centro universitário FEI

AULA 1 - ARMAZENAMENTO E ARQUIVOS

PROFA. DRA. LEILA BERGAMASCO

CC6240 – Tópicos Avançados em Banco de Dados

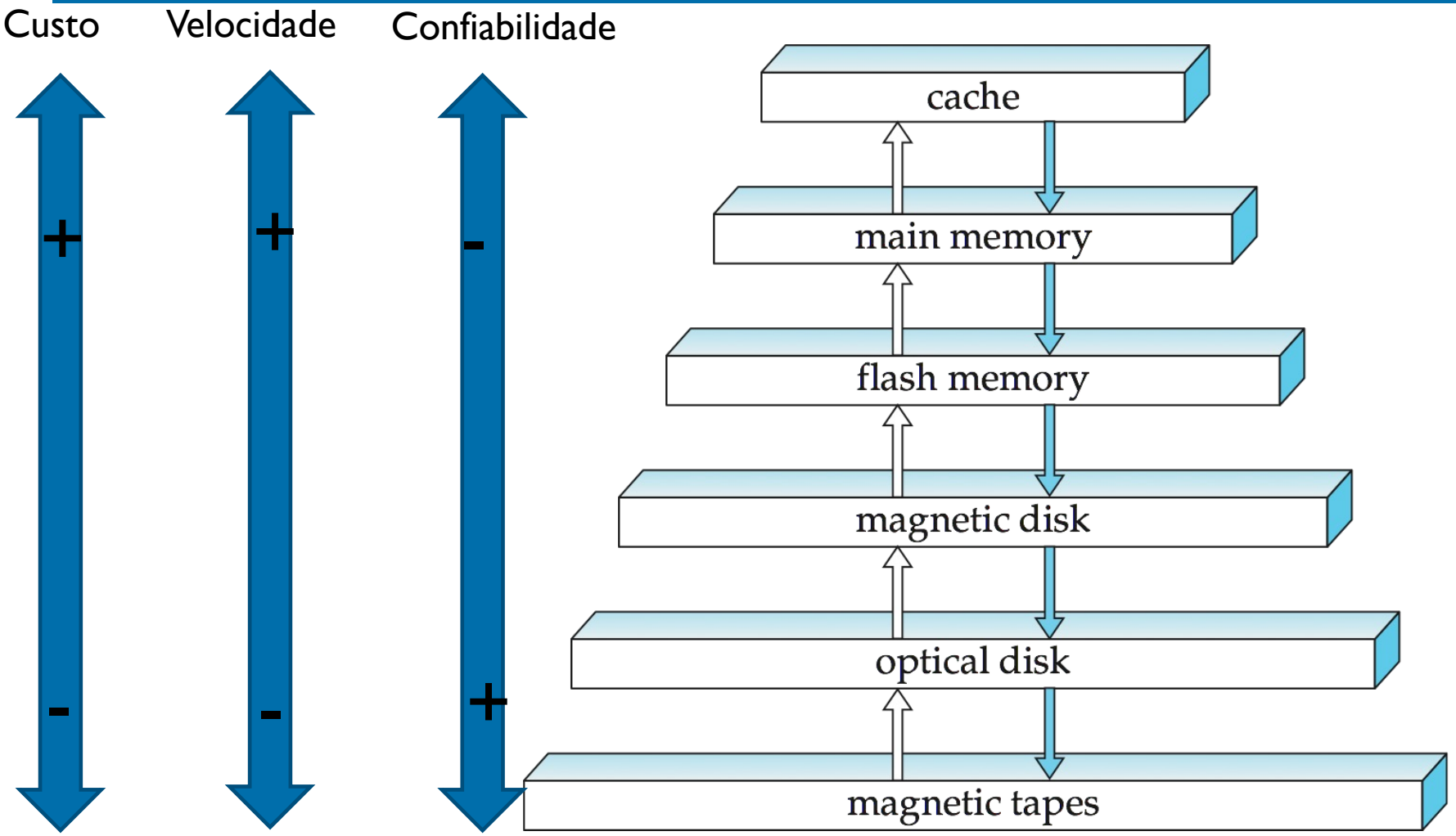
centro universitário FEI

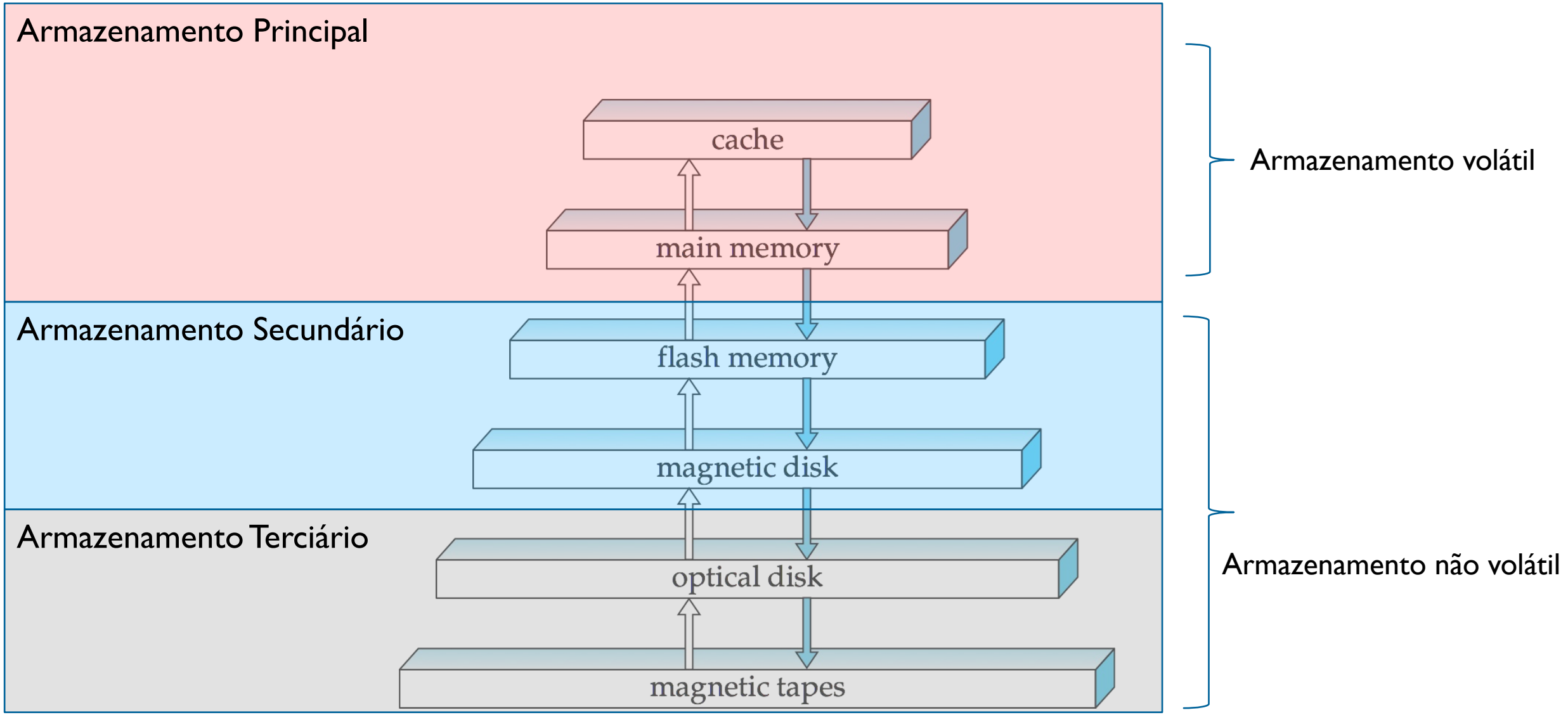
AULA 1: ARMAZENAMENTO E ESTRUTURAS DE ARQUIVOS

ARMAZENAMENTO FÍSICO

- Classificados de acordo com custo, velocidade e confiabilidade
  - Cache
  - Memória principal
  - Memória flash
  - Disco magnético
  - Disco óptico
  - Fitas magnéticas

ARMAZENAMENTO FÍSICO

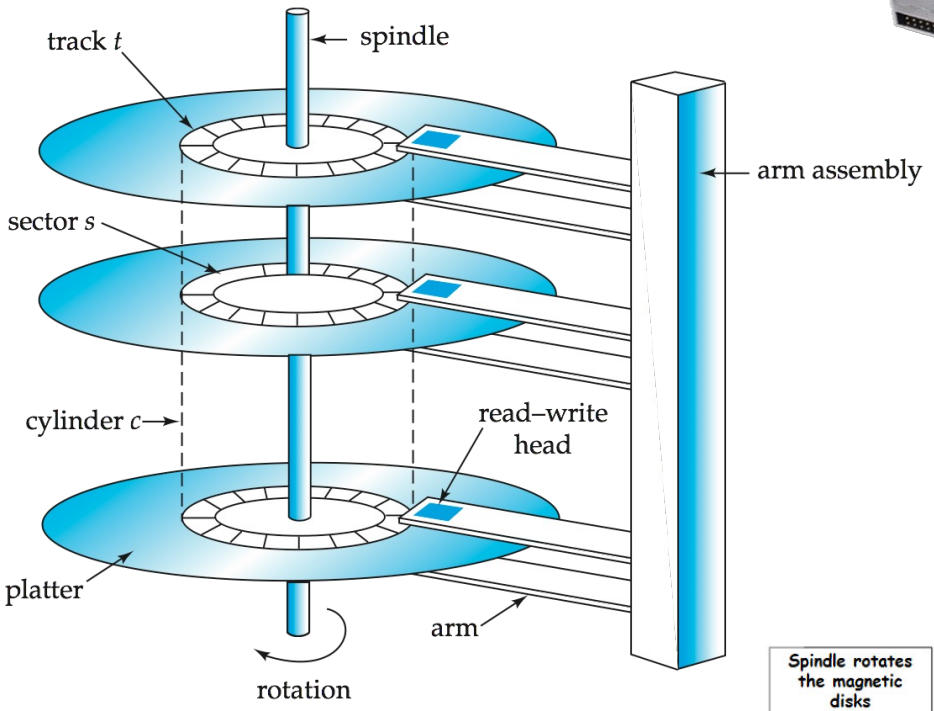



Armazenamento Principal
cache
main memory
Armazenamento Secundário
flash memory
magnetic disk
Armazenamento Terciário
optical disk
magnetic tapes
Armazenamento volátil
Armazenamento não volátil

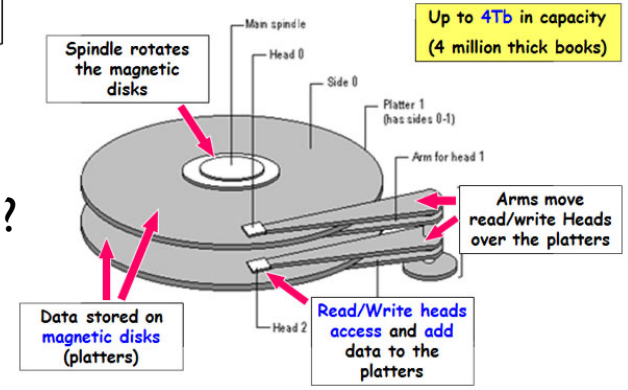
youtube.com/watch?v=lpYfep68xnA

