Arquitetura e Organização de Computadores

Uma Introdução

Gabriel P. Silva – José Antonio Borges

Memória e Hierarquia de Memória

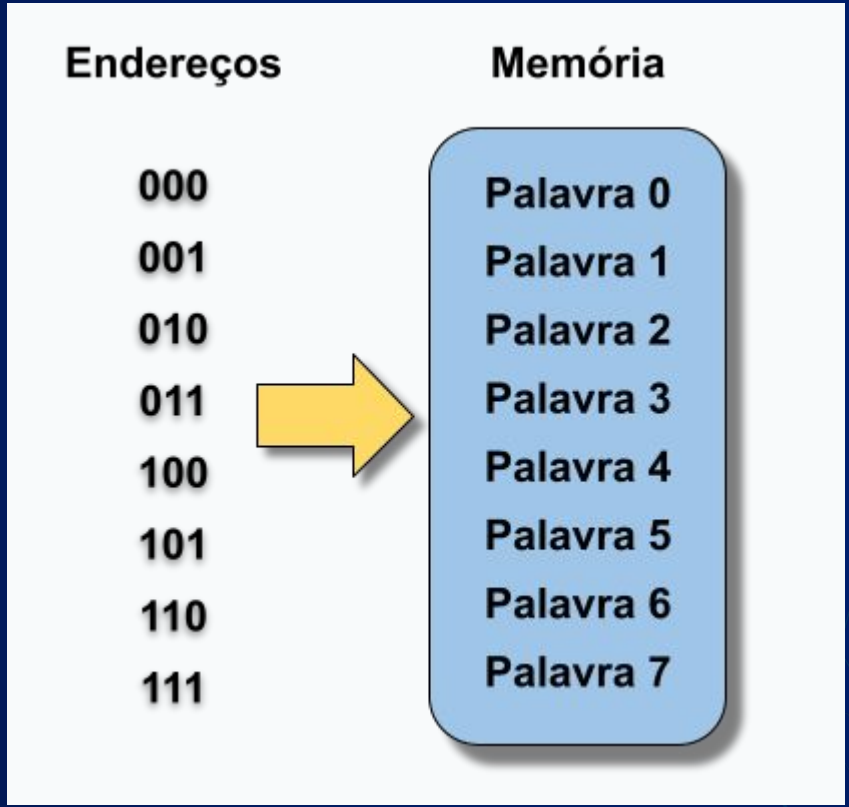
Capítulo 4

4.1 Memória

Memória

- Uma palavra de memória é um grupo de células que, nos computadores atuais, pode armazenar entre 4 e 8 bytes, sendo que normalmente uma palavra de memória é lida ou escrita de uma única vez.
- Cada posição de memória possui um endereço diferente. Por exemplo, uma memória com 32 posições precisaria de um endereço de 5 bits para que todo o seu conteúdo fosse acessado; já uma memória com 64 precisaria de 6 bits e assim por diante.
- Uma memória com N bytes, precisa de $\log_2(N)$ bits para ser endereçada.

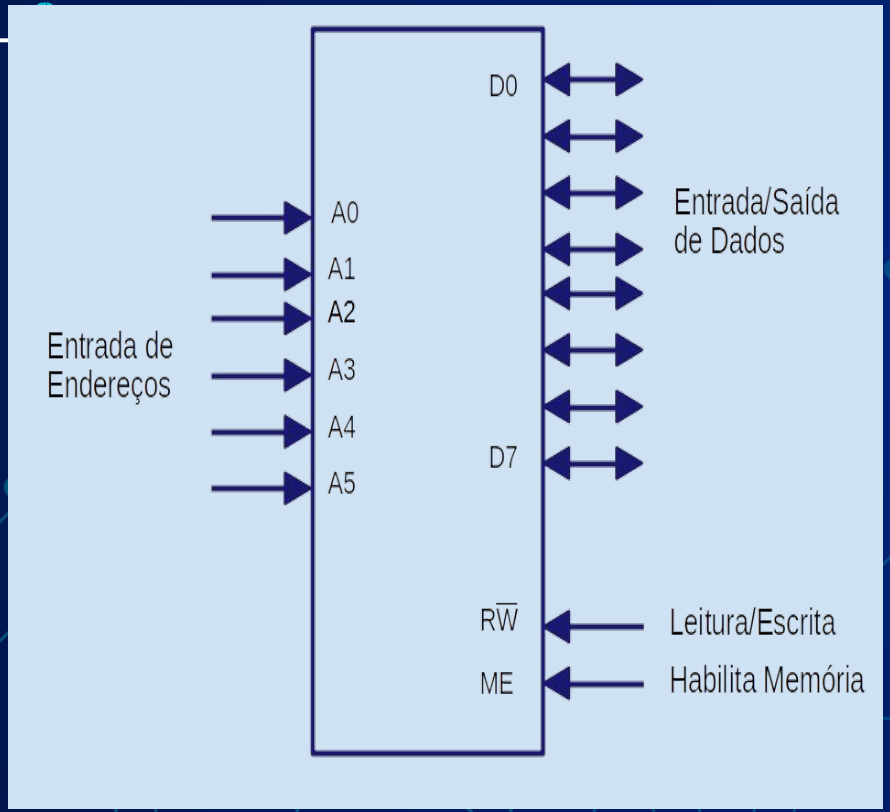
Endereçamento da Memória



Memória

- Entrada/Saída de Dados (D0-D7)}: contém as palavras que serão lidas ou escritas na memória.
- Entradas de Endereço (A0-A5)}: como a memória possui capacidade de armazenar 64 palavras deve ter 6 bits de entradas de endereço.
- Entrada R/ W: Determina qual das operações de memória deverá ser efetuada. .
- Habilitação da Memória (ME)}: responsável pela habilitação e desabilitação da pastilha de memória
- Cada tipo de memória pode ter outras linhas específicas de controle.

Memória



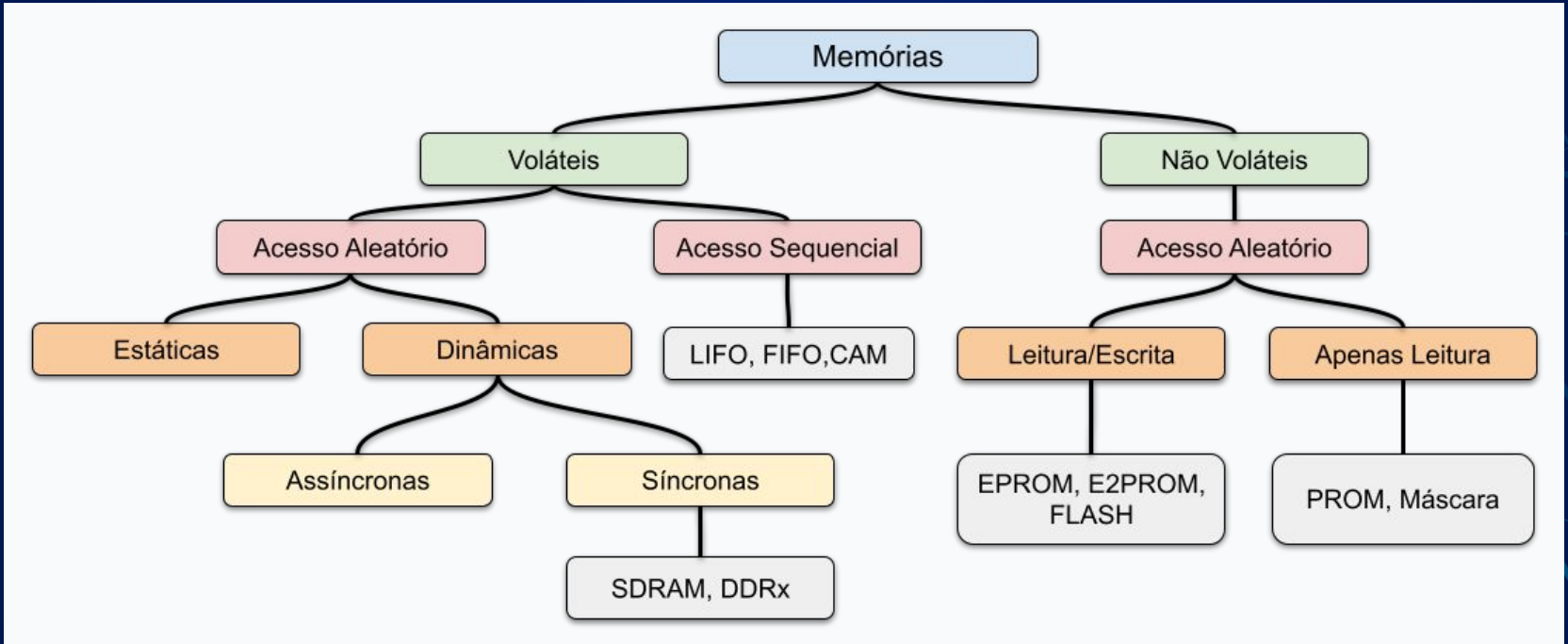
A instrução tem um tamanho de 4 bytes de comprimento, e três formatos diferentes.

Classificação das Memórias

- As memórias podem apresentar propriedades distintas, de acordo com a tecnologia com que são fabricadas e são utilizadas em aplicações diferentes, de acordo com a velocidade de leitura e escrita dos dados, capacidade de armazenamento, volatilidade da informação, consumo, etc.
- A maioria esmagadora das memórias utilizadas no computador são de acesso aleatório, ou seja, necessitam de um endereço para determinar onde as informações estão/serão armazenadas.

4.2 Classificação das Memórias

Classificação das Memória



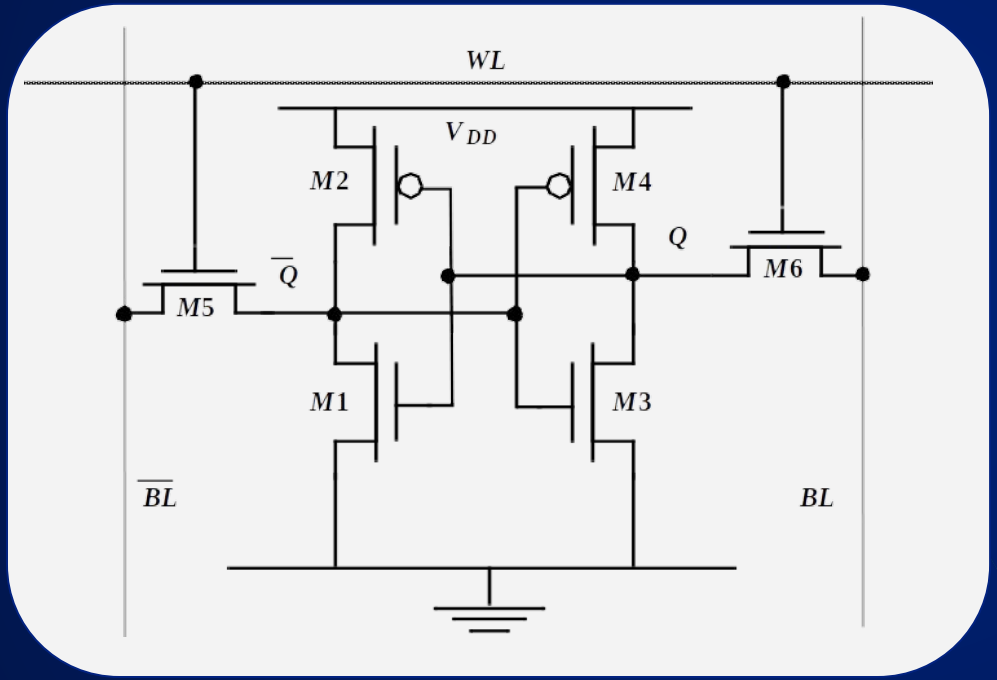
Memórias Voláteis e Não Voláteis

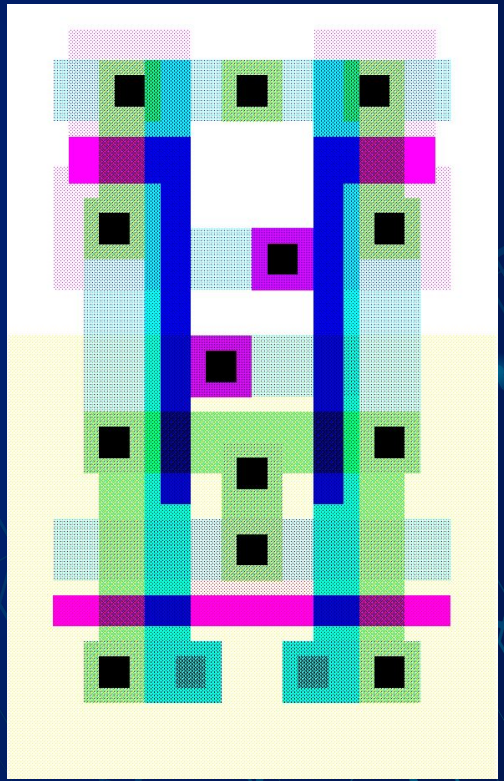
- As memórias também podem ser classificadas com relação à capacidade de manter o seu conteúdo, depois que deixam de ser alimentadas por uma fonte de energia elétrica.
- Podem ser então de dois tipos: voláteis e não-voláteis. A memória principal do computador é formada na maior parte com memória volátil e uma pequena parcela de memória não-volátil.

