Disciplina: Arquitetura de Computadores

Tema: Memória e hierarquia de

Memórias

Prof. Plínio Andrade Passos

Características de memórias

- Localização
- Capacidade
- Método de acesso
- Desempenho
- Tecnologia
- Características físicas
- Organização



Requisitos ideais de uma memória

- Tamanho ilimitado

Acesso instantâneo para escrita ou leitura de informações

"Teoria da Contradição das memórias"

Quanto maior a memória maior será o seu tempo de acesso



Exemplo de Estudante em uma Biblioteca

- Na biblioteca, pesquisa tem o seguinte algoritmo
 - 1. Ir até a estante de livros
 - 2. Procurar livro desejado
 - 3. Levar livro até a cadeira
 - 4. Consultar livro
 - 5. Se não concluiu pesquisa, ir para 1

Considerações

- Consulta com 10 livros
- 1 minuto para ir e voltar da cadeira a estante
- 1 minuto procurando o livro na estante
- 30 segundos para consultar a informação desejada no livro

Tempo de consulta de cada livro

- 2 minutos e 30 segundos
- Tempo total consumido
 - 25 minutos



Exemplo de Estudante em uma Biblioteca

- Novo algoritmo
 - 1. Ir até a estante de livros
 - 2. Procurar livros desejados
 - 3. Levar livros até a mesa
 - 4. Consultar livros
 - 5. Se não terminou ir para 4
- Consideração
 - A mesa tem espaço para os 10 livros
- Tempo total de pesquisa
 - 16 minutos (1 para deslocamento, 10 para procura na estante e 5 para pesquisa de material nos livros)
- Problemas possíveis
 - Nem todos livros desejados cabem na mesa
 - Outro aluno ocupa parte da mesa, fazendo uma pesquisa diferente



Princípio da Localidade

 Localidade Espacial: quando uma localização de memória é acessada, há uma tendência de que as localizações de memória próximas a ela também sejam acessadas em breve.

 Localidade Temporal: os dados que foram usados recentemente tendem a ser usados novamente em breve



Localização da Memória

Memória no processador

Memória interna

Memória externa



Palavra

- É a quantidade de <u>bytes</u> trocados entre a memória principal (RAM) e o processador (CPU), <u>por vez.</u>
- Tamanho da palavra pode <u>variar entre diferentes</u> arquiteturas
- Exemplo as <u>arquiteturas x86 e x64</u>



Bloco

- É a quantidade de <u>bytes</u> trocados com a memória externa
- Os dados geralmente são transferidos em grandes quantidades chamadas de blocos.
- Exemplo troca de dados entre HD e Memória
 RAM



MÉTODO DE ACESSO

- Sequencial
- Direto
- Aleatório
- Associativo



- Tempo de acesso: tempo que leva para recuperar ou armazenar dados em uma localização específica da memória.
- Tempo de ciclo "tempo de clock": intervalo de tempo necessário para a realização de uma operação de leitura ou escrita
- Taxa de transferência: velocidade com a qual os dados podem ser transmitidos