DISCOS MAGNÉTICOS

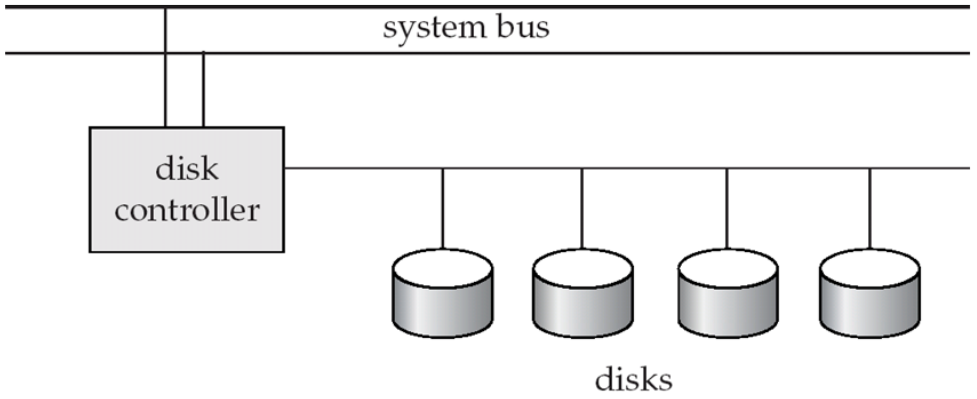
- Placa
  - Forma circular plana, dados são gravados na sua superfície
  - 2-3 placa, com duas faces.
  - Dividida em trilhas
- Trilha
  - 50-100 mil trilhas por placa que por sua vez são divididas em setores
- Setores
  - 500-1000 setores por trilha mais interna
  - 2000 setores por trilha (mais externa)
  - Menor unidade de dados → Cada setor possui 512 bytes
- Cilindro
  - Conjunto das mesmas trilhas pertencente a diferentes placas
  - A cabeça de leitura não funciona de forma independente.
- Cabeça leitura-escrita
  - "Flutua" muito próxima à placa
  - Função de leitura e escrita no disco
- Montagem do braço
  - Muitas placas para cada disco (1-5)
  - Uma cabeça leitura-escrita para cada placa e um braço para cada disco



