Arquitetura e Organização de Computadores

Guilherme Henrique de Souza Nakahata

Universidade Estadual do Paraná - Unespar

25 de Abril de 2023

Memória Cache

Localização	Desempenho
Interna (por exemplo, registradores do processador, memória principal, cache)	Tempo de acesso
Externa (por exemplo, discos ópticos, discos magnéticos, fitas)	Tempo de ciclo
	Taxa de transferência
Método de acesso	**Tipo físico**
Sequencial	Semicondutor
Direto	Magnético
Aleatório	Óptico
Associativo	Magneto-óptico
Unidade de transferência	**Características físicas**
Palavra	Volátil/não volátil
Bloco	Apagável/não apagável
Capacidade	**Organização**
Número de palavras	Módulos de memória
Número de bytes	

Conceitos

- O termo localização indica se a memória é interna ou externa ao computador.
- Uma característica importante da memória é a sua capacidade.
- Um conceito relacionado é a unidade de transferência.
- Para a memória interna, a unidade de transferência é igual ao número de linhas elétricas que chegam e que saem do módulo de memória.

Método de acesso

- **4.1) Quais são as diferenças entre acesso sequencial, acesso direto e acesso aleatório?**
- Os **métodos de acesso** das unidades de dados **inclui**:
  - **Acesso sequencial**
  - **Acesso direto**
  - **Acesso aleatório**
  - **Associativo**

Método de acesso

- Os **métodos de acesso** das unidades de dados **inclui**:
  - **Acesso sequencial:** a memória é organizada em unidades de dados chamadas registros. Sendo acessado de forma linear.
  - **Acesso direto:** envolve um mecanismo compartilhado de leitura-escrita. Seus blocos ou registros contém endereço único, baseado no local físico. Acesso é feito por meio de um acesso direto a uma vizinhança genérica do registo, em seguida, por uma pesquisa sequencial.
  - **Acesso aleatório:** cada local endereçável na memória tem um mecanismo de endereçamento exclusivo, fisicamente interligado. Qualquer posição pode ser selecionada de modo aleatório, sendo endereçada e acessada diretamente.
  - **Associativo:** permite fazer uma comparação de um certo número de bits com uma combinação específica. Uma palavra é buscada na memória com base em uma parte de seu conteúdo, e não de acordo com seu endereço.

Memória cache

- As duas características mais importantes da memória são **capacidade** e **desempenho**.
- Três parâmetros de desempenho são usados:
  - **Tempo de acesso (latência)**;
  - **Tempo de ciclo de memória**;
  - **Taxa de transferência**;
- Variedade de tipos físicos da memória:
  - Memória semicondutora;
  - Memória de superfície magnética (Disco e Fita);
  - Óptica e magneto-óptica;