TECNOLOGIA

- Semicondutores
- Magnética
- Óptica
- Magneto-óptica

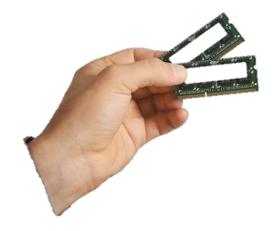


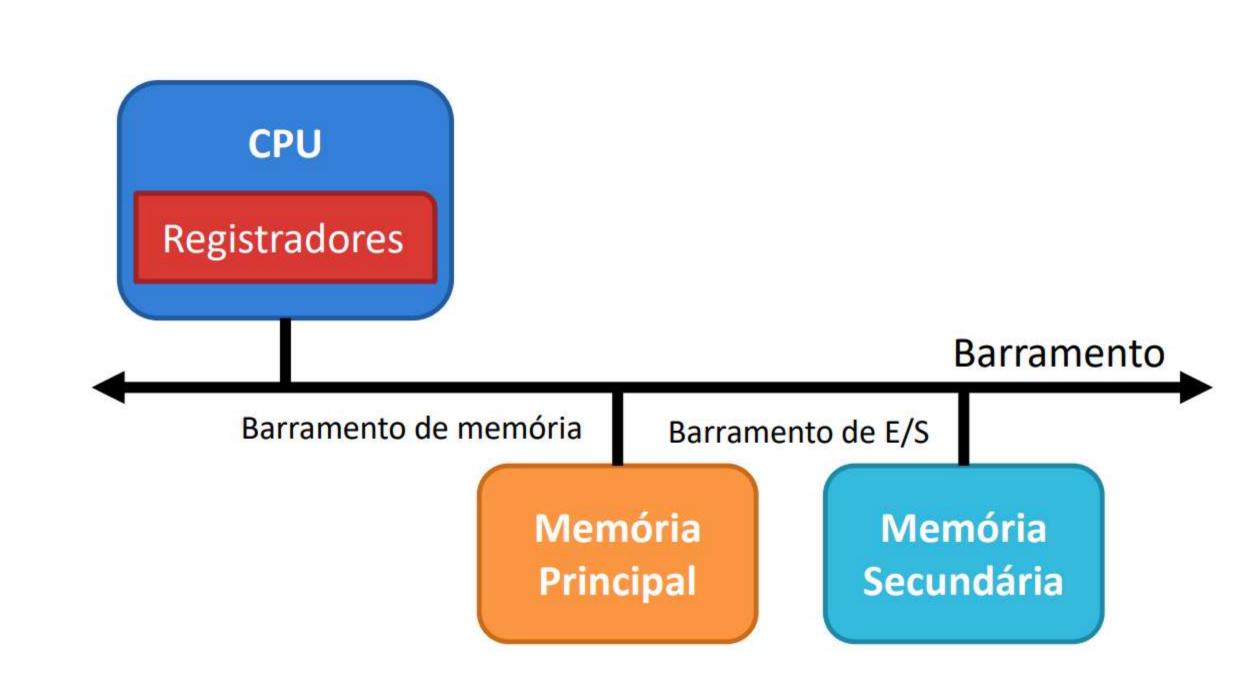
CARACTERÍSTICAS FÍSICAS

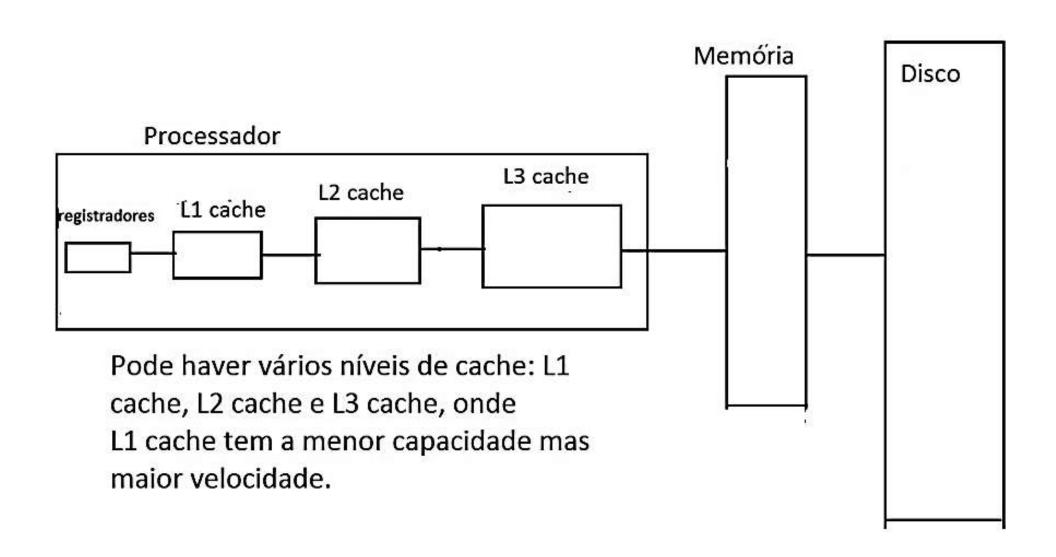
Volátil/Não volátil

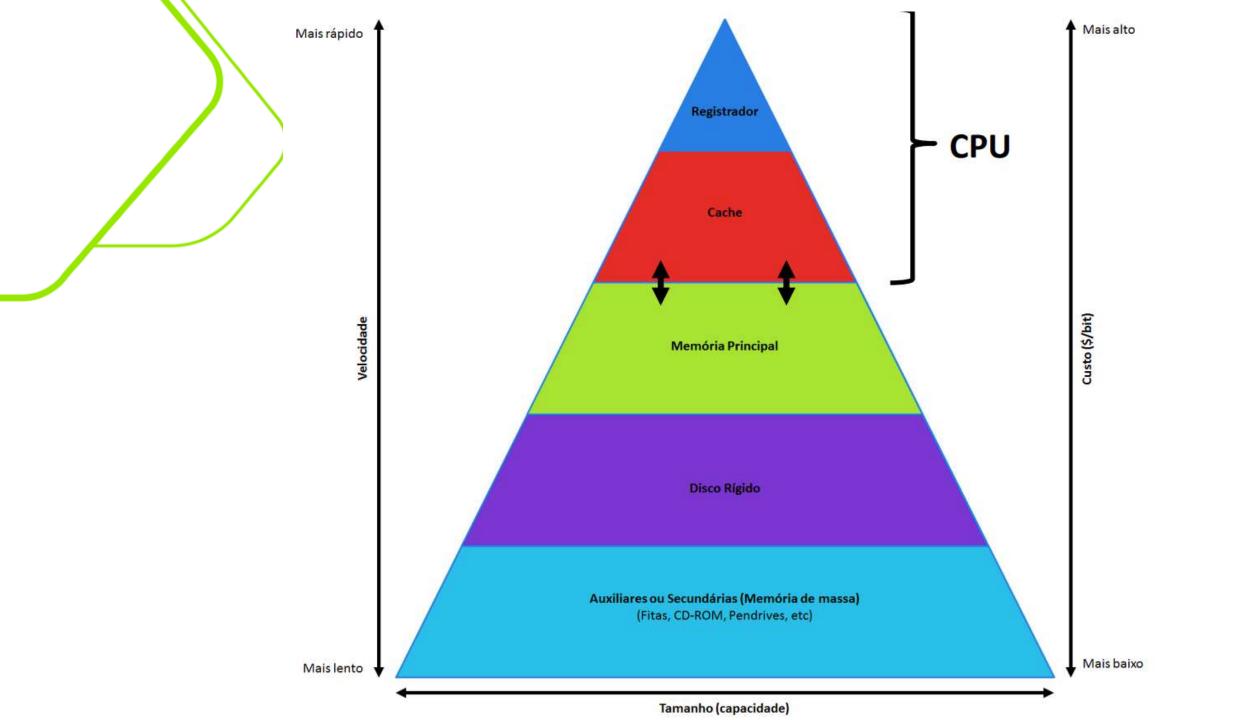
Apagável/Não apagável

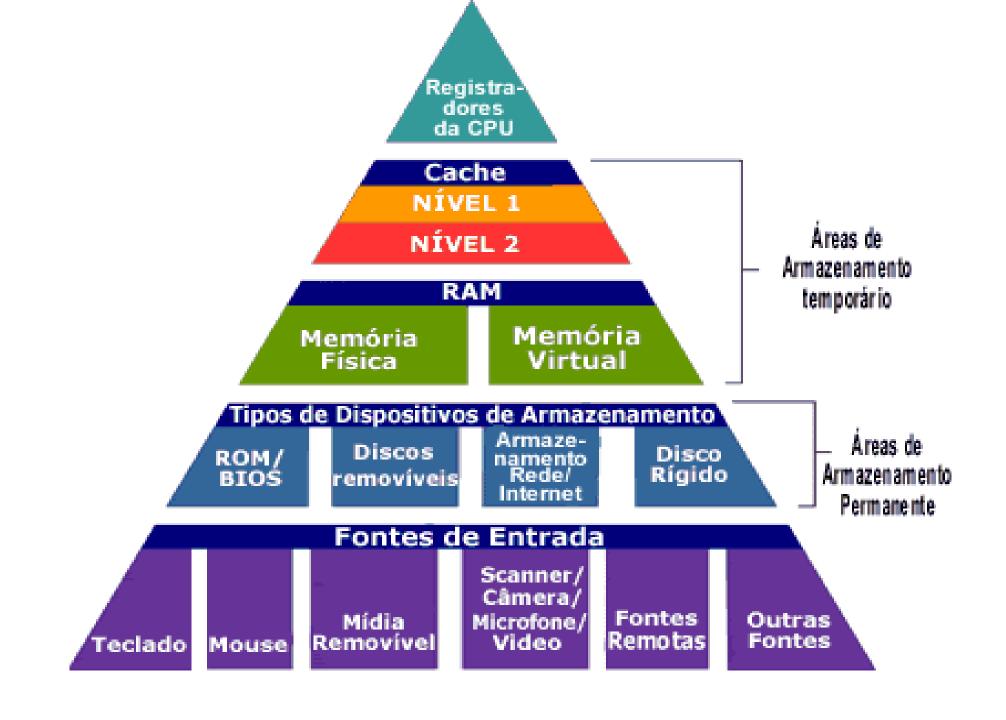


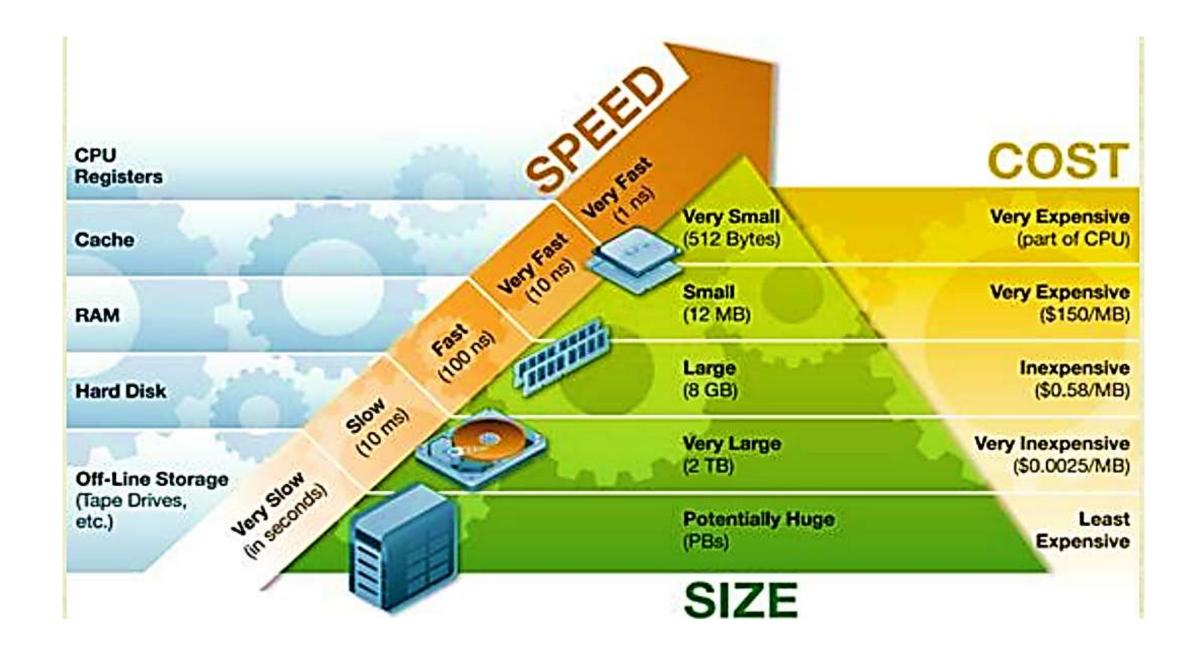


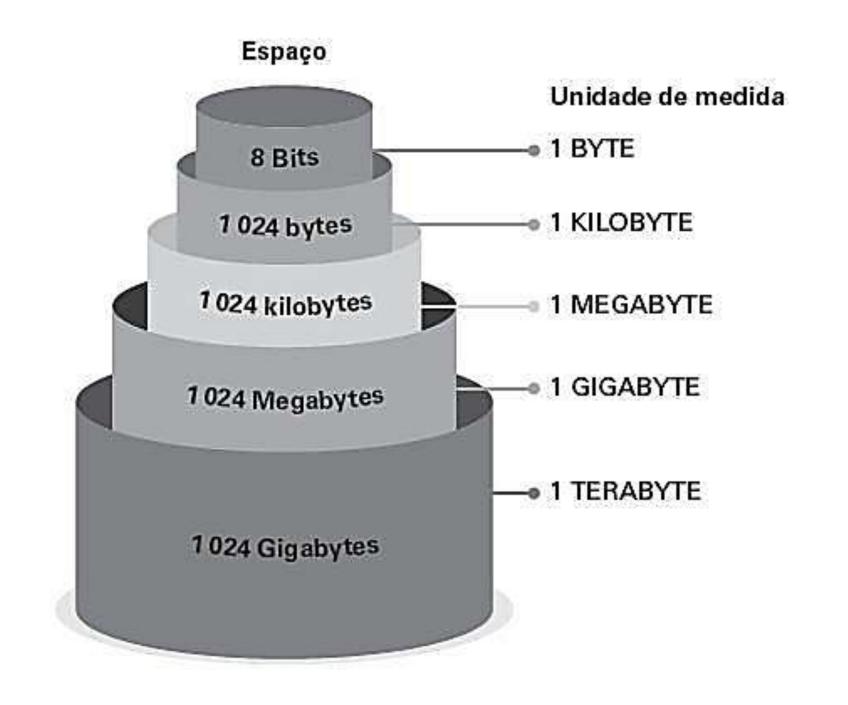












Cisco previu que o tráfego global de internet alcançaria 4,8 zettabytes por mês até 2022

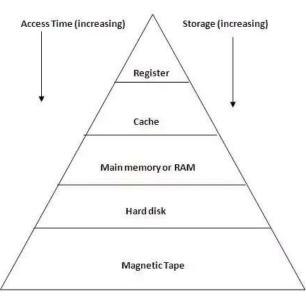
Unidades de Medida da Informática			
Unidade	Símbolo	Valor	
Bit	b		
Byte	В	8 bits	
Kilobyte	KB	1024 Bytes	
Megabyte	MB	1024 Kilobytes	
Gigabyte	GB	1024 Megabytes	
Terabyte	ТВ	1024 Gigabytes	
Petabyte	PB	1024 Terabytes	
Exabyte	EB	1024 Petabytes	
Zettabyte	ZB	1024 Exabytes	
Yottabyte	YB	1024 Zettabytes	





CARACTERÍSTICAS FÍSICAS

À medida que descemos em uma hierarquia de memórias, temos as seguintes relações:



- O custo por bit diminui
- A capacidade aumenta