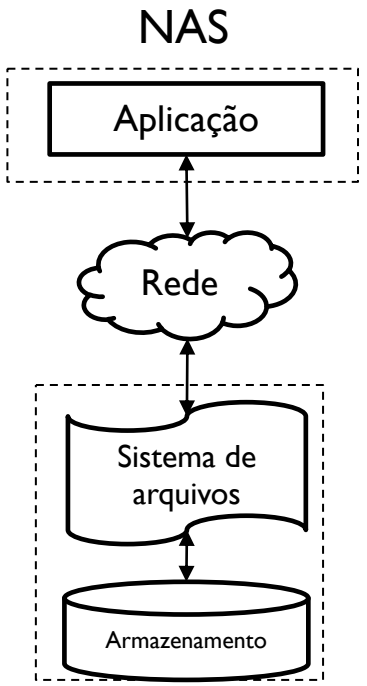
Por que HDs possuem 3.5'?

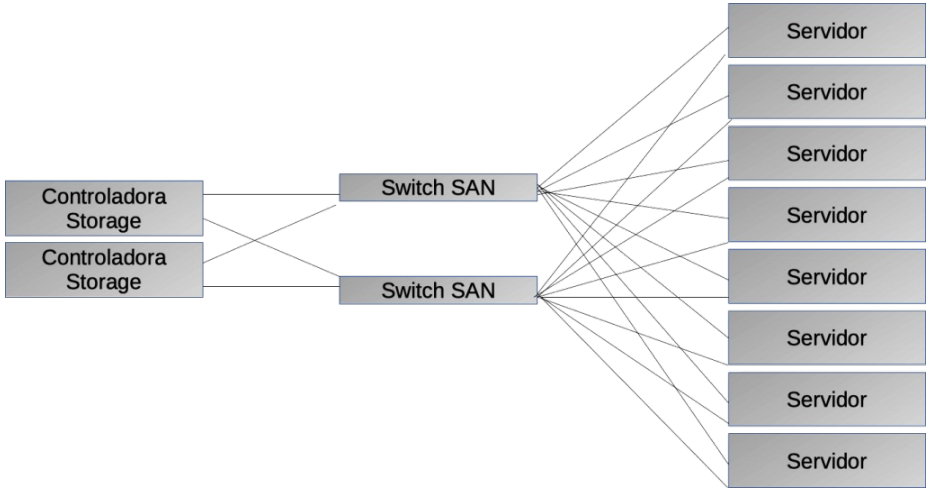


DISCOS MAGNÉTICOS



- Controladora de disco
  - Comunicação entre computador e hardware do disco
    - Aceita commandos ler/escrever e traduz em comandos para a cabeça dos discos
  - Checksum
  - Remapeamento
- Interfaces
  - ATA – Advanced Technology Attachment
  - SATA – Serial Ata
  - SCSI – Small Computer System Connect
  - SAS – Serial SCSI
- Arquiteturas
  - SAN - Storage Area Network
    - Discos conectados por uma rede de alta velocidade
    - Utilizam armazenamento RAID
    - Visão lógica de apenas um disco
  - NAS - *Network Attached Storage*
    - Interface de sistemas de arquivos:



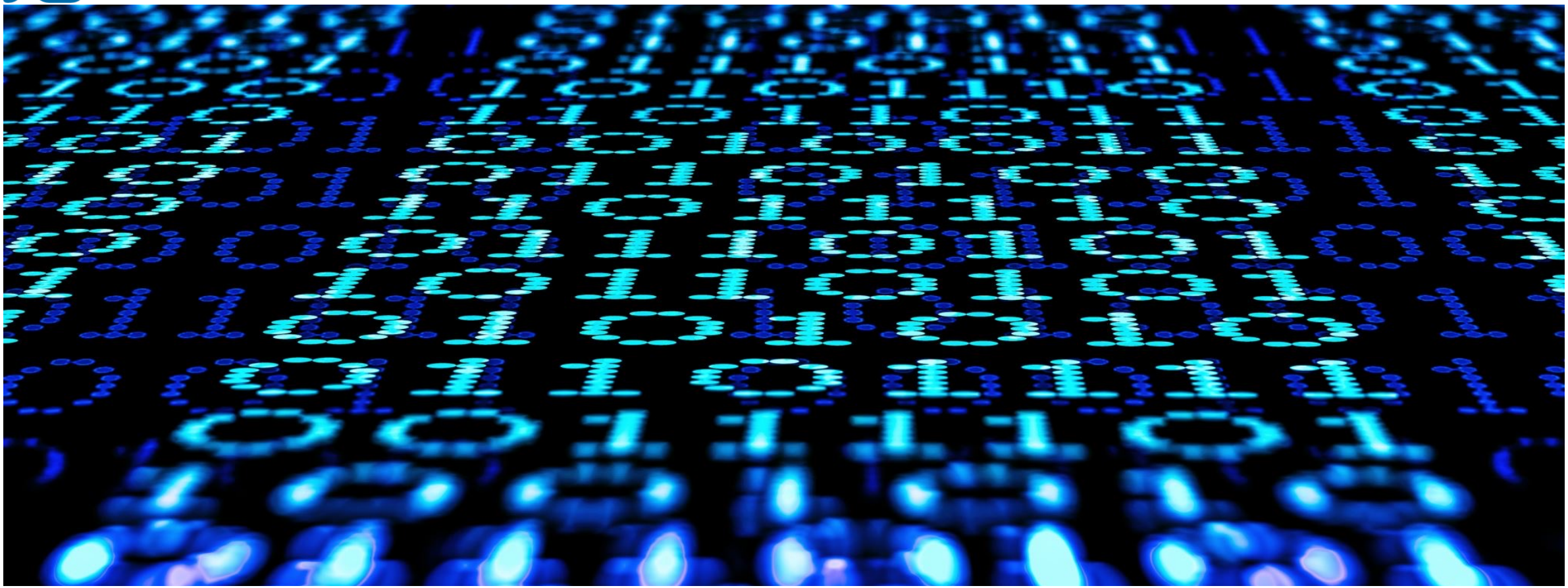


https://www.youtube.com/watch?v=uSb2XG1rmK4

MEDIDAS DE DESEMPENHO DOS DISCOS

- Capacidade
- Tempo de acesso (TA)
  - Tempo de busca médio (TB)
  - Tempo de latência rotacional médio (TL)
  - Tempo de transferência (TR)
  - TA =TB+TL + TR
- Taxa de transferência
- Confiabilidade
  - MTTF = Tempo Médio para falha (Mean time to failure)
    - Alguns fabricantes dizem que o MTTF é 1.200.000 horas = 136 anos!
    - Porém se não usássemos o disco. Horas de utilização diminuem o MTTF.
    - Atualmente um disco tem vida útil de 5 anos.