Memórias Estáticas e Dinâmicas

- As memórias dinâmicas recebem este nome porque necessitam que a informação armazenada seja periodicamente atualizada, isto é, elas precisam ser lidas e novamente escritas sob o risco dessas informações serem perdidas.
- As memórias estáticas não precisam deste tipo de operação, preservando a informação enquanto houver alimentação.

Célula de Memória Estática

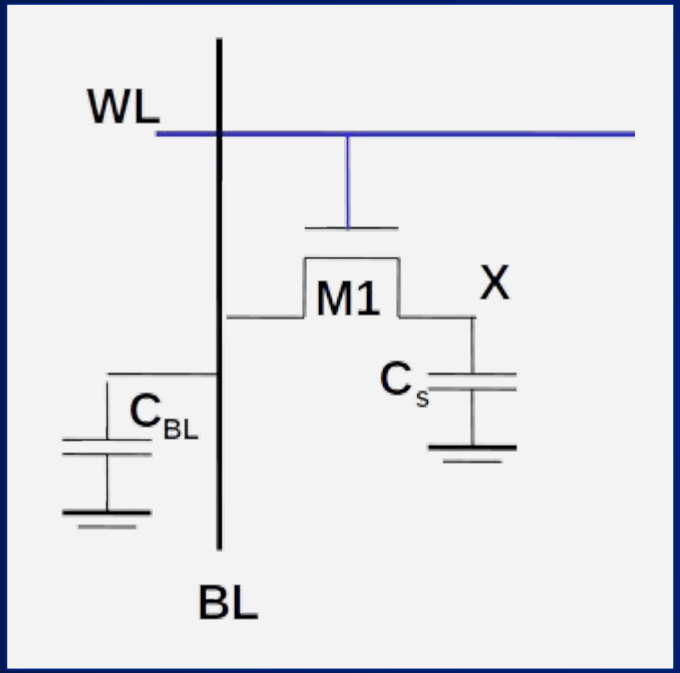


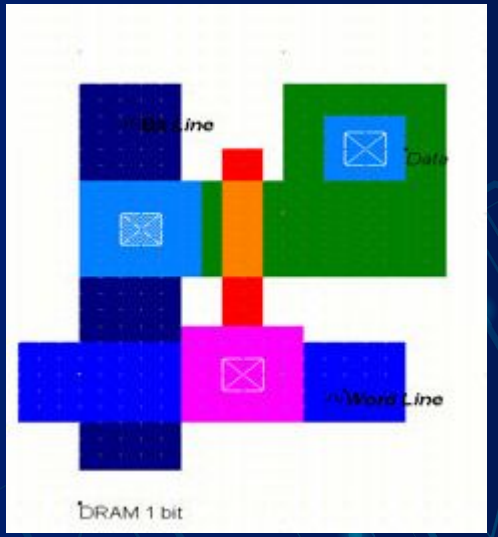


Memórias Estáticas

Resumo	
Vantagens	**Desvantagens**
Os dados permanecem armazenados enquanto houver alimentação	As células de memória são maiores, com cerca de 6 transistores
São mais rápidas	O consumo de energia é maior
O estado da célula é estável, não precisa de ciclos de atualização	A capacidade de armazenamento é menor

Célula de Memória Dinâmica

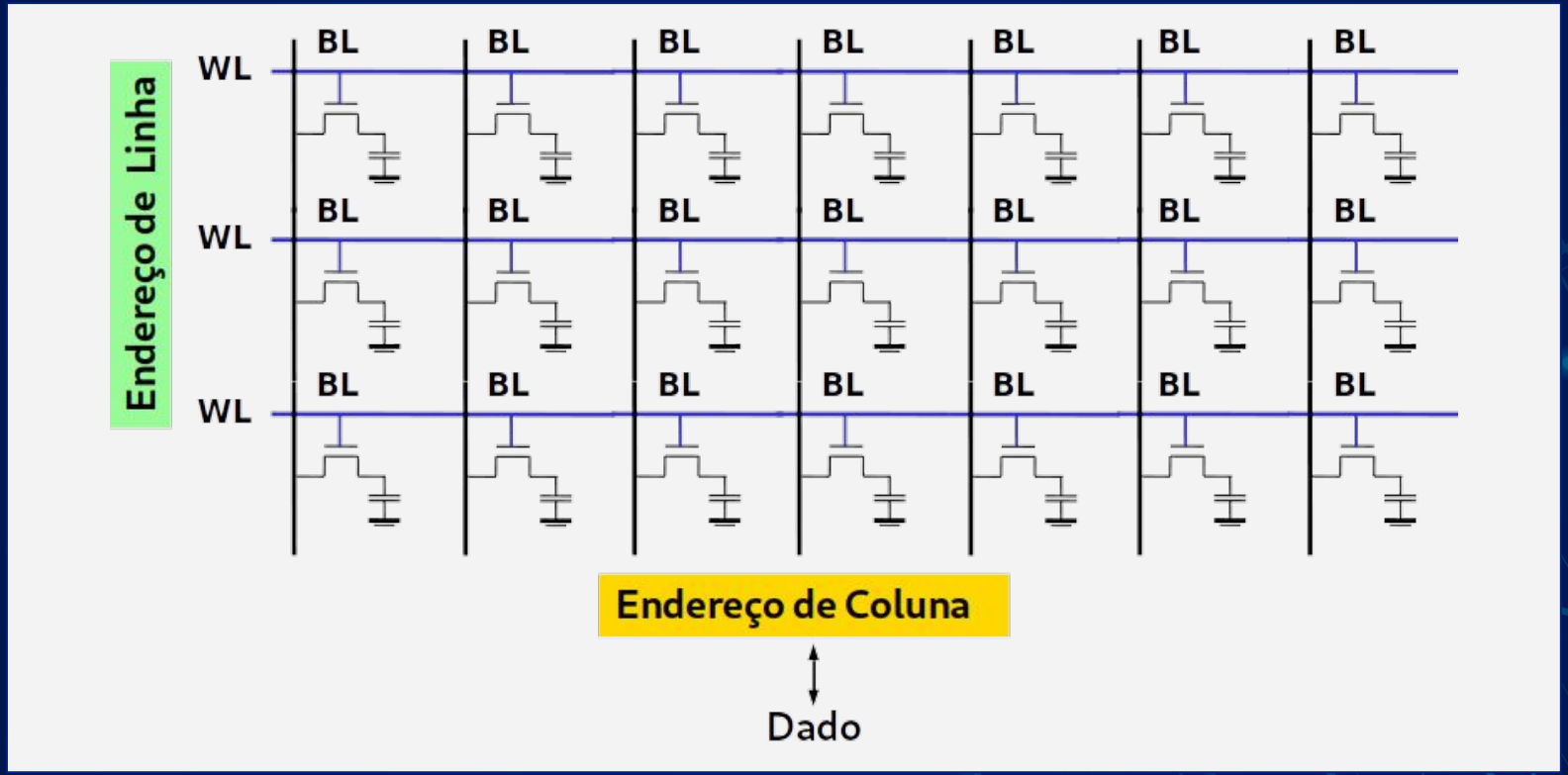




Memórias Dinâmicas

Resumo	
Vantagens	**Desvantagens**
As células de memória são menores, com apenas um transistor e um capacitor	A atualização periódica dos dados é necessária
O consumo de energia é menor	A leitura é destrutiva e requer uma atualização em seguida
A capacidade de armazenamento é maior	São mais lentas que as estáticas

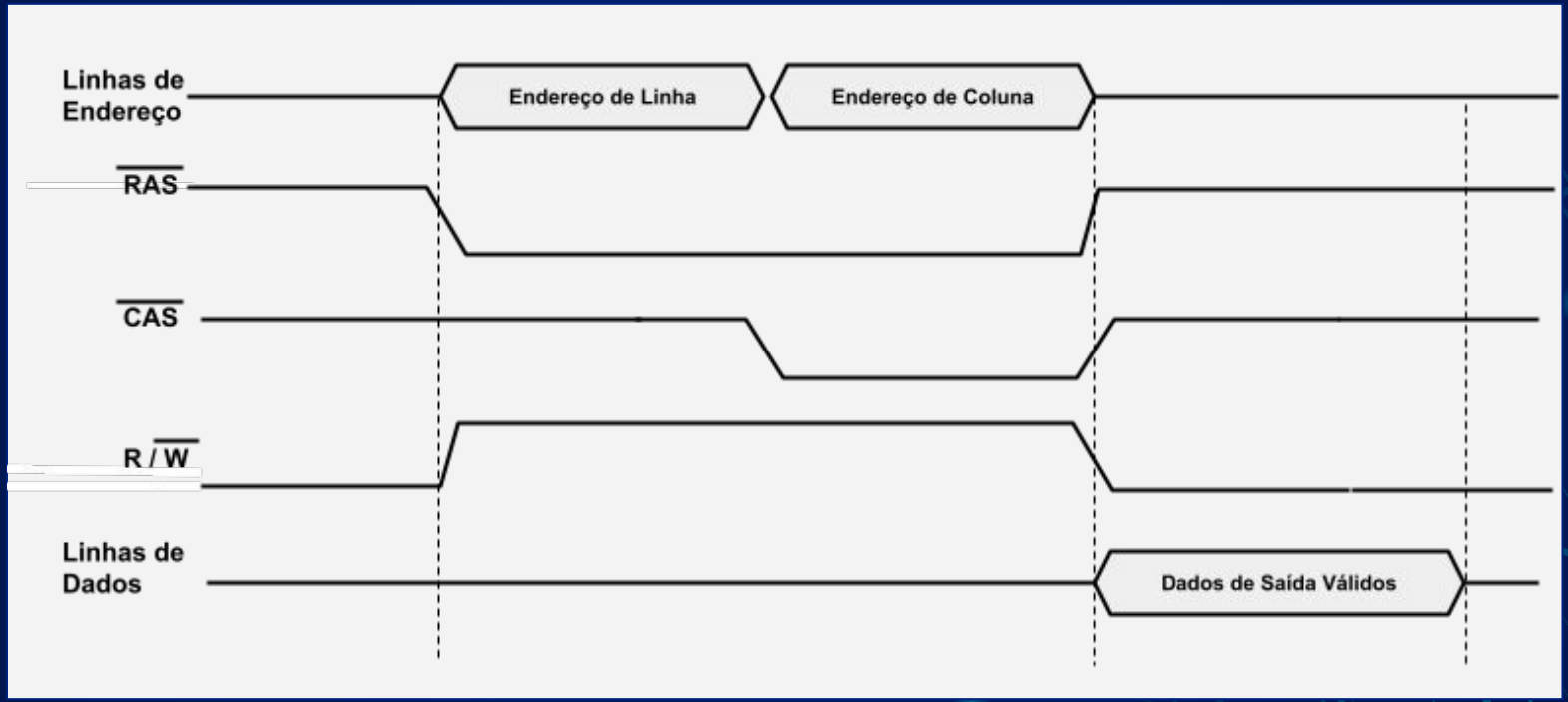
Matriz de Memória



Memórias Assíncronas

- As memórias dinâmicas assíncronas requerem um protocolo mais simples, porém menos eficiente, para serem acessadas.
- O barramento de endereços é compartilhado, ou seja, o endereço de linha é fornecido antes do endereço de coluna, nas mesmas linhas de acesso à memória.
- As memórias dinâmicas assíncronas foram sendo gradativamente substituídas pelas memórias síncronas.