- O tempo de acesso aumenta
- A frequência de acesso à memória pelo



Desafios

Movimentação de dados entre níveis necessita de mecanismos

Ex.: Movimento de dados para um nível superior que já está cheio

Quem retirar?

 Decisão errada pode afetar desempenho do sistema como um todo

Hit – dado <u>encontrado</u> no nível procurado

 Miss – dado <u>não encontrado</u> no nível procurado

Hit-rate – percentual de hits no nível, Ex.:
 70%

- Miss-rate percentual de misses no nível,
 Ex.: 30% (complementar ao Hit-rate)
- Hit-time tempo de acesso ao nível incluindo tempo de ver se é hit ou miss
- Miss-penalty tempo médio gasto para que o dado não encontrado no nível seja transferido dos níveis mais baixos

 Tempo médio de acesso à memória (AMAT)

Tempo médio efetivo para <u>acessar um</u> dado em certo nível de memória

 $AMAT = \text{Hit time} + \text{Miss rate} \times \text{Miss penalty}$

 Tempo médio de acesso à memória (AMAT)

Calcule o tempo médio efetivo de acesso (AMAT) a uma cache com?

- Miss-rate = 20%
- Hit-time = 2 µs e
- Mice-nenalty 10 He

 $AMAT = \text{Hit time} + \text{Miss rate} \times \text{Miss penalty}$

$AMAT = \text{Hit time} + \text{Miss rate} \times \text{Miss penalty}$

$$AMAT = 2 \mu s + 0,20 \times 10 \mu s$$

$$AMAT = 2 \mu s + 2 \mu s$$

$$AMAT = 4 \mu s$$



Disciplina: Arquitetura de Computadores

Tema: <u>Tipos de memória</u>

Prof. Plínio Andrade Passos

Tecnologias de Memórias

Tecnologia de Memória	Volatilidade	Uso Comum
DRAM (Dynamic RAM)	Volátil	Memória Principal
SRAM (Static RAM)	Volátil	Caches, Registradores
Flash Memory	Não Volátil	Armazenamento de Dados
ROM (Read-Only Memory)	Não Volátil	Firmware
EPROM (Erasable PROM)	Não Volátil	Programação de Firmware
EEPROM (Electrically Erasable PROM)	Não Volátil	Armazenamento de Dados



- Memória Dinâmica DRAM (Dynamic Random Access Memory)
 - Tipo de memória Volátil
 - Maior capacidade de armazenamento do que a SRAM
 - > ACESSO ALEATÓRIO
 - > Precisa de **refresh** constantemente
 - Mais lenta que SRAM
 - Mais barata que SRAM

Quem utiliza esse tipo Memória Dinâmica - DRAM (Dynamic Random Access Memory)?