centro universitário FEI



RAID

centro universitário FEI

RAID

- RAID: *Redundant Arrays Independent Disks*.
  - Melhora a confiabilidade por redundância: espelhamento
  - Melhora o desempenho por paralelismo: espalhamento
- Espelhamento:
  - Um disco lógico consiste de dois ou mais discos físicos e cada escrita é feita nos dois discos. Se um dos discos falha, é possível ler os dados a partir do outro.
    - Mesmo assim se durante uma queda de energia o dado estiver sendo escrito nos discos, o dado é corrompido. Como proceder nesse caso?
- Espalhamento:
  - Espalhar dados por vários discos
    - Espalhamento no nível de bit
    - Espalhamento no nível de bloco (mais usado)

centro universitário FEI

RAID

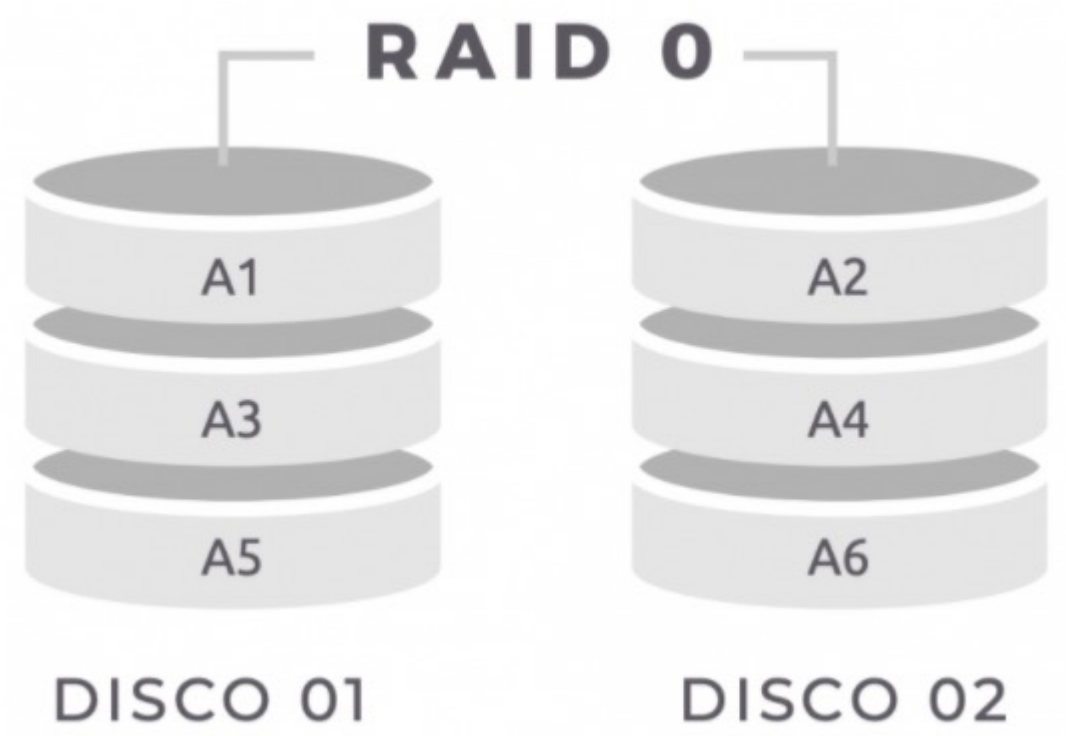
- Espelhamento: + confiabilidade + custo
- Espalhamento: - confiabilidade + taxa de transferência

Não existe uma solução ideal!

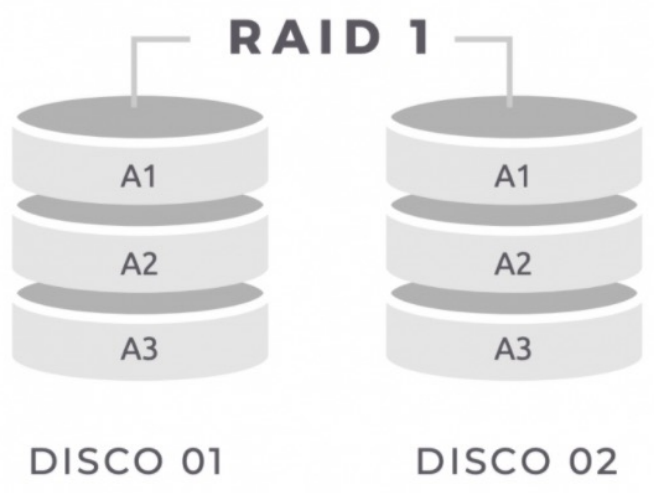
- Combinação das duas soluções geram diferentes esquemas
  - Níveis de RAID

RAID

- RAID nível 0: apenas espalhamento a ní
  - Alta resposta, baixa confiabilidade. Exemplo

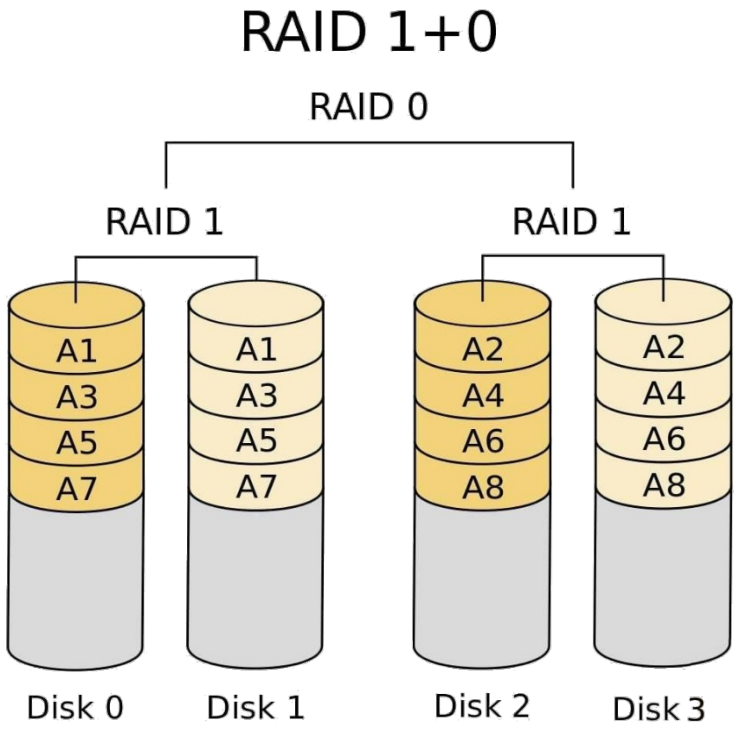


RAID



- RAID nível 1: espelhamento a nível de bloco
  - Melhor custo de escrita, mais tolerante a falhas. Muito usado para armazenamento de logs