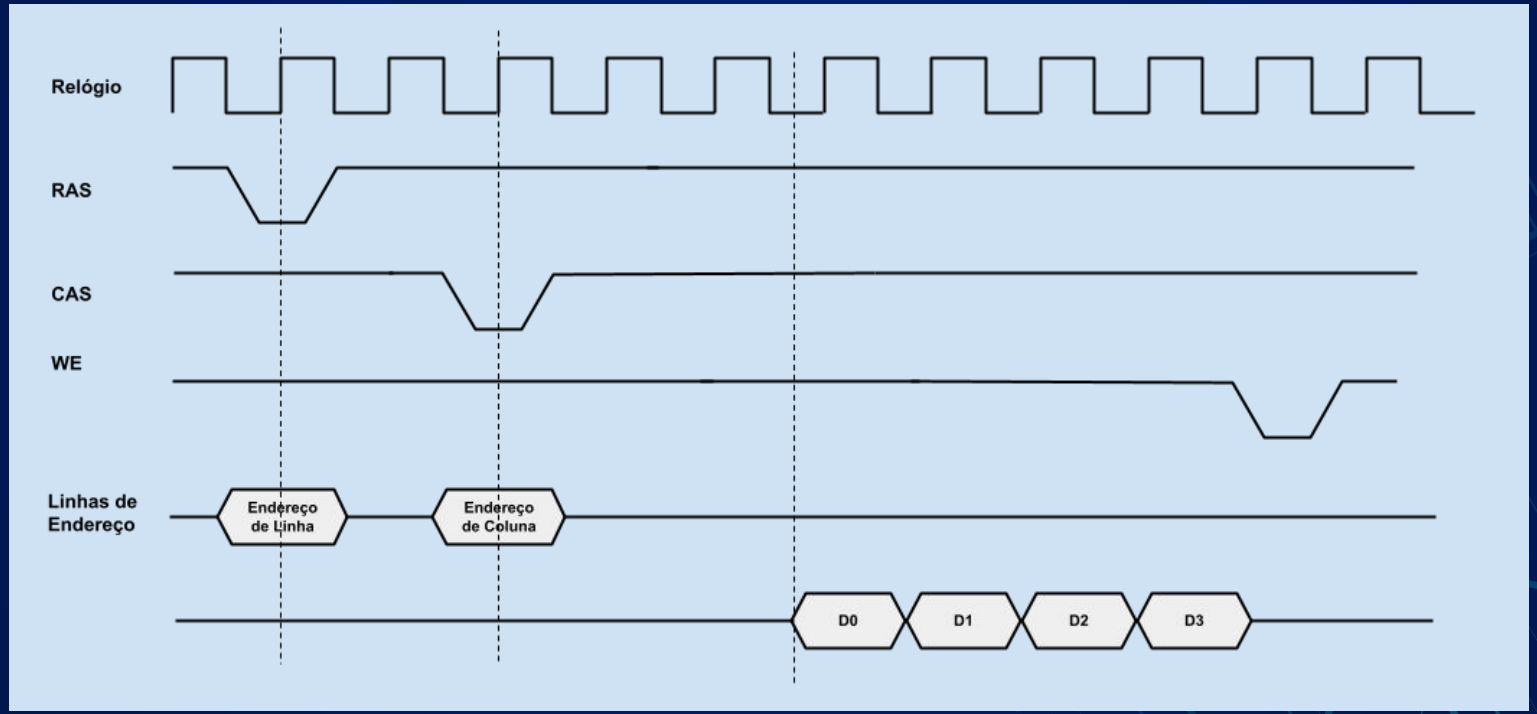
Memória Assíncrona



Memórias Síncronas

- As memórias dinâmicas síncronas são um tipo de memória de acesso aleatório (DRAM), voláteis, onde a leitura ou escrita dos dados é sincronizada por um relógio de sistema ou de barramento.
- São projetadas para permitir a leitura ou escrita, depois da latência inicial, em modo rajada (burst mode) com uma taxa de um ciclo de relógio por acesso.

Memória Síncrona



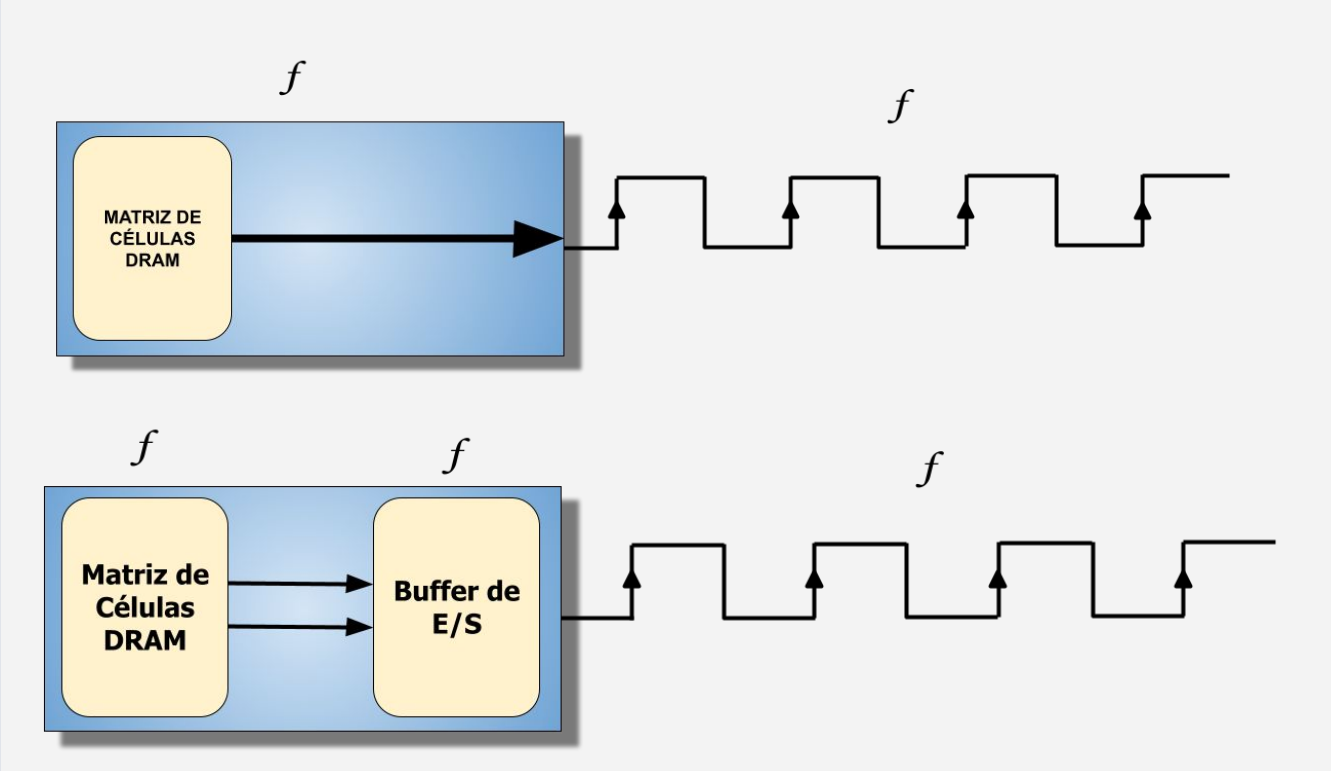
Memórias SDRAM

- As memórias SDRAM (Single Data Rate DRAM) são módulos de memória que podem transferir uma palavra de dados por ciclo de relógio, seja na leitura ou na escrita.
- As velocidades típicas de relógio dessas memórias são 100 ou 133 MHz, cujos módulos são conhecidos como PC-100 e PC-133.
- Na memória SDRAM, a matriz de células de memória opera na mesma frequência do barramento externo, ou seja, 100 ou 133 MHz.

Memórias DDR

- As memórias DDR SDRAM (Double-data-rate Synchronous Dynamic Random Access Memory}) alcançam maior largura de banda através da transferência de dados tanto na subida como na descida do sinal de relógio.
- Efetivamente, isso praticamente dobra a taxa de transferência sem aumentar a frequência da interface de barramento do processador com a memória.
- Assim, uma célula de memória DDR-200 opera na realidade com uma frequência de relógio de apenas 100 MHz e possui uma largura de banda de cerca de 1600 MB/s.

SDRAM e DDR


f
f
MATRIZ DE CÉLULAS DRAM
f
f
f
Matriz de Células DRAM
Buffer de E/S

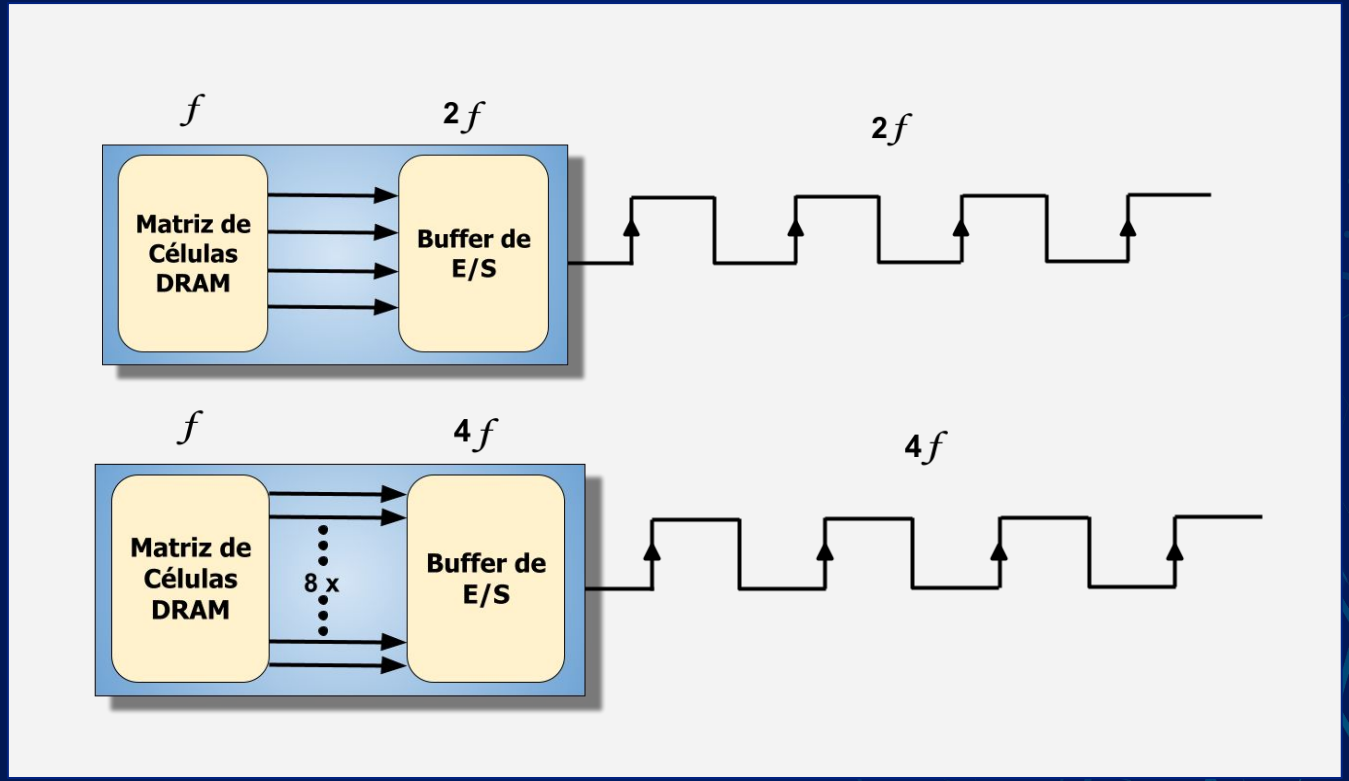
Memórias DDR2

- As memórias DDR2 transferem os dados tanto na subida como na descida do relógio.
- A diferença principal entre elas é que as matrizes de memória são organizadas em quatro bancos, e a frequência interna dos buffers e do barramento externo da DDR2 é o dobro da velocidade das células de memória, permitindo que quatro palavras de dados sejam transferidos por ciclo de memória interna.
- Então, sem acelerar as células de memória propriamente ditas, a DDR2 pode operar efetivamente com o dobro da velocidade de uma memória DDR.