- ✓ Memória Principal (RAM)
- √ RAM dos Smartphones e Tablets
- √ Placas de Vídeo (GPUs)
- √ Roteadores e Switches de Rede etc...

- Memória Estática Static Random-Access Memory (SRAM)
 - > Tipo de memória Volátil
 - Muito mais rápida que DRAM
 - > DISPENSA refresh
 - > Acesso Aleatório
 - > Menor capacidade que DRAM



Quem utiliza esse tipo Memória Estática - SRAM ?

✓ Memória Cache

✓ Registradores

✓ Aplicações Aeroespaciais

Tipo	Tempo de acesso	Custo
SRAM	0,5 ns à 2,5 ns	\$ 2000,00 à \$ 5000,00 por GB
DRAM	50 ns à 70 ns	\$ 20,00 à \$ 75,00 por GB
HD	5 ms à 20 ms	\$ 0,20 à \$ 2,00 por GB



- Memória dentro da própria CPU que armazena temporariamente n bits.
- Estão no topo da hierarquia de memória
- Memória mais rápida e financeiramente mais custosa

Registradores

• Exemplos de registradores da arquitetura

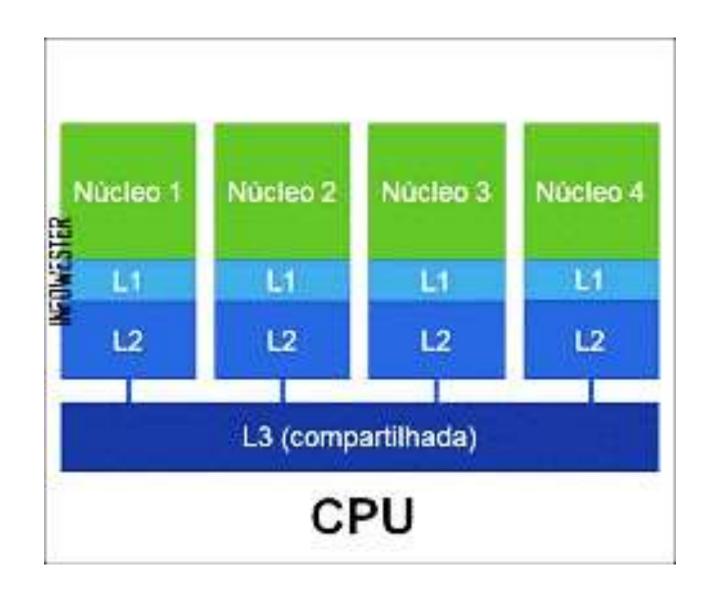
Registrador	Palavras-Chave das Funções
EAX	Acumulador, Operações Aritméticas e Lógicas, Retorno de Função
EBX	Base, Ponteiro para Dados, Endereçamento
ECX	Contador, Controle de Laços, Contagem
EDX	Dados, Operações de Divisão, Controle de Portas
ESI	Índice de Origem, Operações de Cadeia de Caracteres
EDI	Índice de Destino, Operações de Cadeia de Caracteres
ESP	Ponteiro de Pilha, Gerenciamento de Chamadas de Função
EBP	Ponteiro de Base de Pilha, Acesso a Parâmetros de Função
EIP	Ponteiro de Instrução, Endereço da Próxima Instrução
EFLAGS	Flags de Status, Controle de Condições e Interrupções

Memória CACHE

- Memória temporária que atua entre a RAM e o Processador
- Armazena cópias de dados da RAM que são frequentemente usados
- Quanto mais dados o processador puder acessar da memória cache, mais rápido será o desempenho do computador.

Memória CACHE

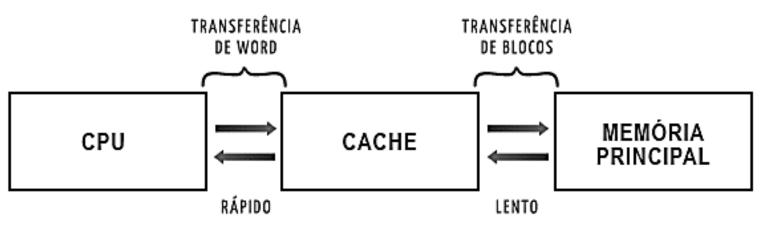
- Memória cache é dividida em diferentes níveis – L1, L2 e L3.
- Cada nível tem uma quantidade específica de armazenamento e uma velocidade de acesso específica.
- O cache L1 é o menor e o mais rápido, enquanto o cache L3 é o maior e o mais lento.