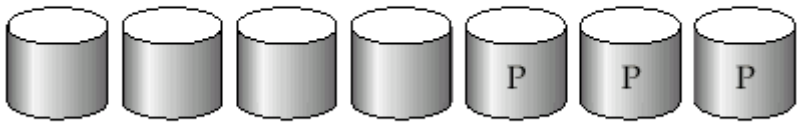
- RAID nível 1+0: espelhamento com espalhamento
  - Mais tolerante a falhas



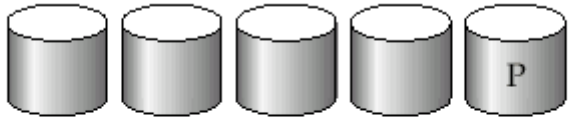
RAID

- RAID nível 2: ECC (Error-Correcting-Code) – código de correção de erro no estilo da memória
  - Checagem de erros. HDs atuais já trazem nativamente o checksum.

- RAID nível 3: um [illegible] tado para cada palavra de dados e escrit



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

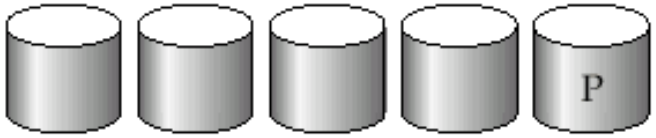


(d) RAID 3: bit-interleaved parity

https://www.youtube.com/watch?v=xPYYKFR3-4A

RAID

- RAID nível 4:
  - Similar ao RAID 3 porém utiliza-se a distribuição e paridade por bloco. Paralelismo maior, bom para arquivos grandes.
- RAID nível 5:
  - Espalhamento no nível de bloco porém cada disco já possui sua respectiva porção de paridade. Logo o paralelismo aumenta ainda mais.



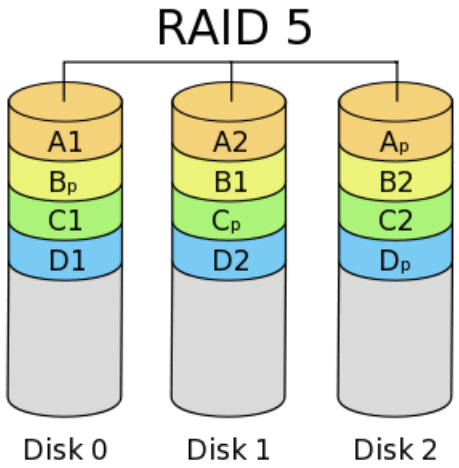
(e) RAID 4: block-interleaved parity

(f) RAID 5: block-interleaved distributed parity

https://www.youtube.com/watch?v=xPYYKFR3-4A

RAID

- RAID nível 5:
  - Espalhamento no nível de bloco porém cada disco já possui sua respectiva porção de paridade. Logo o paralelismo aumenta ainda mais.



Discos lógicos com 3 discos

Distribuição $Pi = B_{2i-1} \rightarrow B_{2i}$

Pi = 1 → Armazena Discos 1 até 2

2 → Armazena Discos 3 até 4

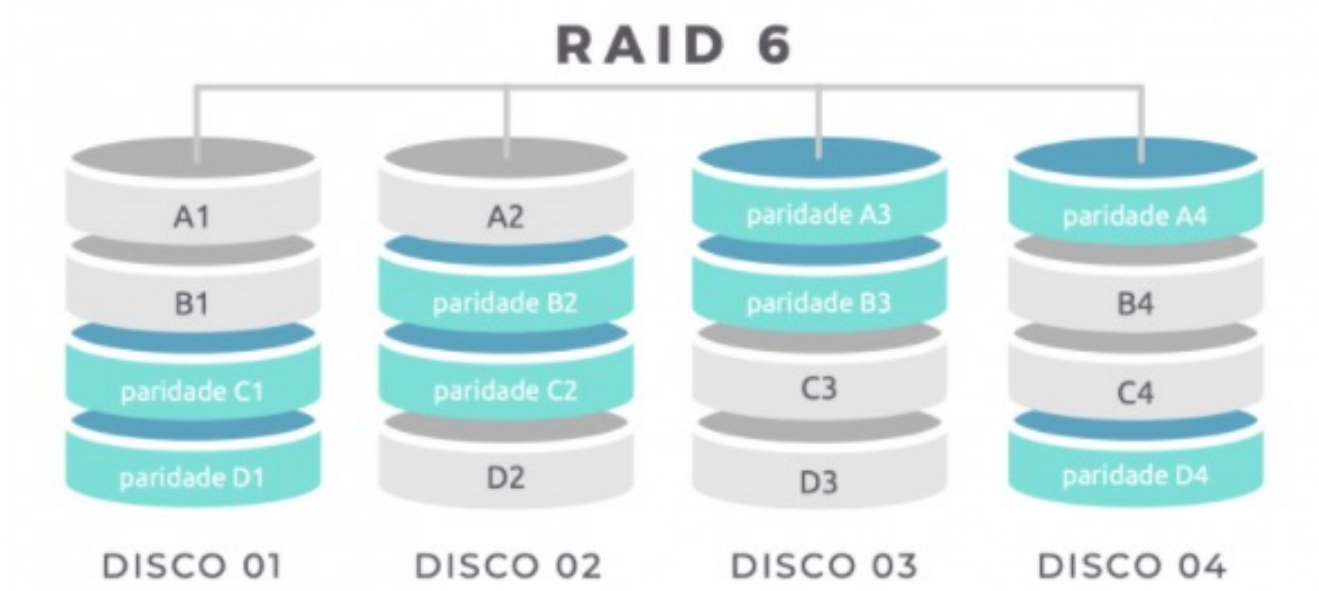
3 →???