Memórias DDR3

- A matriz de memória dos módulos DDR3 estão organizadas internamente em oito bancos que trabalham a um quarto da frequência do barramento externo, o requer um buffer capaz de armazenar 8 bits em comparação com os 4 bits da memória DDR2.
- As memórias DDR3 são alimentadas com 1,5 volts, possuem latência típica de 5 a 17 ciclos de relógio e velocidades de relógio do barramento externo de até 1600 MHz.

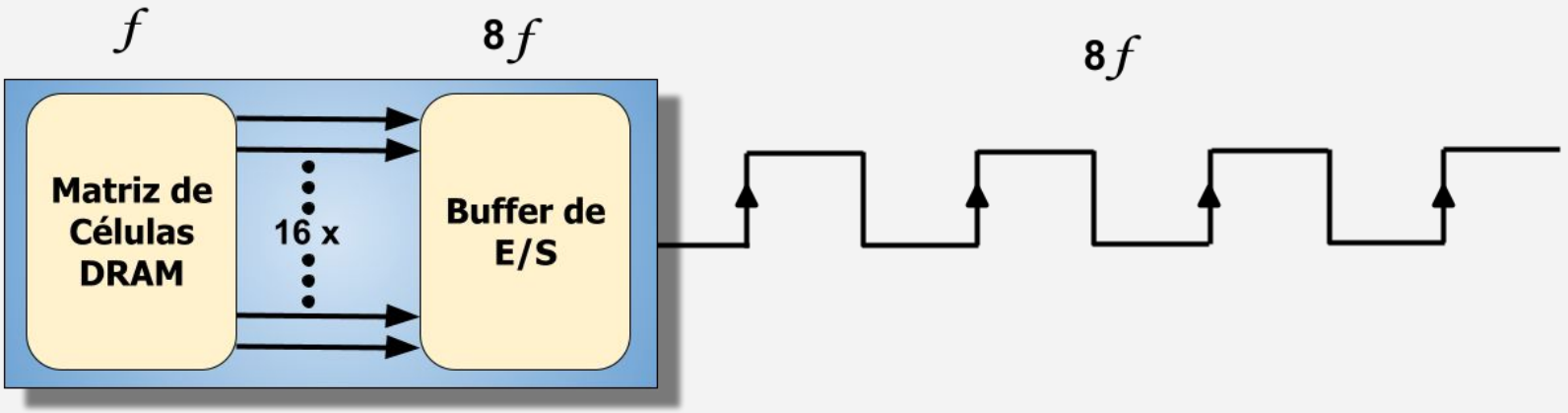
DDR2 e DDR3


f
2f
Matriz de Células DRAM
Buffer de E/S
2f
f
4f
Matriz de Células DRAM
8 x
Buffer de E/S
4f

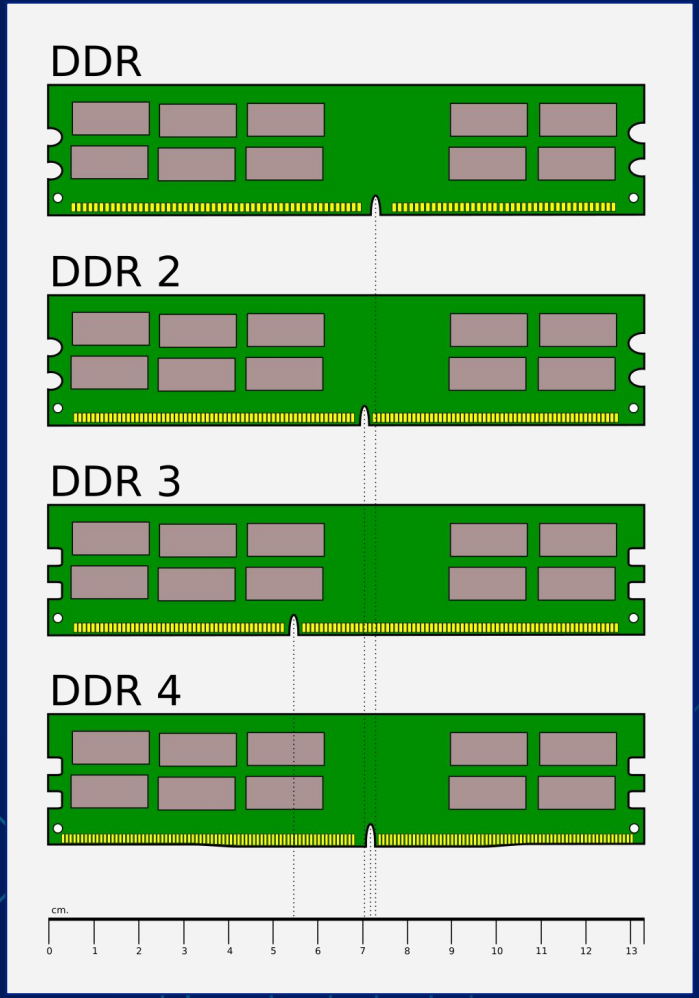
Memórias DDR4

- As matrizes de memória dos módulos DDR4 trabalham a um oitavo da frequência do barramento externo.
- São alimentadas com 1,2 volts e possuem latência inicial típica de 10 a 19 ciclos de relógio, bem maior que os demais tipos de memórias DDRs.
- Em compensação, as frequências de relógio, de até 2400 MHz, e as taxas de transferências estão entre as maiores atingidas, conferindo um grande desempenho a essas memórias.

DDR4



Comparação DDRs



Comparação DDRs

Fonte: https://en.wikipedia.org/wiki/DDR_SDRAM

Padrão	Relógio Célula (MHz)	Relógio Barramento (MHz)	Largura de Banda(MB/s)
DDR-200	100	100	1600
DDR-266	133	133	2133
DDR-333	166	166	2666
DDR-400	200	200	3200
DDR2-400	100	200	3200
DDR2-533	133	266	4266
DDR2-667	166	333	5333
DDR2-800	200	400	6400
DDR2-1066	266	533	8533
DDR3-800	100	400	6400
DDR3-1066	133	533	8533
DDR3-1333	166	666	1066
DDR3-1600	200	800	12800
DDR3-1866	233	933	14933
DDR3-2133	266	1066	17066
DDR4-1600	200	800	12800
DDR4-1866	233	933	14933
DDR4-2133	266	1066	17066
DDR4-2400	300	1200	19200
DDR4-2666	333	1333	2133
DDR4-2933	366	1466	23466
DDR4-3200	400	1600	25600

Correção de Erro

- Nos esquemas mais simples é adicionado um bit de paridade adicional para cada palavra armazenada.
- Assim, o bit de paridade assegura que o número total de bits em '1' na palavra armazenada seja par ou ímpar.
- O código de Hamming é capaz de detectar e corrigir erros simples (1 bit) e detectar erros duplos (2 bits) se um bit adicional de paridade for inserido para toda a palavra.

Código de Hamming

$$k + m + 1 \leq 2^m \qquad (4.1)$$

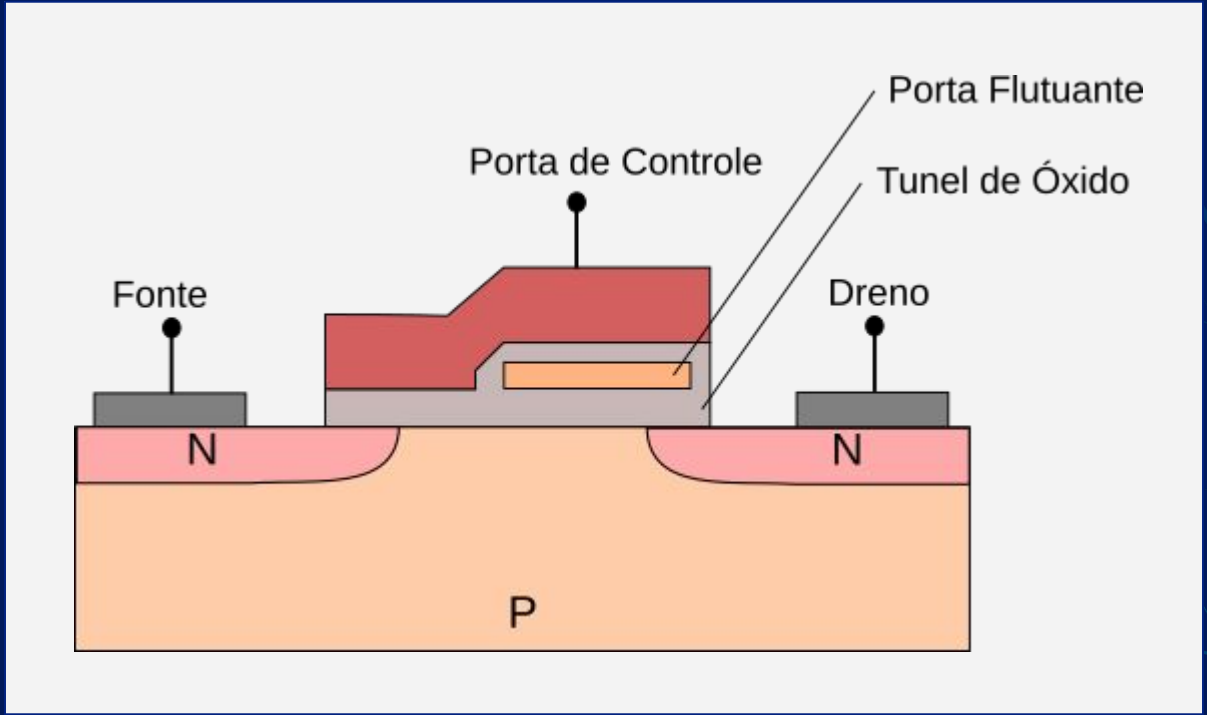
Fonte: https://en.wikipedia.org/wiki/Hamming_code

Bits Paridade	Bits de Dados	Total de bits
m	k	m + k
2	1	3
3	4	7
4	11	15
5	26	31
6	57	63
7	120	127
8	247	255

Memória Flash

- A memória Flash é uma memória de leitura e escrita que mantém o seu conteúdo mesmo sem alimentação.
- A memória Flash evoluiu das memórias EEPROM (Electrical Erasable PROM) e seu nome foi criado pela empresa Toshiba para expressar o quão rápido ela poderia ser apagada e re-escrita.
- A memória Flash é amplamente utilizada para armazenamento em módulos como dispositivos de estado sólido (SSD), pendrives e cartões de memória.

Memória Flash

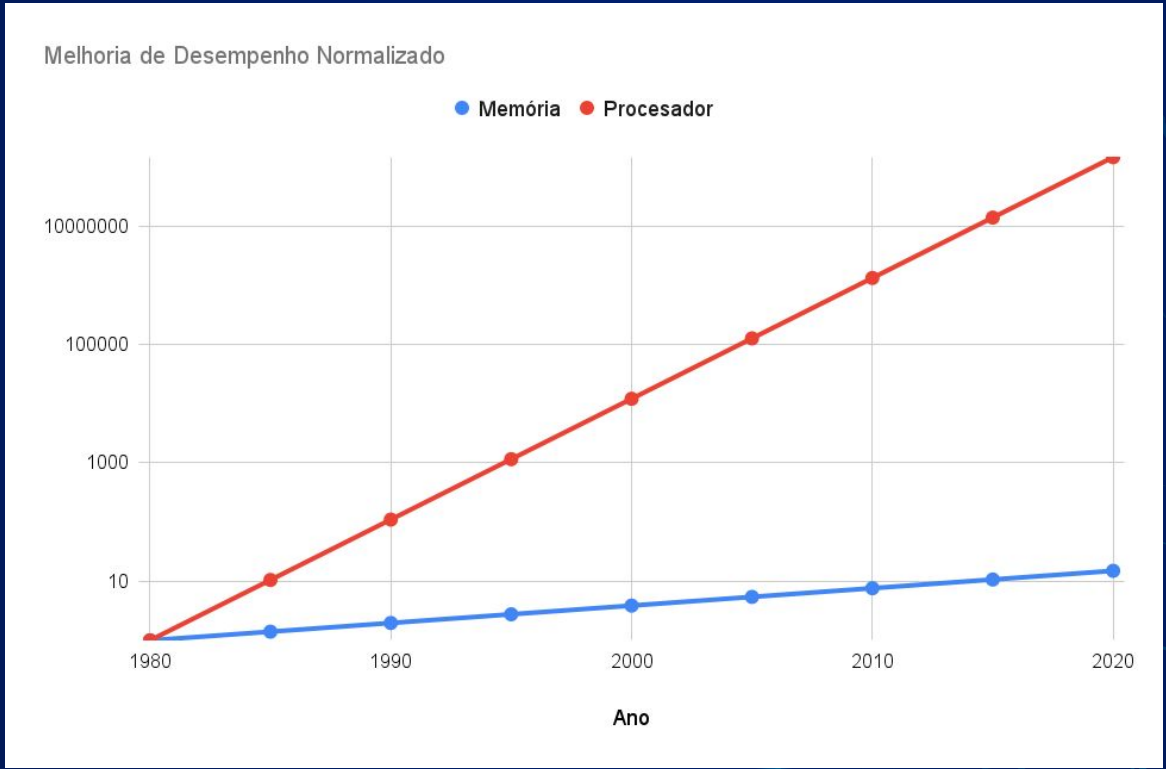


4.3 Hierarquia de Memória

Hierarquia de Memória

- As memórias principais dos modernos computadores, que são construídas com pastilhas de memórias dinâmicas têm um tempo de acesso e capacidade limitadas.
- Mais do que isso, a diferença de velocidade entre os processadores e o tempo de acesso à memória principal tem aumentado ao longo dos anos.
- Hoje em dia, a duração de um ciclo de relógio do processador é muito menor que o tempo de acesso à memória principal.