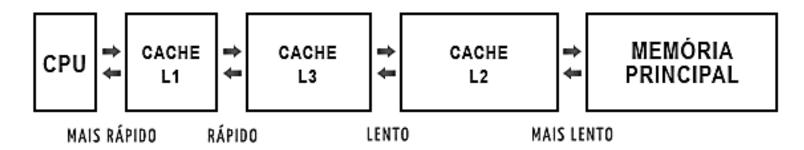


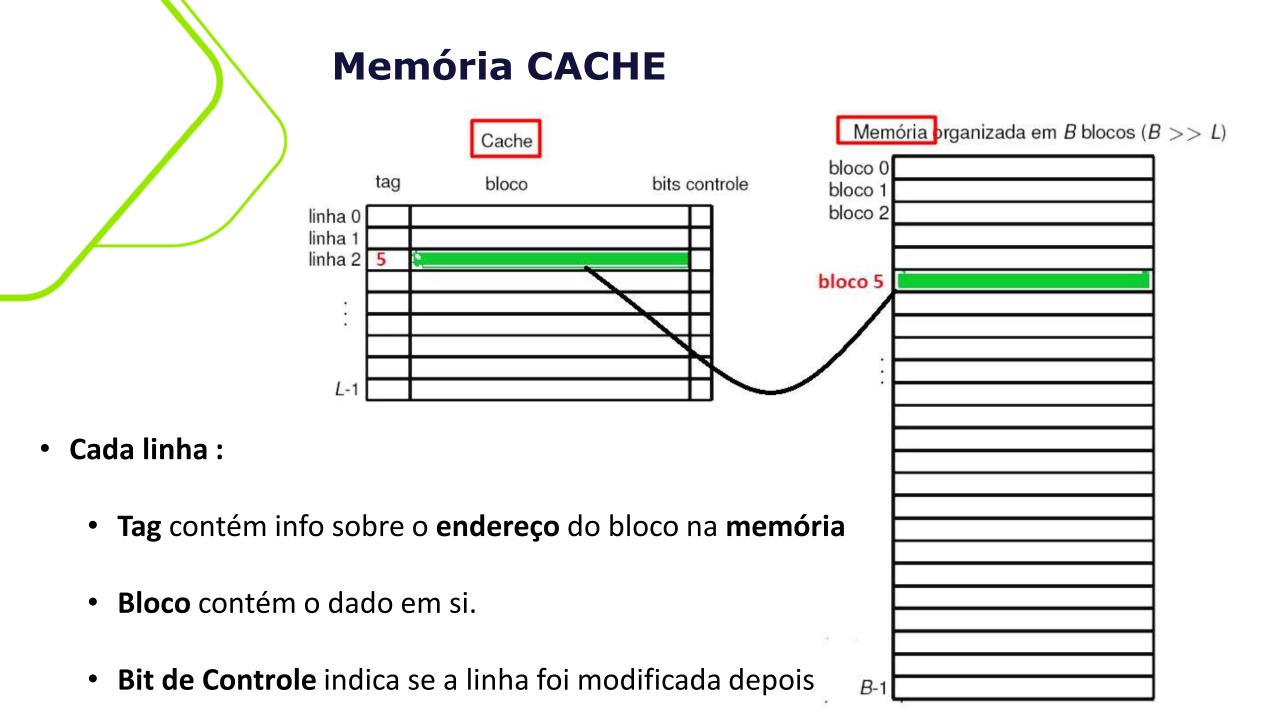
Memória CACHE

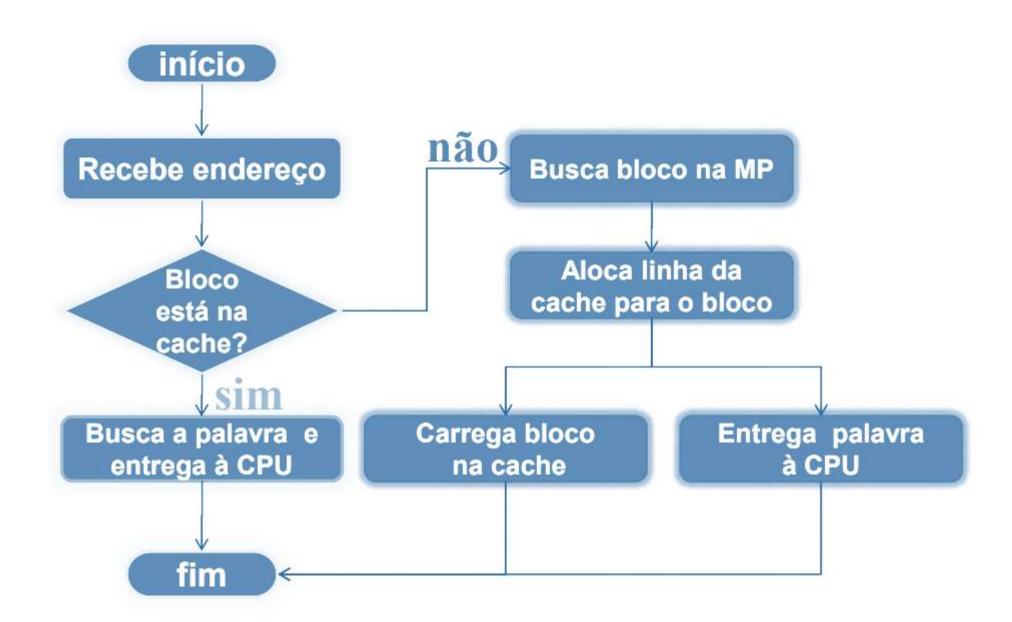
Funcionamento:

- Quando o processador precisa de dados, ele primeiro verifica a memória cache. Se achar o dado ocorre "cache hit", e o processador pode pegá-los rapidamente.
- Se os dados não estiverem no cache ocorre "cache miss", o processador terá que buscar os dados na memória RAM, que é um processo mais lento.