DL 1	DL2	DL3
$Ap_{(1,2)}$	A1	A2
A3	$Ap_{(3,4)}$	A4
A5	A6	$Ap_{(5,6)}$

https://www.youtube.com/watch?v=xPYYKFR3-4A

RAID

- RAID nível 6:
  - Dobra a quantidade de discos de paridade. "Aguenta" falha em dois discos simultaneamente.



DL 1	DL2	DL3	DL4
1	2	$P_{(1-2)}$	$P_{(1-2)}$
3	$P3_{(3-4)}$	$P4_{(3-4)}$	4
$P5_{(5-6)}$	$P6_{(5-6)}$	5	6
$P8_{(7-8)}$	7	8	$P7_{(7-8)}$

CONFIGURAÇÃO DE UM RAID NO STORAGE DELL

- https://www.youtube.com/watch?v=11NZKtkW54U
- https://www.delltechnologies.com/en-us/servers/specialty-servers/poweredge-xe-servers.htm#video-overlay=6179217788001

centro universitário FEI

RAID – QUAL ESCOLHER?

- Fatores para escolha
  - Custo
  - Performance
  - Performance durante a falha
  - Performance durante a recuperação
- RAID 0 segurança do dado não é importante. Recuperação rápida
- RAID 2 e 4 Não são usados já que o RAID 3 e RAID 5 superaram suas características
- RAID 3 é baseada em busca a nivel de bit. Muito dispendisioso e não é utilizada hoje em dia. Nível 5, baseado em blocos é mais usado.
- Level 6 é pouco usado já que usar uma solução 5 ou 1 oferecem o mesmo nível se segurança.

centro universitário FEI

RAID – QUAL ESCOLHER?

- RAID 1 tem performance melhor que RAID 5
- RAID 1 tem custo maior de armazenamento.
- Porém: Enquanto capacidade de disco aumenta 50%/ano, acesso ao disco diminui 3x a cada 10 anos!
- Logo o custo maior diminui ano a ano.

- RAID 5 é preferível para aplicações com baixa taxa de acesso porém grande quantidade de dados
- RAID 1 para o restante.

centro universitário FEI

BIBLIOGRAFIA

- ABRAHAM SILBERSCHATZ, HENRY F. KORTH, S. SUDARSHAN. Sistema de Banco de Dados. 6. Campus. 0. ISBN 9788535245356.

- ELMASRI, RAMEZ, SHAMKANT B. NAVATHE. Sistemas de banco de dados. Vol. 6. São Paulo: Pearson Addison Wesley, 2011.

- DATE, CHRISTHOPER J. Introdução a Sistemas de Bancos de Dados, 5ª. Edição. Campus, Rio de Janeiro (2004).

centro universitário FEI

ARQUIVOS

centro universitário FEI

TAMANHOS DE REGISTROS

ORGANIZAÇÃO DE ARQUIVOS

Um banco de dados pode ser considerado uma coleção de tabelas → Cada tabela pode ser considerada um arquivo → Cada arquivo é sequência de registros (linhas da tabela) → Registro é uma sequência de campos

- Como organizar um banco de dados dentro de um disco?
- Lembrando que blocos tem tamanhos fixos
- Estratégia mais fácil de implementar
  - Registros com tamanho fixo
  - Registros do mesmo tipo

Um bloco pode armazenar uma sequência de registros

REGISTROS DE TAMANHO FIXO

```
create table disciplina (
    matricula number (5),
    nome char (22),
    nome_disciplina (22),
    num_disciplina number (5) )
```

Se char = 1 byte e number (5) = 8 bytes.
Cada registro da tabela terá 60 bytes

	8	+ 22	+ 22	+ 8
record 0	10101	Srinivasan	Comp. Sci.	65000
record 1	12121	Wu	Finance	90000
record 2	15151	Mozart	Music	40000
record 3	22222	Einstein	Physics	95000
record 4	32343	El Said	History	60000
record 5	33456	Gold	Physics	87000
record 6	45565	Katz	Comp. Sci.	75000
record 7	58583	Califieri	History	62000
record 8	76543	Singh	Finance	80000
record 9	76766	Crick	Biology	72000
record 10	83821	Brandt	Comp. Sci.	92000
record 11	98345	Kim	Elec. Eng.	80000

- Desafios:
  - Exclusão
  - Limite de bloco

REGISTROS DE TAMANHO FIXO

record 0	10101	Srinivasan	Comp. Sci.	65000
record 1	12121	Wu	Finance	90000
record 2	15151	Mozart	Music	40000
record 3	22222	Einstein	Physics	95000
record 4	32343	El Said	History	60000
record 5	33456	Gold	Physics	87000
record 6	45565	Katz	Comp. Sci.	75000
record 7	58583	Califieri	History	62000
record 8	76543	Singh	Finance	80000
record 9	76766	Crick	Biology	72000
record 10	83821	Brandt	Comp. Sci.	92000
record 11	98345	Kim	Elec. Eng.	80000

REGISTROS DE TAMANHO FIXO

Exclusão registro 3 e "subir" registros subsequentes

record 0	10101	Srinivasan	Comp. Sci.	65000
record 1	12121	Wu	Finance	90000
record 2	15151	Mozart	Music	40000
record 4	32343	El Said	History	60000
record 5	33456	Gold	Physics	87000
record 6	45565	Katz	Comp. Sci.	75000
record 7	58583	Califieri	History	62000
record 8	76543	Singh	Finance	80000
record 9	76766	Crick	Biology	72000
record 10	83821	Brandt	Comp. Sci.	92000
record 11	98345	Kim	Elec. Eng.	80000