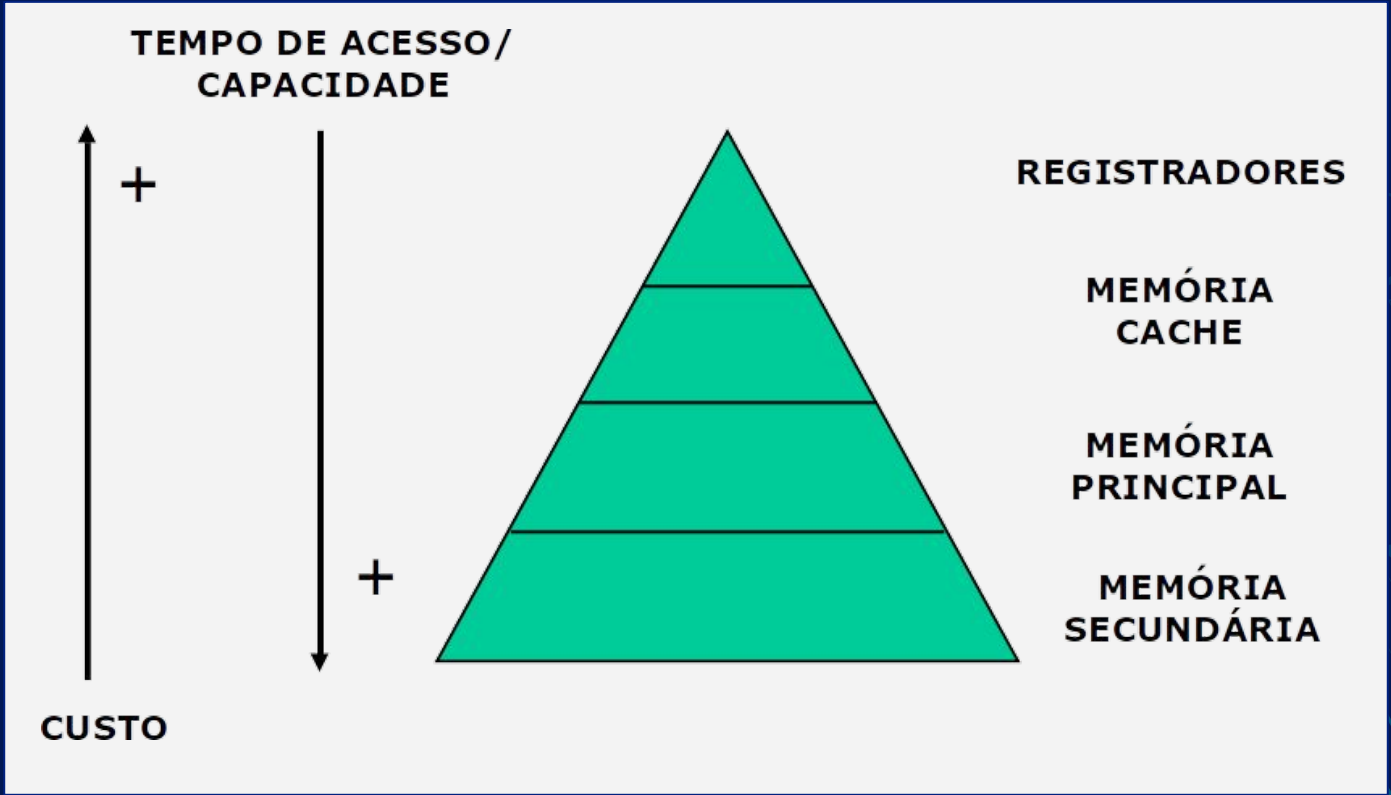
Gap Memória


Melhoria de Desempenho Normalizado
Memória
Procesador
10000000
100000
1000
10
1980
1990
2000
2010
2020
Ano

Hierarquia de Memória

- Os projetistas de computador propuseram o uso de uma hierarquia de memória, onde componentes de memória, de menor capacidade porém mais rápidos, são colocados junto ao processador para fornecer os dados mais necessários naquele momento.
- Esses dados são uma cópia parcial da informação que está nas hierarquias inferiores de memória, sendo que só a informação mais relevante é copiada dos níveis inferiores para os níveis mais altos.

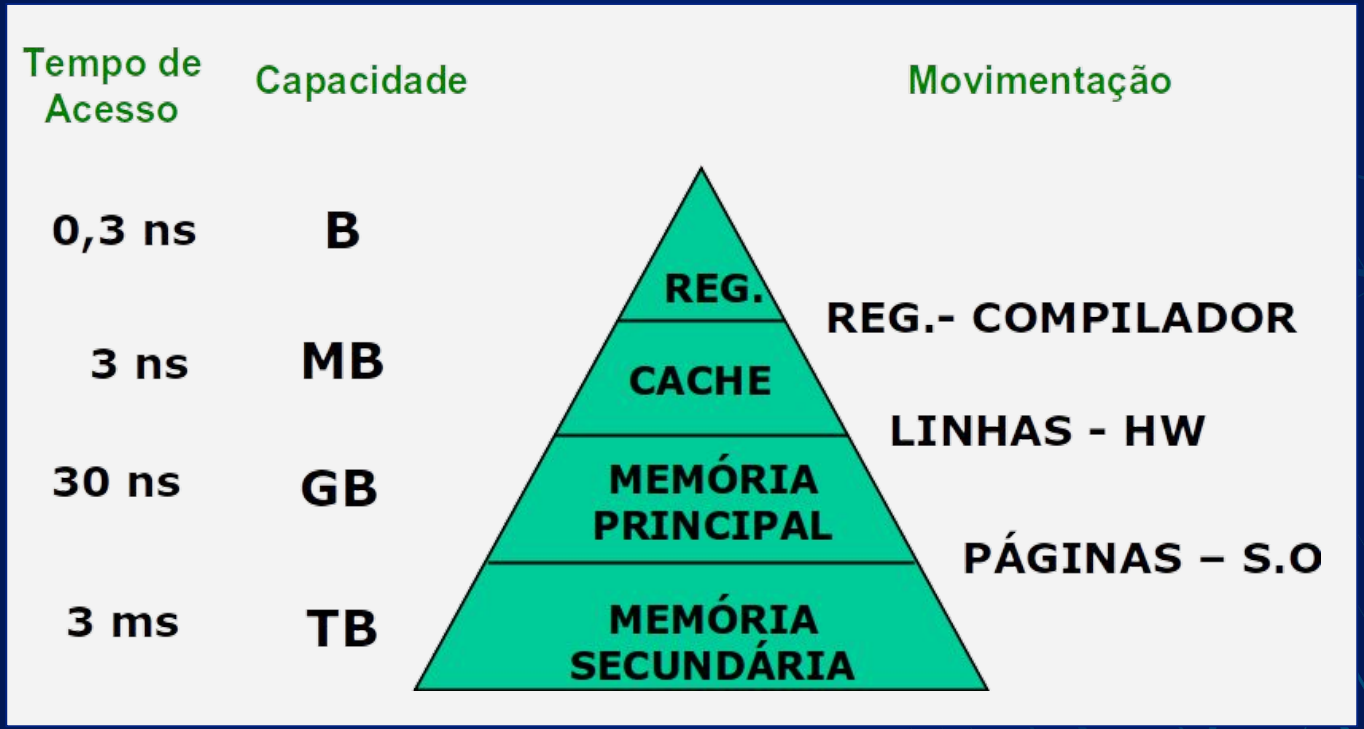
Hierarquia de Memória



Hierarquia de Memória

- Os elementos mais importantes nesta proposta de hierarquia de memória são os seguintes:
  - Registradores
  - Memória Cache
  - Memória Principal
  - Memória Secundária

Hierarquia de Memória



Conceito de Localidade

- Localidade temporal
  - Durante a execução de um programa, as posições de memória, relativas aos dados ou instruções, que são referenciadas (lidas ou escritas), tendem a ser novamente referenciadas em um curto intervalo de tempo.
- Localidade Espacial
  - Durante a execução de um programa, se uma posição de memória, relativa aos dados ou instruções, é referenciada, as posições de memória em endereços adjacentes tendem a ser referenciadas logo em sequência.

Taxa de Acerto

- A taxa de acerto (h) é definida como a relação entre o número de acertos e o número total de acessos em um determinado nível da hierarquia de memória.
- Ou seja, a probabilidade com que uma posição referenciada seja encontrada em determinado nível da hierarquia de memória. O total de acessos inclui tanto os acessos de leitura como os de escrita.

$$h = \frac{numero\ de\ acertos}{total\ de\ acessos} \qquad (4.3)$$

Tempo médio de acesso

$$T_{ma} = h \times T_a + (1 - h) \times T_f \quad (4.4)$$

Onde:

- T_{ma} = tempo médio de acesso
- h = taxa de acerto
- T_a = tempo de acesso quando há acerto
- T_f = tempo de acesso quando ocorre uma falha

4.4 Memória Cache

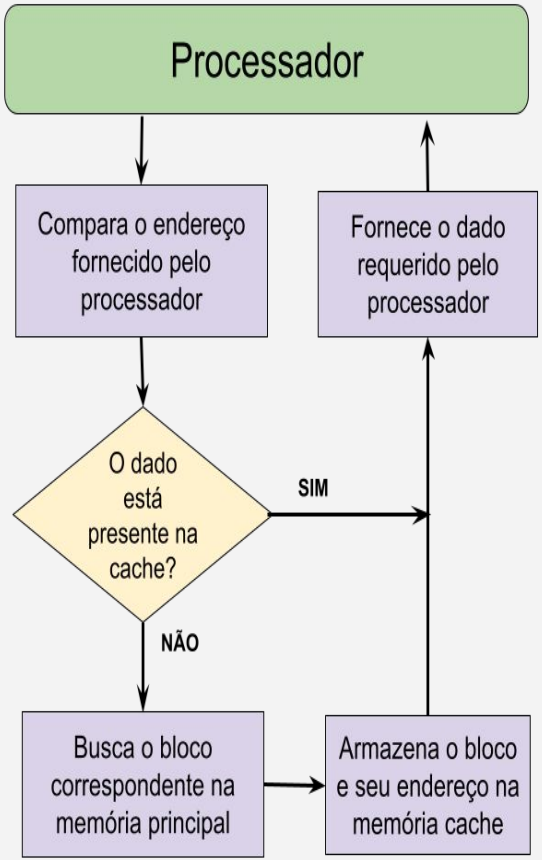
Memória Cache

- Uma cópia de parte do programa que está sendo executado é colocada em um dispositivo de memória mais rápido.
- O restante do programa, que não está sendo utilizado no momento, fica em uma memória mais lenta.
- A velocidade de acesso resultante é próxima à da memória mais rápida, mas com uma capacidade total de armazenamento igual à da memória lenta.