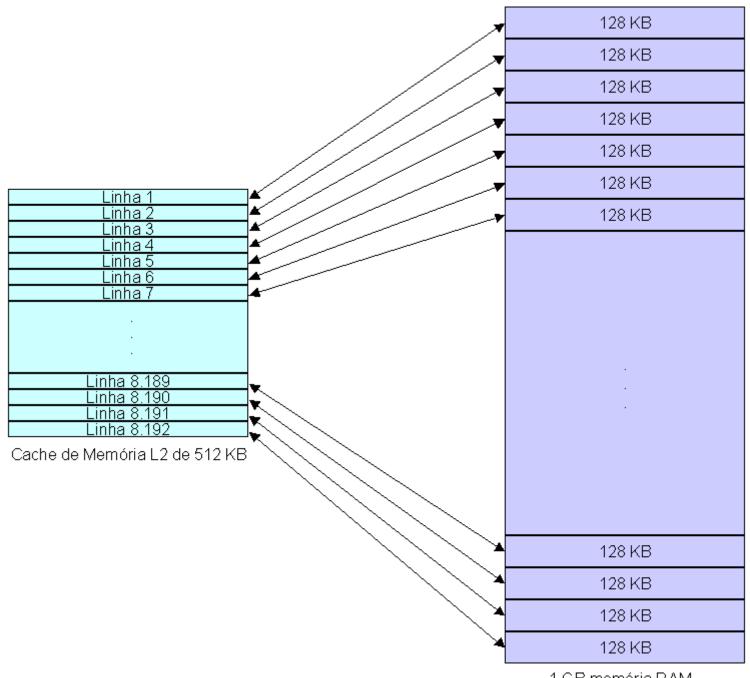


- É necessário definir como <u>mapear blocos de memória a</u> <u>linhas da memória cache</u>.
- A escolha da função de mapeamento <u>determina</u> como a cache é organizada

Três tipos de mapeamento:

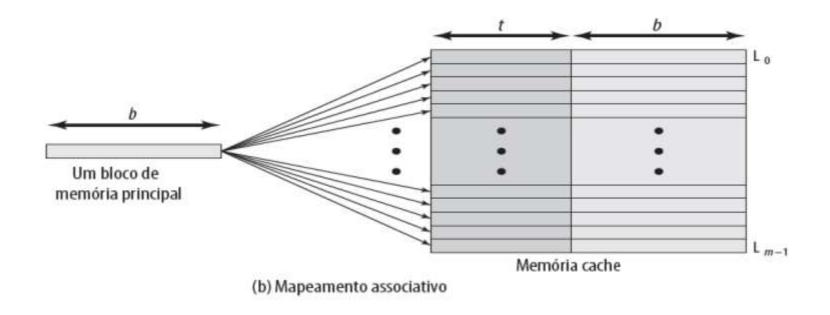
- Mapeamento direto
- Mapeamento associativo
- Mapeamento associativo por conjunto

- Mapeamento direto
 - Cada <u>bloco</u> na memória principal é mapeado em um <u>único bloco da cache</u>
 - Cada dado possui uma posição fixa
 - A técnica de mapeamento direto é simples e tem custo de implementação baixo;
 - Gera muito misses!



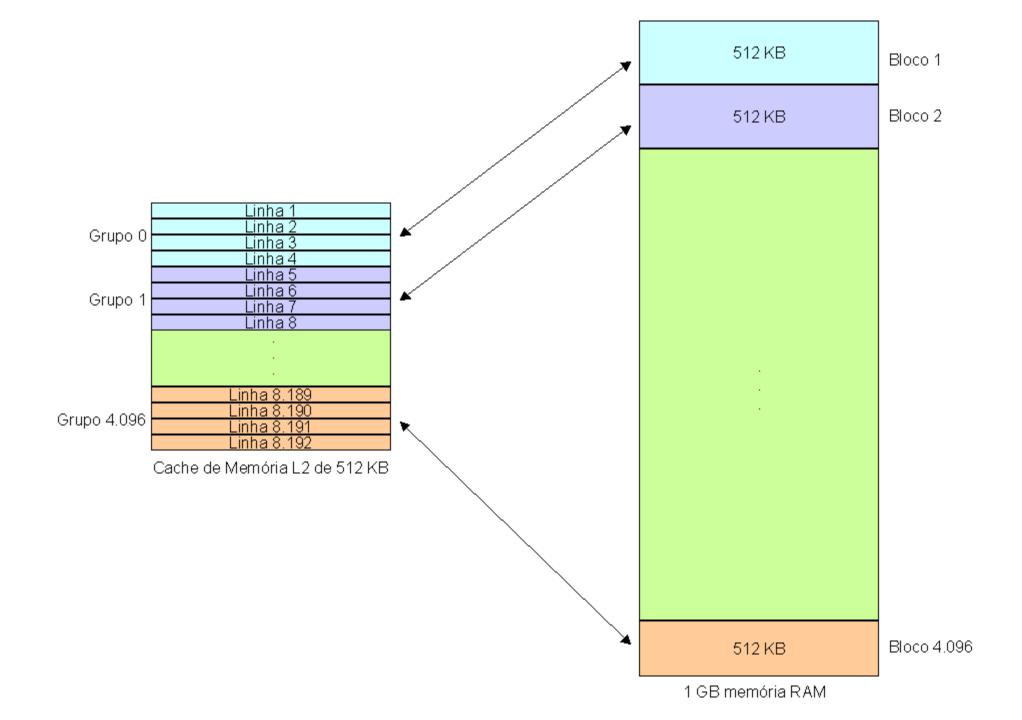
1 GB memória RAM

- Mapeamento Associativo
 - Não existe uma posição fixa entre as linhas do cache de memória e as localizações da memória RAM.
 - A cache pode armazenar dado de <u>qualquer</u> endereço da RAM.
 - Controle complexo



- Mapeamento Associativo por Conjunto
 - O cache de memória é <u>dividido em vários</u> <u>blocos (grupos)</u> contendo "n" linhas cada..