Exclusão registro 3 e mover último registro

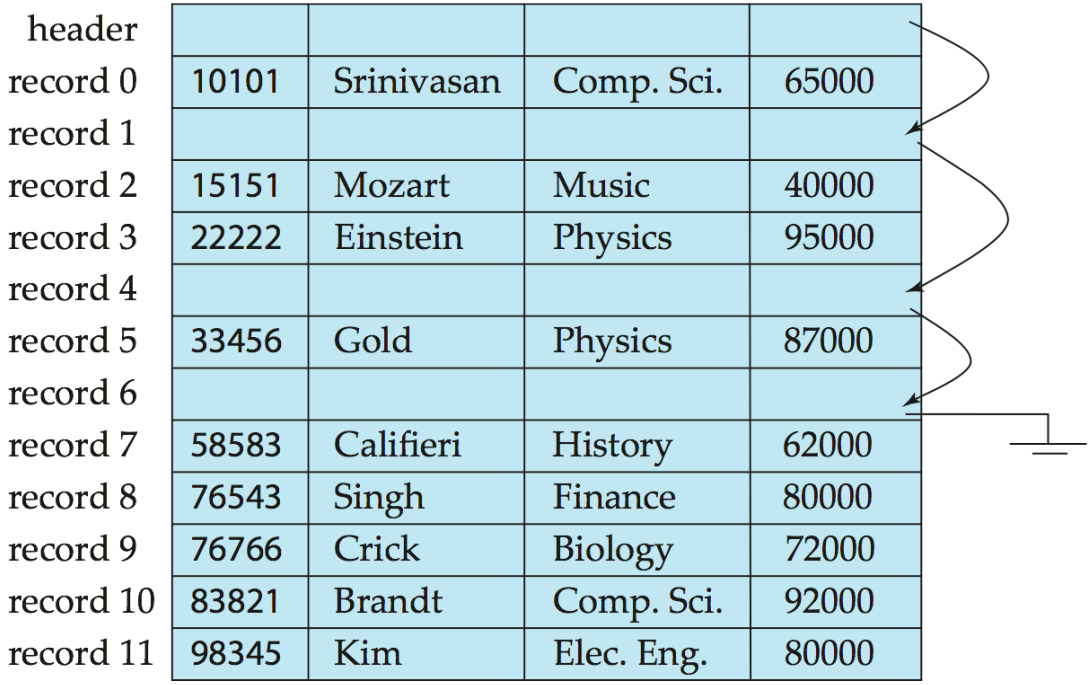
record 0	10101	Srinivasan	Comp. Sci.	65000
record 1	12121	Wu	Finance	90000
record 2	15151	Mozart	Music	40000
record 11	98345	Kim	Elec. Eng.	80000
record 4	32343	El Said	History	60000
record 5	33456	Gold	Physics	87000
record 6	45565	Katz	Comp. Sci.	75000
record 7	58583	Califieri	History	62000
record 8	76543	Singh	Finance	80000
record 9	76766	Crick	Biology	72000
record 10	83821	Brandt	Comp. Sci.	92000

3ª opção: Excluir e esperar nova inserção, porém "varrer" todos os registros atrás de espaço livre é custoso

Soluções pouco eficientes e custosas! → em todas elas há acessos a blocos = ++tempo

REGISTROS DE TAMANHO FIXO

- Possível solução:
  - Cabeçalho com ponteiro apontando próximo registro livre.
  - Dentro do registro excluído, reutilizar o espaço com ponteiro para o próximo registro excluído ou se não tiver, armazenar um marcador de fim e inserir como último registro

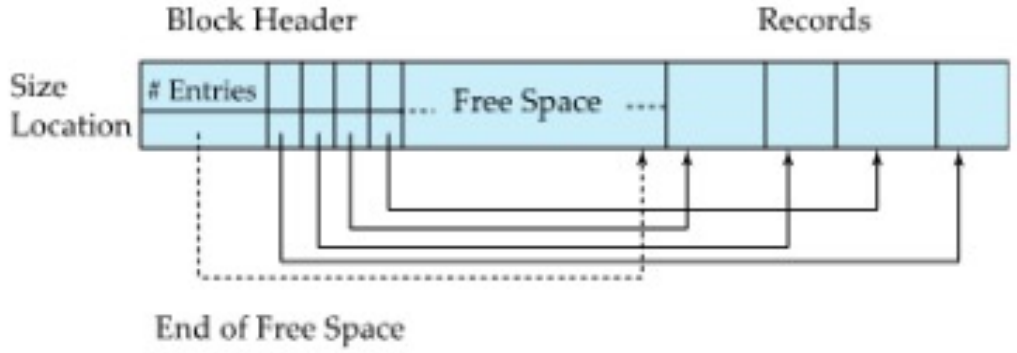


header				
record 0	10101	Srinivasan	Comp. Sci.	65000
record 1				
record 2	15151	Mozart	Music	40000
record 3	22222	Einstein	Physics	95000
record 4				
record 5	33456	Gold	Physics	87000
record 6				
record 7	58583	Califieri	History	62000
record 8	76543	Singh	Finance	80000
record 9	76766	Crick	Biology	72000
record 10	83821	Brandt	Comp. Sci.	92000
record 11	98345	Kim	Elec. Eng.	80000

Listas interligadas/livres

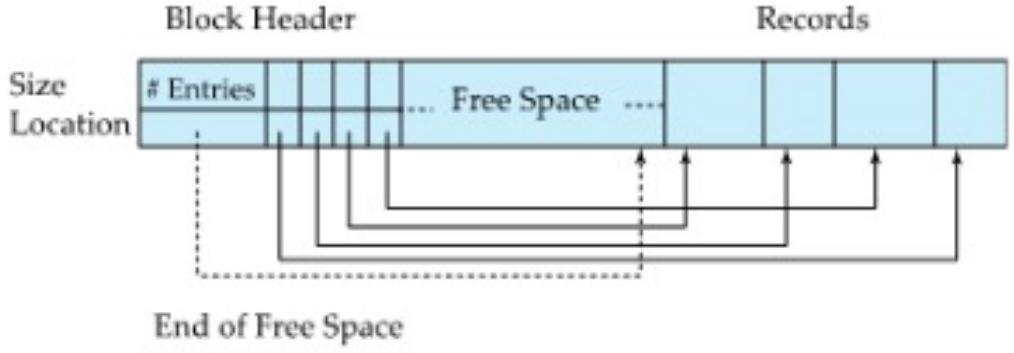
REGISTROS DE TAMANHO VARIÁVEL

- Armazenamento de vários tipos de registro em um arquivo
  - Tamanhos variáveis de campos
  - Campos repetidos (arrays, por exemplo →atributos multivalorados)
- Solução:
  - Estruturas de páginas em slot
  - cabeçalho no início de cada bloco com: número de entradas de registro; final do espaço livre, array com local e tamanho de cada registro