Funcionamento da Memória Cache

- As instruções e dados vão sendo então gradativamente copiados dos níveis inferiores, como a memória principal e secundária, para a memória cache, na medida que são utilizados pelo processador.
- Durante a busca da informação que está faltando na cache, no lugar de trazer somente a instrução ou dados solicitados pelo processador, um bloco inteiro é copiado da memória principal e armazenado em uma linha da cache, que então é marcada como válida.
- Junto com os dados propriamente ditos, é armazenado o endereço que o bloco ocupava na memória, para sua posterior identificação.

Funcionamento da Memória Cache



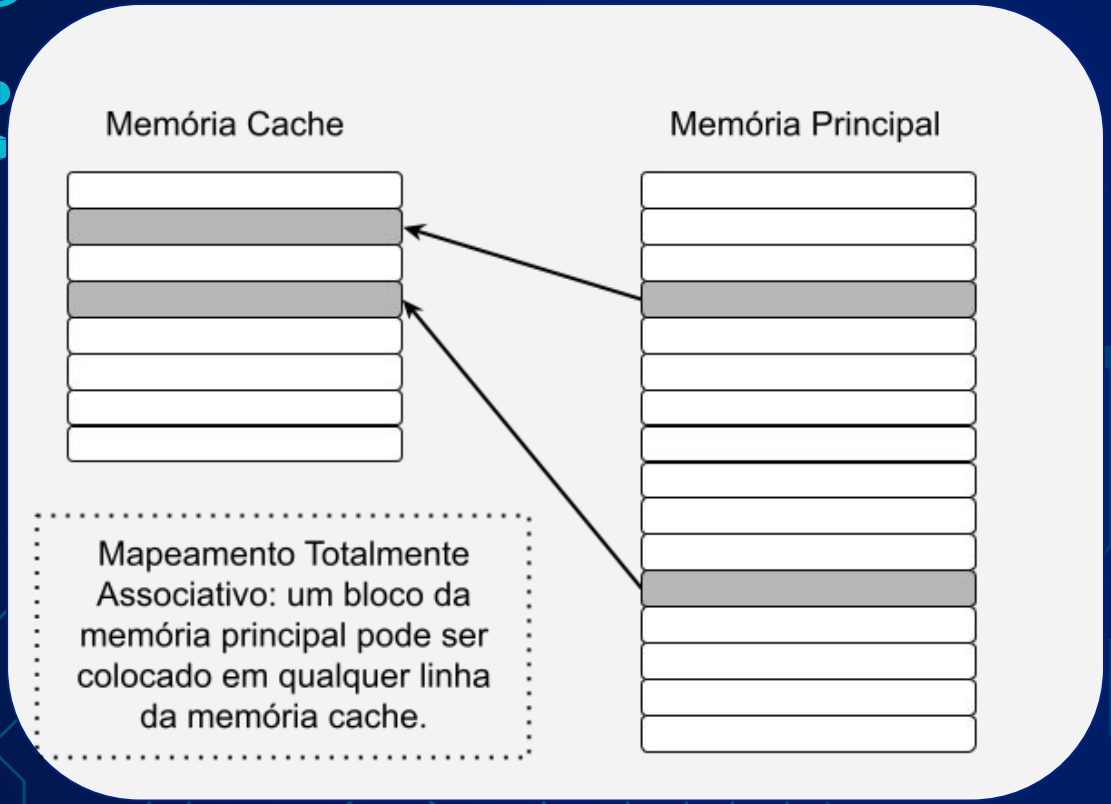
Mapeamentos Memória Cache

- Mapeamento completamente associativo: Um bloco da memória principal pode ser armazenado indistintamente em qualquer linha na memória cache.
- Mapeamento direto: Cada bloco da memória principal só pode ser armazenado em uma única linha na memória cache, segundo um critério definido pelo endereço do bloco na memória principal.
- Mapeamento associativo por conjunto: Cada bloco da memória principal pode ser armazenado indistintamente em um determinado grupo de linhas na memória cache, chamado de conjunto.

Mapeamento Totalmente Associativo

- Vantagens:
  - Máxima flexibilidade no posicionamento dos blocos.
  - Melhor aproveitamento da capacidade de armazenamento.
  - Maiores taxas de acerto.
- Desvantagens
  - O custo em hardware para a comparação simultânea dos endereços é alto;
  - Se a comparação for seqüencial, o tempo gasto torna proibitivo o uso desta opção
  - Quando a cache está cheia, o custo do algoritmo de substituição é significativo.

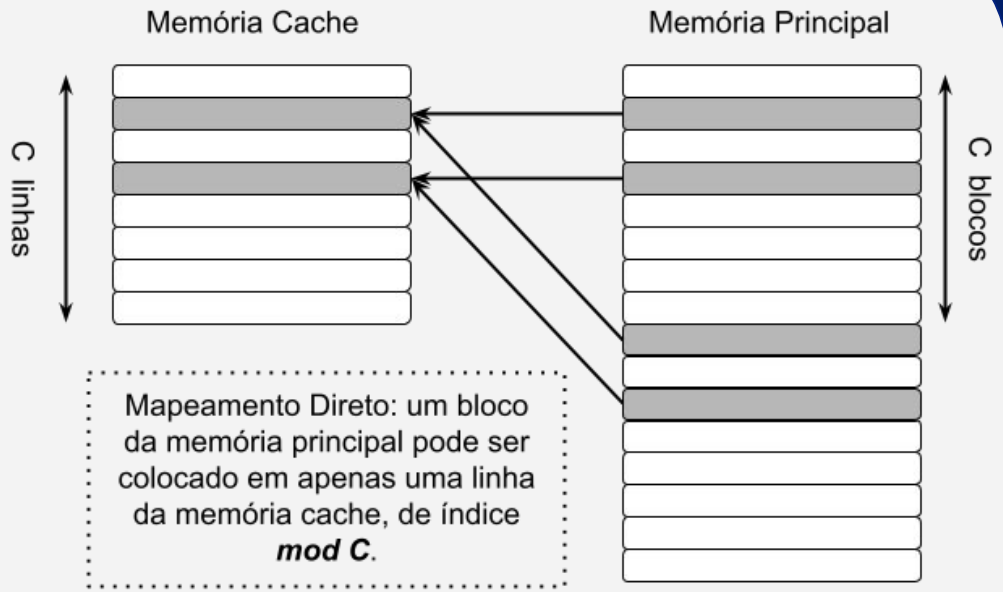
Totalmente Associativa



Mapeamento Direto

- Vantagens:
  - Não há necessidade de algoritmo de substituição.
  - O hardware é simples e de baixo custo.
  - Alta velocidade de operação.
- Desvantagens
  - O desempenho diminui se acessos consecutivos são feitos em palavras com mesmo índice.
  - A taxa de acerto é inferior ao de memórias caches com mapeamento associativo.
  - Contudo, a taxa de acerto pode ser melhorada com o aumento do tamanho da cache.

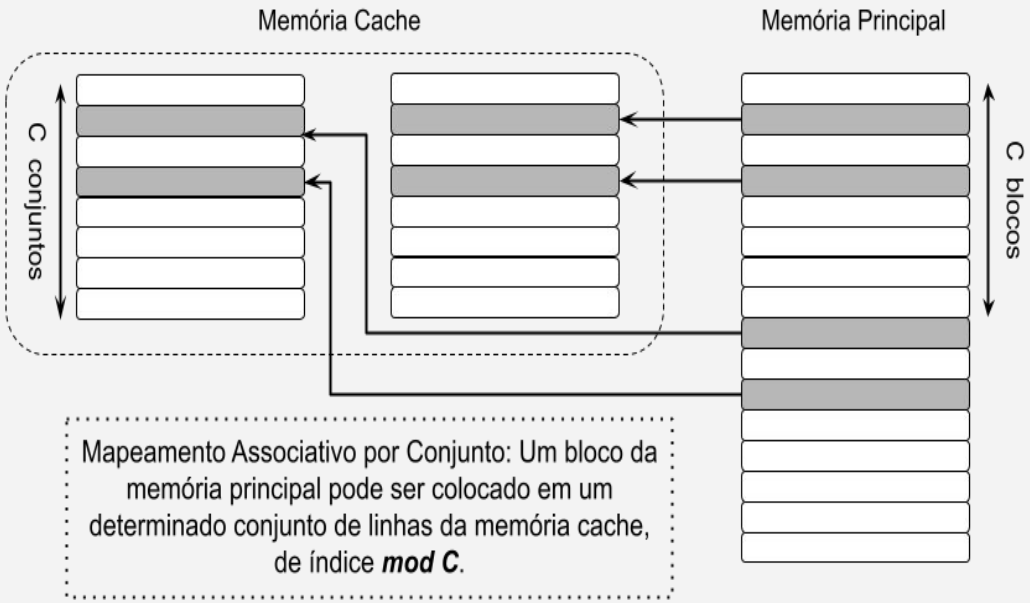
Mapeamento Direto



Mapeamento Associativo por Conjunto

- Vantagens:
  - Reduz as chances de conflito.
  - É rápido para determinar se um bloco está na cache.
  - As taxas de acerto são altas.
- Desvantagens
  - Necessita de algoritmo de substituição implementado em hardware.
  - Possui maior complexidade de hardware que o mapeamento direto.

Associativo por Conjunto



Tamanho do Bloco ou Linha

- Tamanhos de bloco maiores tendem a fazer uso melhor da localidade espacial, aumentando a taxa de acerto.
- Blocos de maior tamanho nas memórias caches tendem a aumentar a penalidade por falha, ou seja, aumentam o tempo necessário para trazer este bloco da memória principal para a memória cache.
- Uma solução de compromisso é encontrada após exaustivas simulações na fase de projeto dos processadores.
- Tamanhos típicos de blocos ou linhas para as memórias caches estão entre 16 e 128 bytes.

Tamanho da Linha ou Bloco

- Um processador A com uma cache com linhas de 32 bytes e um outro processador B com uma cache com linhas de 128 bytes, mas ambas com mapeamento direto e capacidade igual a 64 Kibytes.
- A taxa de acerto observada da cache do processador A é de 95% e do processador B é 98%.
- O tempo de acerto é de 5 ns em ambos os casos, mas que o tempo de falha da cache A é de 50 ns e da cache B é 110 ns, em função do maior tempo necessário para a busca de um bloco maior na memória.
- Qual das caches é a mais eficiente?

Tamanho da Linha ou Bloco

$$T_{ma} = h \times T_a + (1 - h) \times T_f \quad (4.6)$$

Vamos resolver essa equação para cada um dos casos:

$$T_{ma}A = 0,95 \times 5 + (0,05) \times 50 = 7,25ns \quad (4.7)$$

$$T_{ma}B = 0,98 \times 5 + (0,02) \times 110 = 7,1ns \quad (4.8)$$

Escrita com Acerto

- **Write-through**: a escrita é realizada simultaneamente na cache e na memória principal. Neste caso as operações de escrita são mais lentas que as de leitura, e a escrita adicional na memória principal reduz tempo médio de acesso à memória cache.

- **Write-back**: o bloco é atualizado na memória cache imediatamente. A memória principal é atualizada apenas quando o bloco modificado for substituído na memória cache.

Escrita com Falha

- **Write allocate**: O bloco onde vai ser feita a operação de escrita é trazido primeiramente para a memória cache e a operação de escrita é então realizada.

- **No write allocate**: O bloco a ser escrito não é trazido para a memória cache e, portanto, a operação de escrita sempre se realiza apenas na memória principal.