- "Mapeamento Direto" para cada Grupos/Bloco
- Modelo mais usado atualmente.

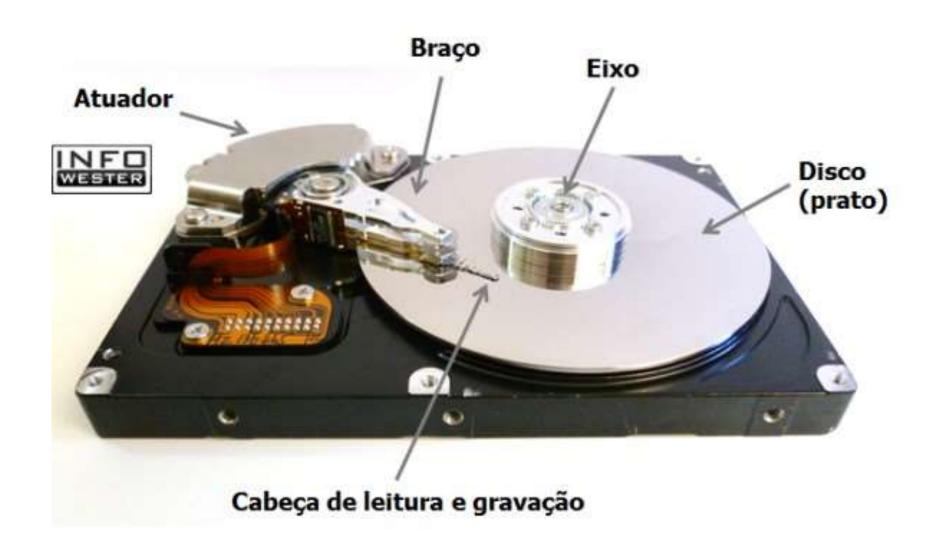


Memória ROM (Read Only Memory)

- Via de regra usada somente para leitura.
- Tipos de memória ROM são:
 - MASK-ROM (MROM)
 - PROM (Programmable Read-Only Memory):
 - > EPROM (Erasable Programmable Read-Only Memory):
 - EEPROM (Electrically Erasable Programmable Read-Only Memory) SSD

HD (hard disk – disco rígido, ou winchester)

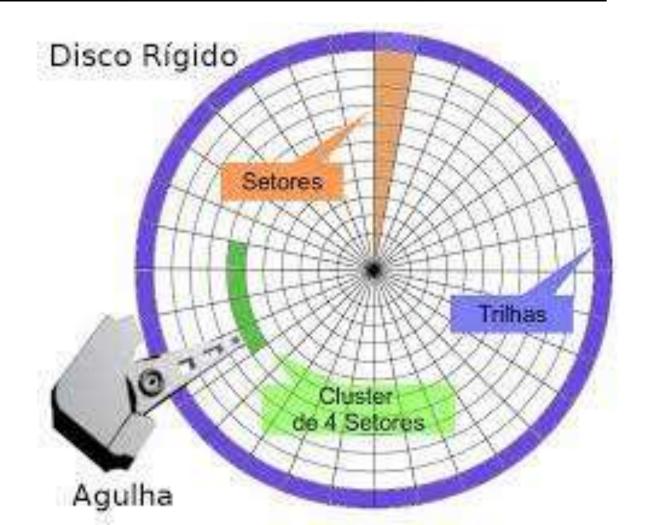
- ☐ Armazena dados de forma "permanente"
- Computador pode funcionar sem ele.
- ☐ Custo reduzido
- ☐ Grande espaço de armazenamento
- ☐ Desvantagem maior é o **tempo de leitura/escrita**





HD (hard disk – disco rígido, ou winchester)

Via de regra usada somente para leitura.





Não possui partes eletromecânicas.



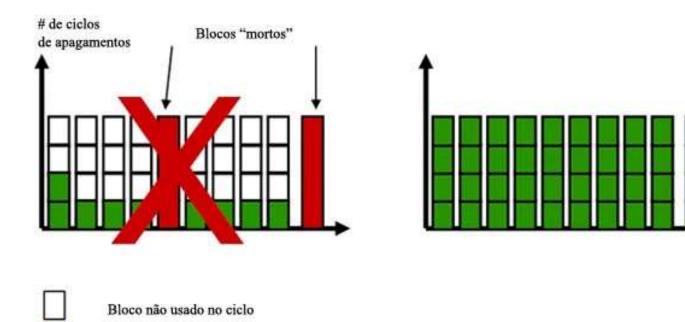
- Vida útil pode ser um pouco menor que o HD
- Células de armazenamento dos dados são "gastas a cada escrita/reescrita. O número de gravações em cada célula é limitado.
 - Necessário <u>nivelamento de desgaste</u>

Sem Nivelamento

Bloco usado no ciclo

Bloco desgastado

Com Nivelamento





centro universitário de excelência