REGISTROS DE TAMANHO VARIÁVEL

- Os registros começam a ser alocados no fim do bloco
  - Não necessita de ponteiros em cada registro
  - Apenas aponta para o inicio dos registros
  - Registros próximos se movem, objetivando menos fragmentação
- BD relacionais



centro universitário FEI

ORGANIZAÇÃO DE REGISTROS

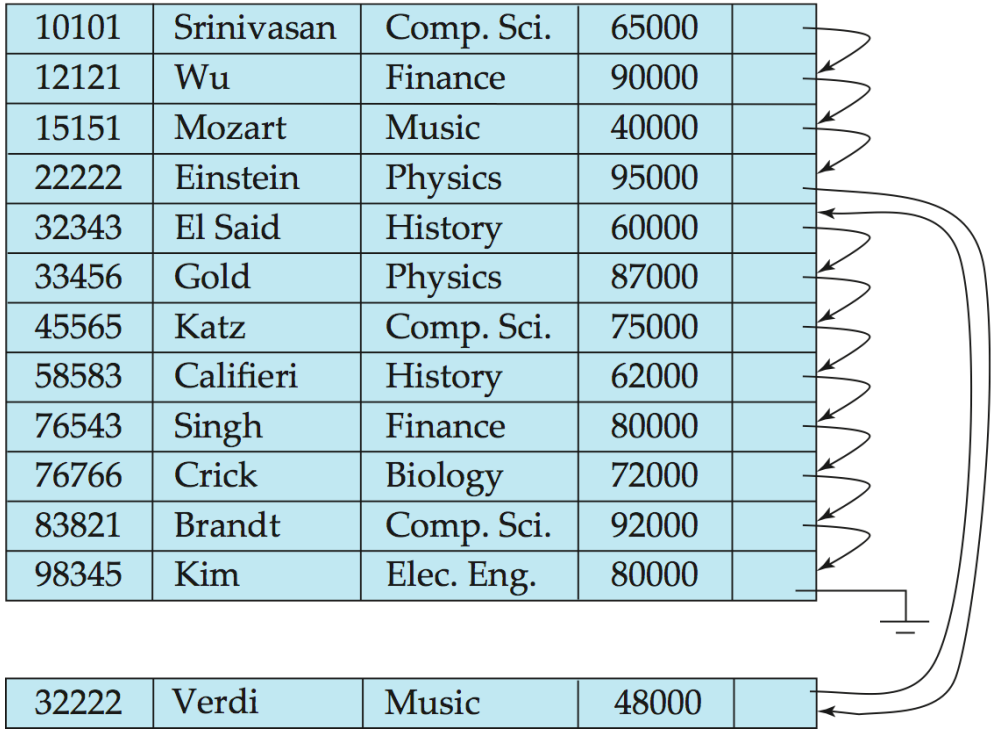
centro universitário FEI

ORGANIZAÇÃO DE REGISTROS EM ARQUIVOS

- Como os registros são organizados dentro dos arquivos?
  - Sequencial
  - Heap
  - Hashing
- Sequencial
  - Projetada para processamento eficiente de registros em ordem, dada alguma chave de busca (qualquer atributo ou conjunto de atributos)
  - Registros encadeados por ponteiros.
  - Registros fisicamente armazenados em ordem de chave de busca
- Problemas:
  - Manter ordem física após muitas inserções e exclusões.
  - Geralmente inserções feitas em bloco de estouro.

ORGANIZAÇÃO DE REGISTROS EM ARQUIVOS - SEQUENCIAL

- Se houver espaço livre, insira no espaço.
- Se não houver espaço livre, adicione no bloco de estouro
- Em qualquer caso, atualize os ponteiros



10101	Srinivasan	Comp. Sci.	65000	
12121	Wu	Finance	90000	
15151	Mozart	Music	40000	
22222	Einstein	Physics	95000	
32343	El Said	History	60000	
33456	Gold	Physics	87000	
45565	Katz	Comp. Sci.	75000	
58583	Califieri	History	62000	
76543	Singh	Finance	80000	
76766	Crick	Biology	72000	
83821	Brandt	Comp. Sci.	92000	
98345	Kim	Elec. Eng.	80000	

32222	Verdi	Music	48000	

centro universitário FEI

E O QUE ISSO IMPACTA NAS CONSULTAS SQL?

- Normalmente os BD armazenam cada relação (tabela) em um arquivo.
- Cada tupla (registro) possui tamanho fixo

- Funciona bem para BDs pequenos, médios
- BDs em grande escala: um arquivo único, grande é dedicado ao SGBD
- O SGBD armazena todas as relações neste arquivo e gerencia o próprio arquivo.

E O QUE ISSO IMPACTA NAS CONSULTAS SQL?

```
SELECT dept_name, building, budget, id, name, salary
FROM department, instrutor
WHERE department.dept_name = instrutor.dept_name
```

department

dept_name	*building*	*budget*
Comp. Sci.	Taylor	100000
Physics	Watson	70000

instructor

ID	*name*	*dept_name*	*salary*
10101	Srinivasan	Comp. Sci.	65000
33456	Gold	Physics	87000
45565	Katz	Comp. Sci.	75000
83821	Brandt	Comp. Sci.	92000

Comp. Sci.	Taylor	100000	
10101	Srinivasan	Comp. Sci.	65000
45565	Katz	Comp. Sci.	75000
83821	Brandt	Comp. Sci.	92000
Physics	Watson	70000	
33456	Gold	Physics	87000

Trazer dados do disco para RAM:

1. Traz o bloco de cada relação → registro próximos
2. Aplica-se where em um bloco só → uma leitura
3. Ponteiros entre os registros

PARTICIONAMENTO

- E quando uma relação é muito grande? Extrapolando um bloco?
- Particionamento de tabela: os registros em uma relação podem ser particionados em relações menores que são armazenadas separadamente
  - Por exemplo, a relação "transação" pode ser particionada em transaction_2018, transaction_2019, etc.
- As consultas devem acessar os registros em todas as partições, a menos que a consulta tenha uma seleção como ano = 2019, caso em que apenas uma partição é necessária
- Particionamento reduz os custos de algumas operações, como gerenciamento de espaço livre e permite que diferentes partições sejam armazenadas em diferentes dispositivos de armazenamento
- Por exemplo, partição de transação para o ano atual no disco, para anos mais antigos no disco magnético

centro universitário FEI

ORGANIZAÇÃO DE REGISTROS EM ARQUIVOS

- Organização em Heap:
  - Qualquer registro pode ser colocado em qualquer lugar onde existe espaço.
  - Não existe ordenação de registro.
  - Normalmente um único arquivo para cada relação.
  - Busca linear para encontrar registro procurado: problemática

- Organização em Hashing:
  - Função de hash é calculada sobre algum atributo.
  - Resultado da função especifica o bloco do arquivo em que registro será colocado.
  - Veremos mais sobre isso na próxima aula.

centro universitário FEI

BIBLIOGRAFIA

- ABRAHAM SILBERSCHATZ, HENRY F. KORTH, S. SUDARSHAN. Sistema de Banco de Dados. 6. Campus. 0. ISBN 9788535245356.

- ELMASRI, RAMEZ, SHAMKANT B. NAVATHE. Sistemas de banco de dados. Vol. 6. São Paulo: Pearson Addison Wesley, 2011.

- DATE, CHRISTHOPER J. Introdução a Sistemas de Bancos de Dados, 5ª. Edição. Campus, Rio de Janeiro (2004).

centro universitário FEI

OBRIGADO E ATÉ A PRÓXIMA AULA!