/4	6/3	6/3 4/2	2/1
Bytes por setor	4.096	4.096	4.096	4.096

Exercício

1. Considere um disco com tamanho de setor igual a 512 bytes, 2000 trilhas por superfície, 50 setores por trilha, cinco pratos e tempo de busca médio de 10 ms.

a. Qual a capacidade de uma trilha em bytes? Qual a capacidade de cada superfície? Qual a capacidade do disco?

b. Dê exemplos de tamanhos válidos de blocos. 256 bytes? 2048 bytes? 51200 bytes?

c. Se os pratos do disco girarem a 5.400 rpm, qual a latência rotacional máxima?

d. Se uma trilha de dados puder ser transferida por rotação, qual a taxa de transferência?

e. Quantos cilindros o disco tem?

Otimização de Acesso de Blocos de Disco

- **Bloco** = é uma sequência contígua de bytes de uma única trilha de um prato
- Dados são transferidos do disco para a memória principal em blocos
- Os tamanhos dos blocos variam de 512 bytes a vários kb
- Blocos menores – mais transferências do disco
- Blocos maiores – mais espaço desperdiçado
- O tamanho mais comum varia de 4 a 16 kbytes

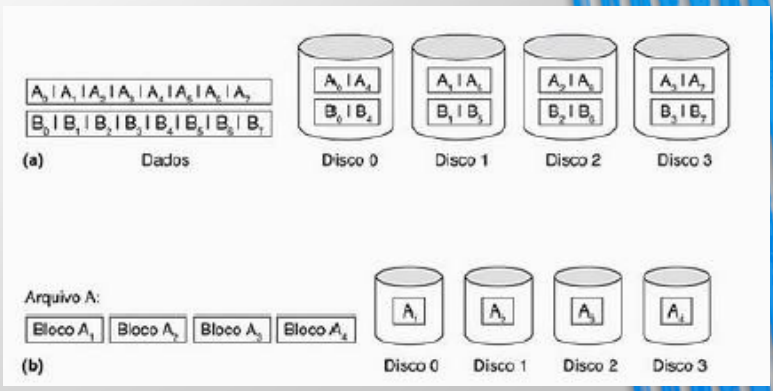
Melhoria na Confiabilidade e Desempenho

RAID (Arrays Redundantes de Discos Independentes)

- Array de pequenos discos independentes que atuam como único disco lógico de maior desempenho
- Conjunto de discos de dados + um conjunto de discos de verificação
- Duas técnicas principais:
  1. **Data striping**: particionamento de dados
  2. **Redundância**: informação redundante permite reconstrução de dados caso disco falhe

Striping de dados

- Emprega o paralelismo para melhorar o desempenho do disco
- O tamanho da partição é chamado unidade striping
- Partições de mesmo tamanho são distribuídas em vários discos.
- Para D discos a partição i é escrita no disco (i mod D)
- Permite leitura em paralelo
- Partição pode ser por **bit** ou **bloco**



Níveis de RAID - 0

- **Sem redundância**
- Este nível também é conhecido como "Striping" ou "Fracionamento".
- Os dados são divididos em pequenos segmentos e distribuídos entre os discos.
- Não oferece tolerância a falhas, pois não existe redundância.
- Isso significa que uma falha em qualquer um dos HDs pode ocasionar perda de informações.

Níveis de RAID - 1

- **Espelhamento**
- Consiste em espelhar os discos.
- A informação gravada num disco será gravada em dois discos componentes do array.
- Caso um deles falhe, o array continua funcionando.

Níveis de RAID - 2

- **Código de correção e detecção de erros**
- É direcionado para uso em discos que não possuem detecção de erro de fábrica.
- É muito pouco usado uma vez que os discos modernos já possuem de fábrica a detecção de erro embutida.

Níveis de RAID - 3

- **Paridade de Bits Interlaçados**

- os dados são divididos em bits entre os discos, exceto um, que armazena informações do bit de paridade.
- Assim, todos os bytes dos dados tem sua paridade (acréscimo de apenas 1 bit, que permite identificar erros) armazenada em um disco específico.
- Através da verificação desta informação, é possível assegurar a integridade dos dados, em casos de recuperação.
- Por isso e por permitir o uso de dados divididos entre vários discos, o RAID 3 consegue oferecer altas taxas de transferência e confiabilidade das informações.
- Para usar o RAID 3, pelo menos 3 discos são necessários.

Níveis de RAID - 4

- **Paridade de Blocos Interlaçados**
- O nível 4 divide os dados, a nível de "blocos", entre múltiplos discos.
- A paridade é gravada em um disco separado.
- O RAID 4 é indicado para o armazenamento de arquivos grandes, onde é necessário assegurar a integridade das informações.
- Isso porque, neste nível, cada operação de gravação requer um novo cálculo de paridade, dando maior confiabilidade ao armazenamento (apesar de isso tornar as gravações de dados mais lentas).