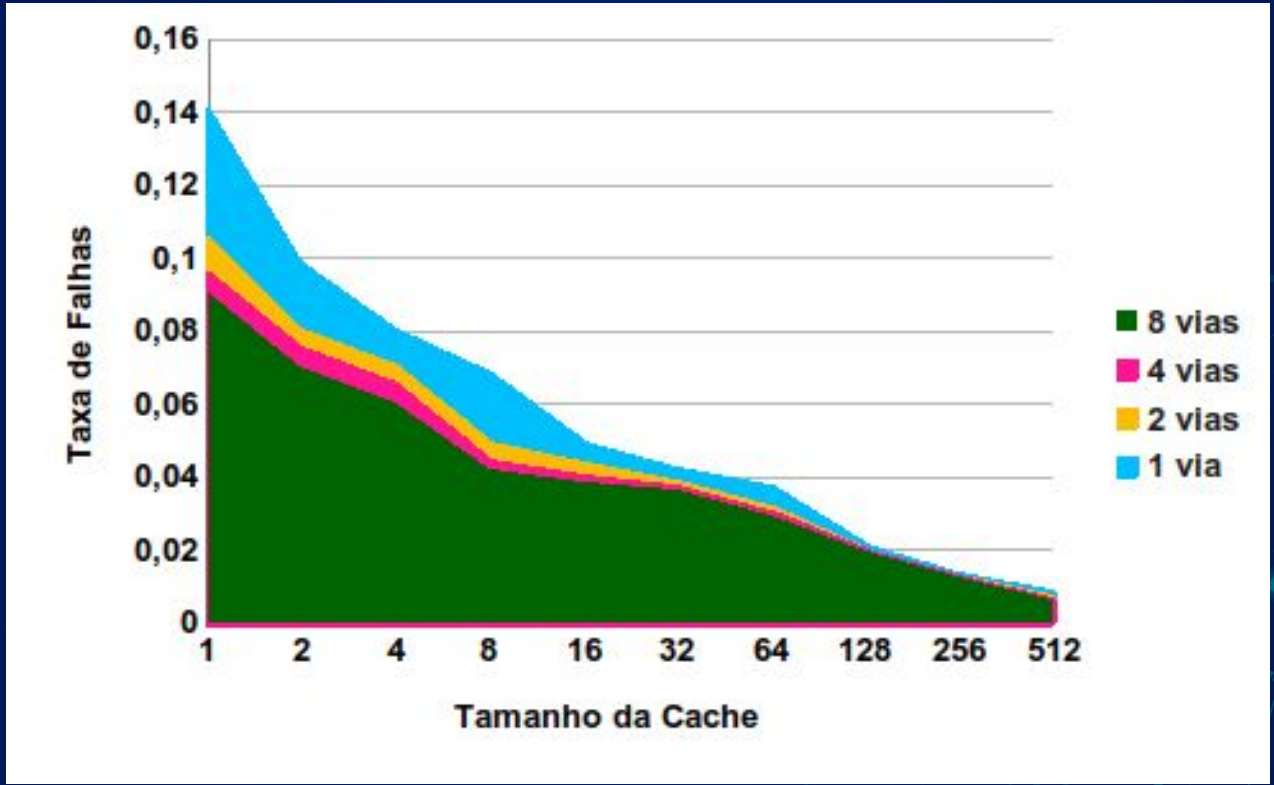
Algoritmos de Substituição

- Aleatório (randômico)}: Um bloco é escolhido aleatoriamente para ser substituído no conjunto. Fácil implementação, mas diminui a taxa de acerto.
- FIFO (first-in first-out --- primeiro a entrar, primeiro a sair)}: O bloco que está há mais tempo no conjunto é removido. Menos simples e pode diminuir a taxa de acerto.
- LRU (least recently used --- menos recentemente utilizado): O bloco a ser substituído no conjunto é aquele que não é referenciado (lido ou escrito) há mais tempo. É o esquema de melhor desempenho, mas cuja implementação é a mais complexa.

O três Cs

- Falhas compulsórias: São falhas que ocorrem sempre no primeiro acesso a um bloco, já que o bloco nunca foi armazenado na memória cache.
- Falhas devido à capacidade: São falhas que ocorrem quando a cache não consegue armazenar todos os blocos necessários à execução de um programa, por não ter capacidade suficiente.
- Falhas por conflitos ou colisão: São falhas que ocorrem quando diversos blocos competem pela mesma posição (conjunto) na memória cache. Esse tipo de falha não ocorre em caches totalmente associativas.

Três Cs


0,16
0,14
0,12
0,1
0,08
0,06
0,04
0,02
0
Taxa de Falhas
1
2
4
8
16
32
64
128
256
512
Tamanho da Cache
8 vias
4 vias
2 vias
1 via

Caches Virtuais e Físicas

- Os endereços devem ser traduzidos, em tempo de execução, para endereços físicos, que refletem a real posição que esses programas ocupam na memória principal.
- Essa tradução é feita pela gerência de memória
- Em algumas situações, quando o processador é muito rápido em relação à memória cache, pode ser que não haja tempo para a tradução dos endereços virtuais em endereços físicos e ainda verificar se o bloco está presente ou não na memória cache.

Caches Virtuais

- Armazenar na cache os endereços virtuais relativos aos rótulos dos blocos, antes da tradução pela gerência de memória.
- Dificuldades:
  - Possibilidade de ocorrência de aliasing entre dois processos.
  - É necessário que seja realizado um descarte de todos os dados armazenados na cache, a cada troca de processo.
  - A manutenção da coerência de dados da memória cache virtual é muito difícil de ser realizada em ambientes com múltiplos processadores, já que isso normalmente é feito com o uso de endereços físicos.

Caches Separadas para Dados e Instruções

- A política de escrita só precisa ser aplicada à cache de dados.
- Existem caminhos separados entre memória principal e cada cache, permitindo transferências simultâneas de dados e instruções, quando o processador possui um pipeline.
- Estratégias diferentes podem ser utilizadas para cada tipo de memória cache, como por exemplo, capacidades diferentes, tamanho do bloco e associatividade.

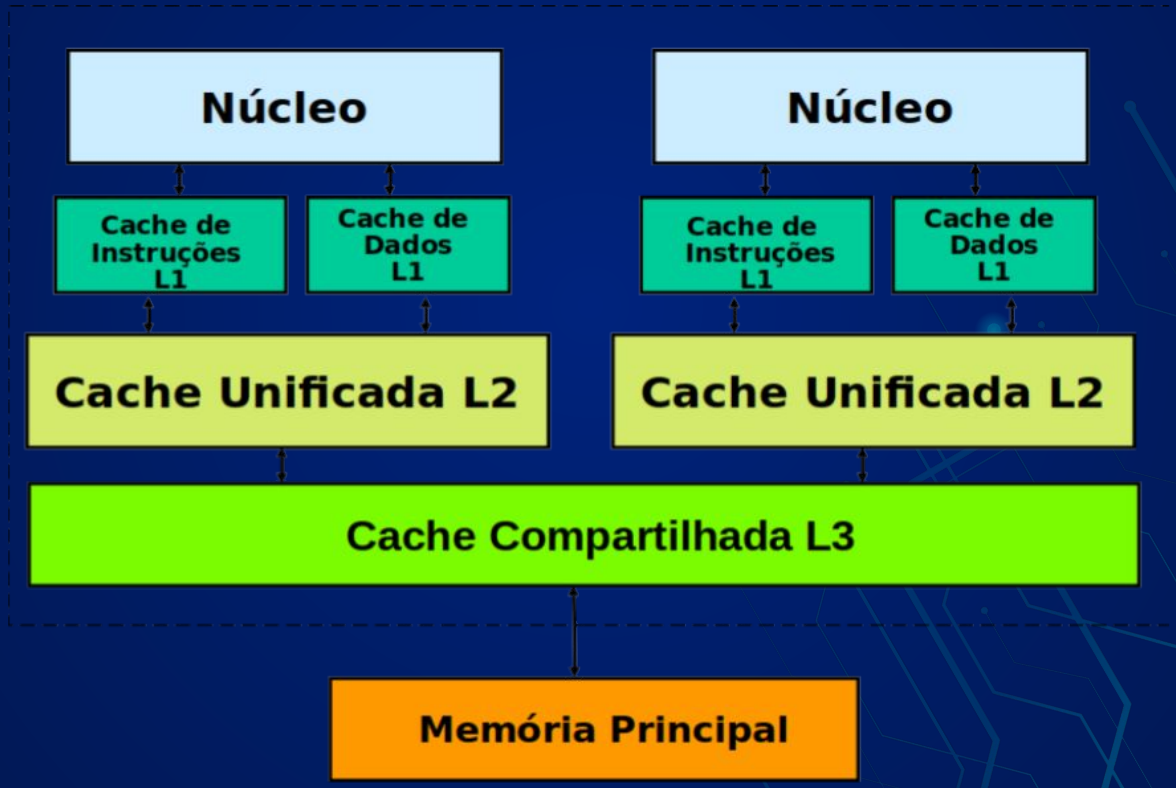
Caches Multinível

- A implementação de uma cache pequena e muita rápida junto ao processador é uma implementação muito utilizada para que os dados e instruções sejam fornecidos em apenas um ciclo de relógio.
- Essa cache recebe o nome de cache de nível 1 (L1). Uma outra cache, chamada de nível 2 (L2), é inserida entre a memória principal e a cache de nível 1.
- Essa cache não é tão rápida quando a cache L1, mas possui maior capacidade e uma taxa de acerto boa o suficiente para reduzir o tempo de acesso aos dados, quando houver uma falha na cache de nível 1.

Caches Multinível

- A implementação de uma cache pequena e muita rápida junto ao processador é uma implementação muito utilizada para que os dados e instruções sejam fornecidos em apenas um ciclo de relógio.
- Essa cache recebe o nome de cache de nível 1 (L1). Uma outra cache, chamada de nível 2 (L2), é inserida entre a memória principal e a cache de nível 1.
- Essa cache não é tão rápida quando a cache L1, mas possui maior capacidade e uma taxa de acerto boa o suficiente para reduzir o tempo de acesso aos dados, quando houver uma falha na cache de nível 1.

Cache Multinível



Caches Multinível

- Nos processadores modernos, com mais de um núcleo, é comum a existência de um terceiro nível de cache (L3), compartilhada por todos os núcleos.
- Esse nível é utilizado para comunicação entre as threads ou processos executando nos diversos núcleos, evitando a ida, muito mais demorada, à memória principal.
- O tempo médio de acesso é menor com uso de uma cache multinível, como pode ser visto na equação para o tempo médio de acesso em uma cache de dois níveis unificados.

Caches Multinível

$$T_{ma} = h_1 \times T_a^1 + (1 - h_1) \times (h_2 \times T_a^2 + (1 - h_2) \times T_f) \quad (4.9)$$

- T_{ma} = tempo médio de acesso
- h_1 = taxa de acerto da cache de nível 1 (L1)
- T_a^1 = tempo de acesso à cache de nível 1 (L1) quando há acerto
- h_2 = taxa de acerto da cache de nível 2 (L2)
- T_a^2 = tempo de acesso à cache de nível 2 (L2) quando há acerto
- T_f = tempo de acesso com falha

Caches Multinível

Hierarquia de Memória				
Processador	Intel Core i7-7820X	Intel Core i9-12900KS	AMD Ryzen 9 5950X	AMD Ryzen 7 5800X
Frequência máxima	4,5 GHz	5,5 GHz	4,9 GHz	4,7 GHz
L1 (Dados)	32 KiB	32 KiB	64 KiB	32 KiB
L1 (Instruções)	32 KiB	32 KiB	64 KiB	32 KiB
L2	1 MiB	256 KiB	512 KiB	256 KiB
L3	11 MiB	30 MiB	64 MiB	32 MiB
Associatividade	16-way	16-way	16-way	8-way

8.5 Memória Virtual

Memória Virtual

- A memória virtual realiza três funções principais:
  - Controle da hierarquia entre a memória principal e a memória secundária.
  - Proteção, evitando que um programa modifique informações que pertençam a algum outro.
  - Mapeamento dos endereços de um espaço de endereçamento virtual em endereços físicos.

Memória Virtual

- O método mais usual é a divisão do espaço de endereçamento do programa em blocos de tamanho fixo.
- Esses blocos são chamados de páginas no espaço de endereçamento virtual e de quadros (frames) na memória principal.
- Tamanhos usuais para as páginas em diversos sistemas operacionais se situam entre 1 Kibytes e 8 Kibytes, mas podendo chegar até 64 Kibytes.