Níveis de RAID - 5

- **Paridade Distribuída de Blocos Interlaçados**
- Comparável ao nível 4 mas ao invés de gravar a paridade em um disco separado, ela é distribuída entre os discos disponíveis.
- Para cada bloco, um dos discos armazena a paridade e os outros armazenam os dados.
- O nível 5 aumenta a velocidade em gravações pequenas, uma vez que não há um disco separado de paridade como gargalo.
- Porém como o dado de paridade tem que ser distribuído entre todos os discos disponíveis, durante a leitura, a performance possui tendência de ser um pouco mais lenta que a do nível 4.
- É o tipo mais comum de RAID

Resumindo

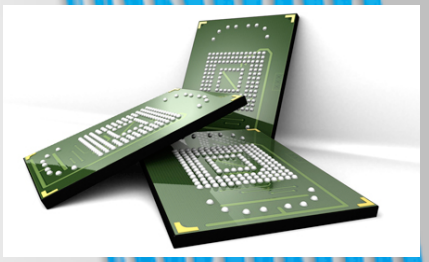
Níveis de RAID			
Nível	Redundância	Stripping	Comentários
0	Sem	Sim	melhora write
1	espelhamento	Não	melhora segurança e leitura otimizada
2	Detector de erros	Unidade = bit	Substituído pelo 3
3	Paridade	Unidade = bit	identifica disco que falhou
4	Paridade	Unidade = bloco	Substituído pelo 5
5	Paridade distribuída	Unidade = bloco	Elimina gargalo

Para Pensar

Considere o seguinte arranjo de quatro discos de blocos de dados e de paridade em que Bis representam blocos de dados e Pis representam os blocos de paridade. O bloco de paridade Pi é o bloco de paridade para os blocos de dados de B 4i-3 até B4i. Qual o problema (se houver) que esse arranjo pode representar?

Disco 1	Disco 2	Disco 3	Disco 4
B_1	B_2	B_3	B_4
P_1	B_5	B_6	B_7
B_8	P_2	B_9	B_{10}
.	.	.	.
.	.	.	.
.	.	.	.

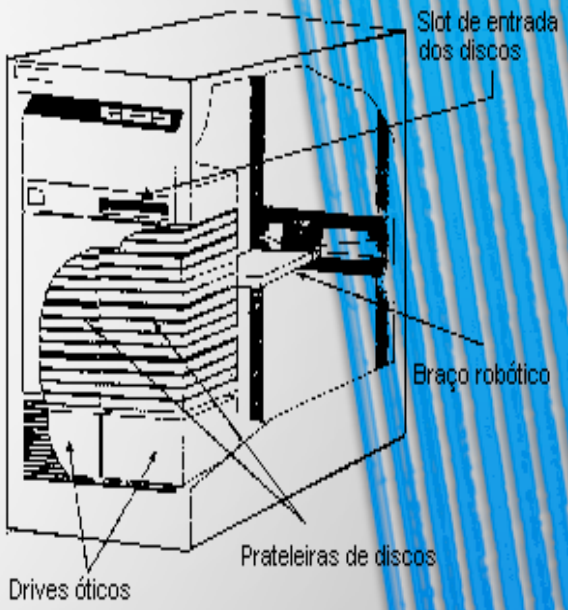
Memória Flash



- Opção de armazenamento secundário
    1. não volátil
    2. acesso rápido à memória RAM
    3. usada em dispositivos USB, câmeras, celulares, laptops
- Está "substituindo" os discos HD nos armazenamentos de dados, porém o custo ainda é bem maior (3 a 4 vezes)

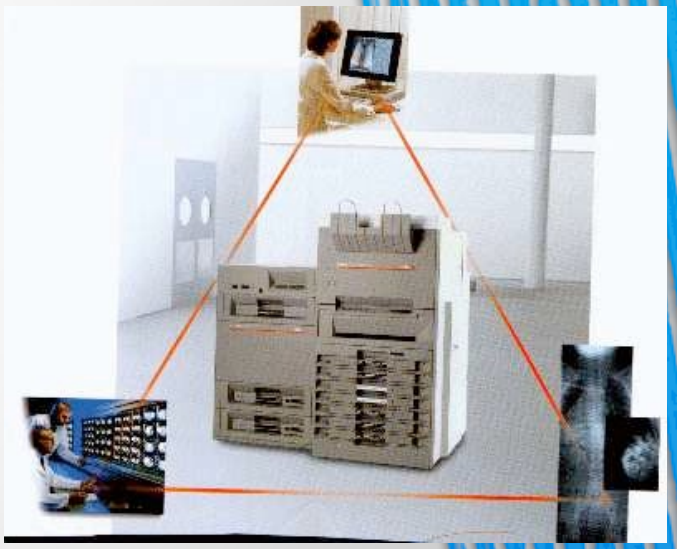
Discos Óticos

- Os dados são armazenados oticamente nos discos e lidos a laser (CDs e DVDs)
- Os discos podem ser carregados e removidos facilmente do acionador
- Grande capacidade de armazenamento, custo baixo
- Sistemas junkebox = contém várias unidades de discos que podem ser trocadas automaticamente por meios de braços mecânicos.



Fitas Magnéticas

- Usadas primordialmente para para backups
- Acesso muito lento (sequencial)
- Barato e de fácil armazenamento
- Junkeboxes de fitas = mantém um grande números de fitas (na casa das centenas) com troca automática entre elas.



Questão

Por que os discos e não as fitas são usados para armazenar arquivos de banco de dados on-line?

Novos Sistemas de Armazenamento

1. SAN - Storage Area Network (Área de Armazenamento em Rede)

- Periféricos de armazenamento online são configurados como nós em uma rede de alta velocidade e podem ser conectados/desconectados dos servidores com flexibilidade

2. NAS - Network-Attached Storage (Armazenamento Conectado à rede)

- São servidores que permitem o acréscimo de armazenamento para o compartilhamento de arquivos

3. iSCSI - Internet SCSI

- Novo protocolo de rede que permite que os clientes enviem comandos para dispositivos de armazenamento SCSI em canais remotos