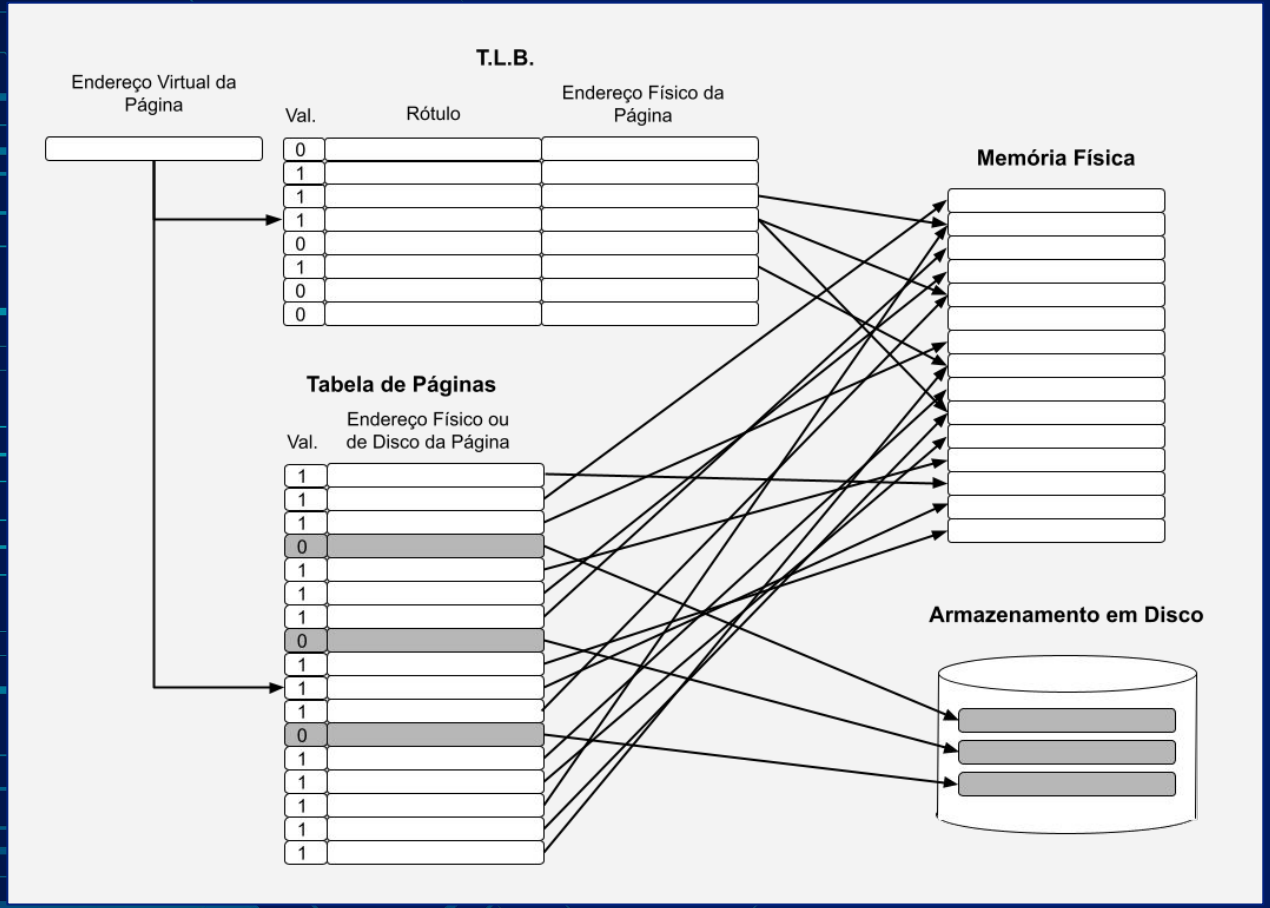
Memória Virtual

- Cada página virtual do processo possui um descritor que contém:
  - Um bit de validade indicando que o descritor tem informação válida.
  - Um bit para indicar se a página foi modificada
  - O endereço físico atual da página, que pode ser um endereço na memória principal ou a sua localização na memória secundária.
- O conjunto de descritores de um processo se chama tabela de páginas e fica armazenada em uma estrutura hierárquica na memória principal.

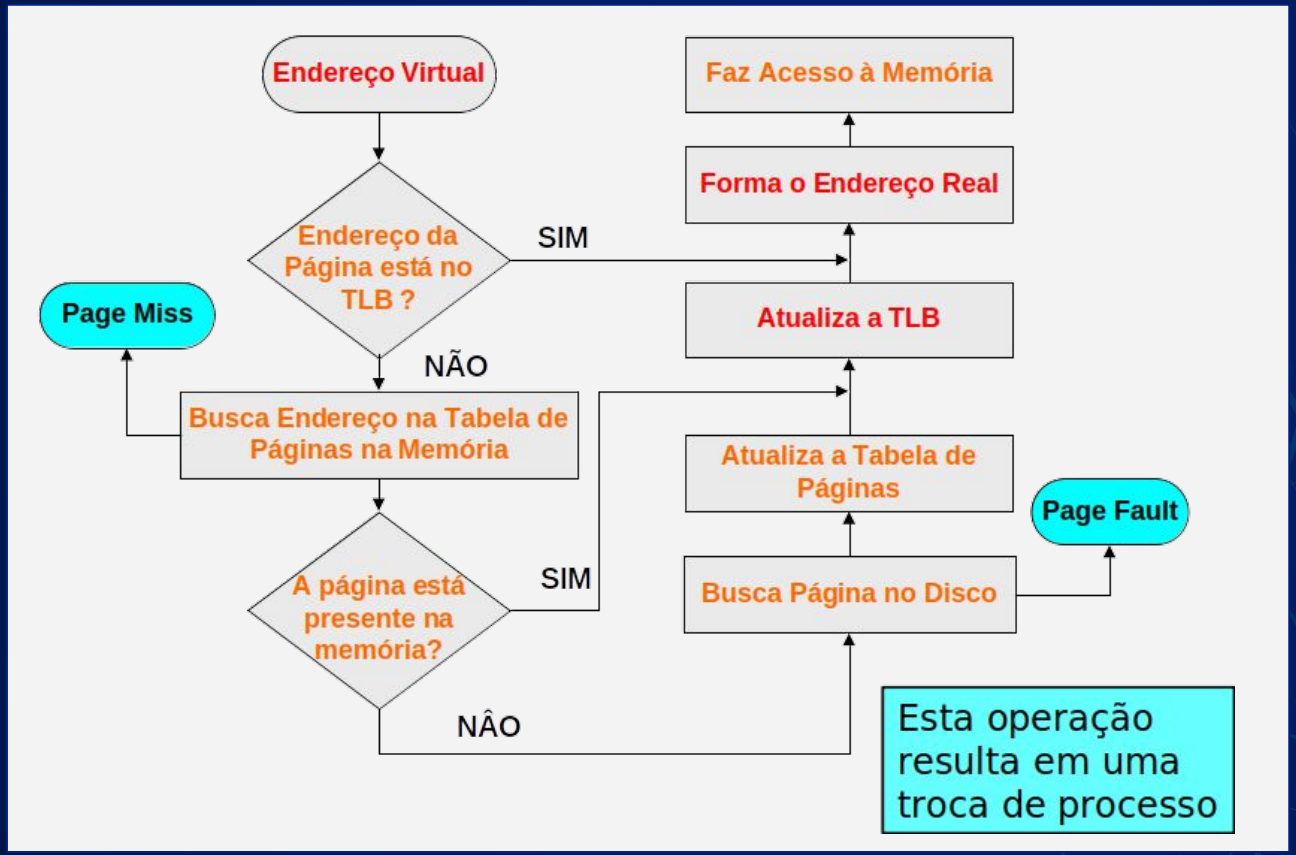
Memória Virtual

- O endereço físico (ou endereço do quadro) é buscado no descritor na tabela de páginas e armazenado em uma pequena memória associativa junto do processador.
- Deste modo, a tabela é consultada e, em caso de acerto, o endereço armazenado é utilizado no acesso à memória, dispensando a ida à tabela de páginas na memória principal.
- Essa tabela recebe o nome abreviado de TLB (Translation Lookaside Buffer) e armazena algumas dezenas de descritores.

Memória Virtual



Funcionamento Memória Virtual



Memória Virtual

Páginas e TLBs			
Processador	Tam. de Página	TLB Instruções	TLB Dados
Alpha 21024	8 KiB	8 entradas (CA[a])	32 entradas (CA)
Alpha 21124	8 KiB	48 entradas (CA)	64 entradas (CA)
Alpha 21224	8 KiB	64 entradas (CA)	128 entradas (CA)
Intel Pentium	4 KiB	32 entradas (4 vias)	64 entradas (4 vias)
intel Pentium II	4 KiB	32 entradas (4 vias)	64 entradas (4 vias)
Intel Pentium IV	4 KiB	64 entradas (4 vias)	128 entradas (4 vias)
Intel Core Duo	4 KiB	64 entradas (CA)	64 entradas (CA)
Intel i3 Sandy Bridge	4 KiB	64 entradas/thread (4 vias)	64 entradas (4 vias)

Onde Colocar uma Página?

- O método utilizado pela memória virtual é sempre o mapeamento totalmente associativo.
- Ou seja, as páginas de um processo podem ocupar qualquer posição na memória física que não esteja sendo utilizada pelo sistema operacional.

Qual o Tamanho de uma Página?

- Tamanhos de páginas maiores tendem a fazer uso melhor da localidade espacial, aumentando a taxa de acerto.
- Uso de páginas maiores tende a diminuir o tamanho da tabela de páginas, diminuindo a penalidade por falha das TLBs.
- Assim, tamanhos típicos de páginas estão entre 1 e 16 Kibytes, sendo 4 Kibytes o tamanho padrão para a maioria dos sistemas operacionais.
- Contundo, alguns sistemas operacionais tem previsão para páginas gigantes (Huge Pages) de vários Mibytes.

Qual Página deve ser Substituída?

- Algoritmos semelhantes aos utilizados na memória cache são empregados para decidir qual a página que deve ser removida da memória: LRU, FIFO, Aleatório e do Relógio são os principais.
- Um bit de modificado no descritor na tabela de páginas indica se a página a ser removida deve ser atualizada antes na memória secundária.
- Nos modernos sistemas operacionais, para cada processo, apenas um subconjunto de suas páginas, que é denominado de ``conjunto de trabalho'', é mantido em memória.

O que Acontece na Escrita?

- O algoritmo utilizado para atualização das páginas na memória secundária é sempre o write-back com write-allocate.
- As páginas só são atualizadas na memória secundária se tiverem sido modificadas, em caso contrário são descartadas, pois sempre haverá uma cópia idêntica no arquivo original.
- No caso de uma escrita sem acerto, a página é trazida para a memória principal para então ser feita a escrita dos dados correspondentes.

O que acontece na escrita?

- A área de armazenamento para as páginas modificadas na memória secundária recebe o nome de espaço de troca (swap), podendo ser uma partição e/ou um arquivo no disco rígido (HDD) ou no dispositivo de estado sólido (SSD).
- A quantidade recomendada de espaço de troca aumentava linearmente com a quantidade de memória principal no sistema.
- Grandes quantidades de memória swap não são efetivas para garantir o funcionamento adequado do sistema nos modernos computadores.

O que acontece na escrita?

- A área de armazenamento para as páginas modificadas na memória secundária recebe o nome de espaço de troca (swap), podendo ser uma partição e/ou um arquivo no disco rígido (HDD) ou no dispositivo de estado sólido (SSD).
- A quantidade recomendada de espaço de troca aumentava linearmente com a quantidade de memória principal no sistema.
- Grandes quantidades de memória swap não são efetivas para garantir o funcionamento adequado do sistema nos modernos computadores.

Tamanho do Swap

Espaço de Troca - CentOS 7		
Quantidade de memória principal	Espaço de troca recomendado	Espaço de troca recomendado com hibernação
menor que 2 GiB	2 vezes a quantidade de memória principal	3 vezes a quantidade memória principal
entre 2 GiB e 8 GiB	Igual à quantidade de memória principal	2 vezes a quantidade de memória principal
entre 8 GiB e 64 GiB	No mínimo 4 GiB	1,5 vezes a quantidade de memória principal
maior que 64 GiB	No mínimo 4 GiB	A Hibernação não é recomendada

ZRAM

- Nos sistemas operacionais mais modernos, a área de swap na memória secundária foi substituída por uma partição comprimida na própria memória principal do computador, chamada de ZRAM.
- A ZRAM cria um dispositivo de bloco na memória onde as páginas são primeiro compactadas e depois armazenadas neste dispositivo de bloco.
- Isso permite uma operação de swap muito mais rápida e também a compactação de dados fornece uma quantidade significativa de economia de memória.

ZRAM

- Uma desvantagem da ZRAM é que ela requer o uso de processamento para compactação dos dados, o que pode causar atrasos em sistemas com poucos núcleos de processamento.
- Geralmente é compensado pelos ganhos obtidos ao evitar a realização swap para a memória secundária e a economia geral de memória com a compactação.

Obrigado !

Arquitetura e Organização de Computadores: Uma Introdução

Mais recursos em:
https://simulador-simus.github.io