Memória cache

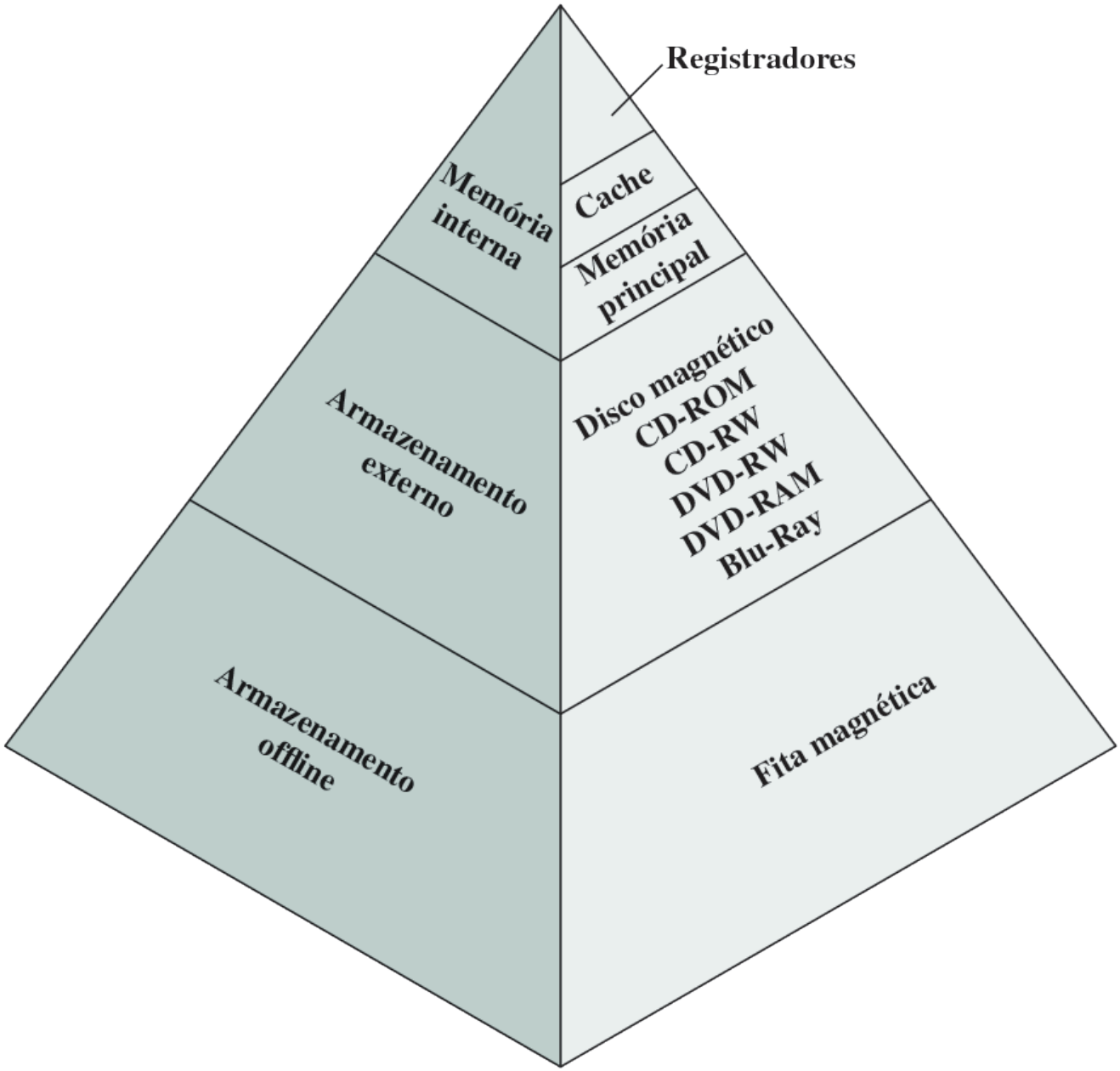
- As restrições de projeto podem ser resumidas por três questões:
  - **4.2) Qual é o relacionamento geral entre tempo de acesso, custo de memória e capacidade?**
  - **Quanto?**
  - **Com que velocidade?**
  - **A que custo?**
- Para conseguir maior desempenho, a memória deve ser capaz de acompanhar a velocidade do processador.
- O custo da memória deve ser razoável em relação a outros componentes.

Memória cache

- Diversas tecnologias são usadas para implementar sistemas de memória, seguinte relações:
  - Tempo de acesso mais rápido, maior custo por bit.
  - Maior capacidade, menor custo por bit.
  - Maior capacidade, tempo de acesso mais lento.
- **Dilema**.

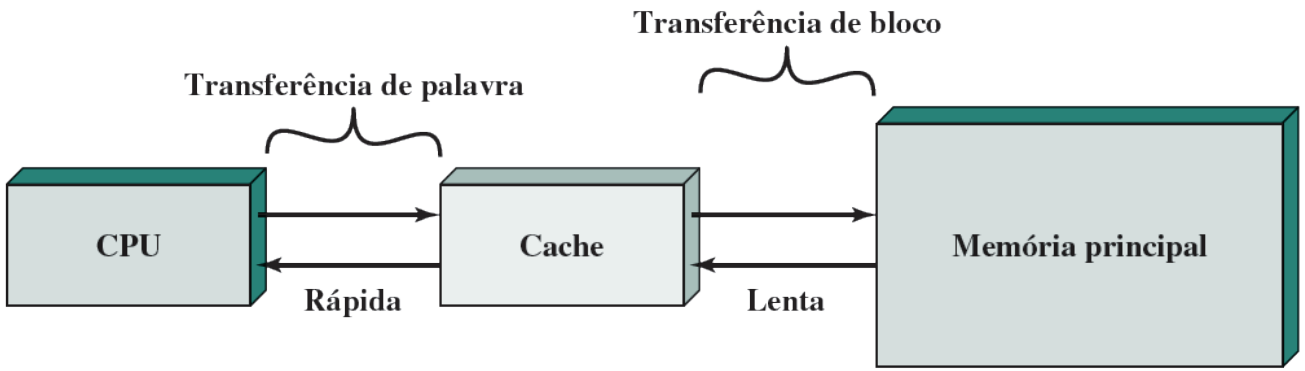
- **4.3) Como o princípio de localidade se relaciona com o uso de múltiplos níveis de memória?**
- O princípio de localidade se baseia na observação de que os programas tendem a acessar repetidamente um conjunto de locais na memória em um curto período de tempo.
- Conforme se desce na hierarquia, ocorre o seguinte:
  - Diminuição do custo por bit.
  - Aumento da capacidade.
  - Aumento do tempo de acesso.
  - Diminuição da frequência de acesso à memória pelo processador.

Memória cache

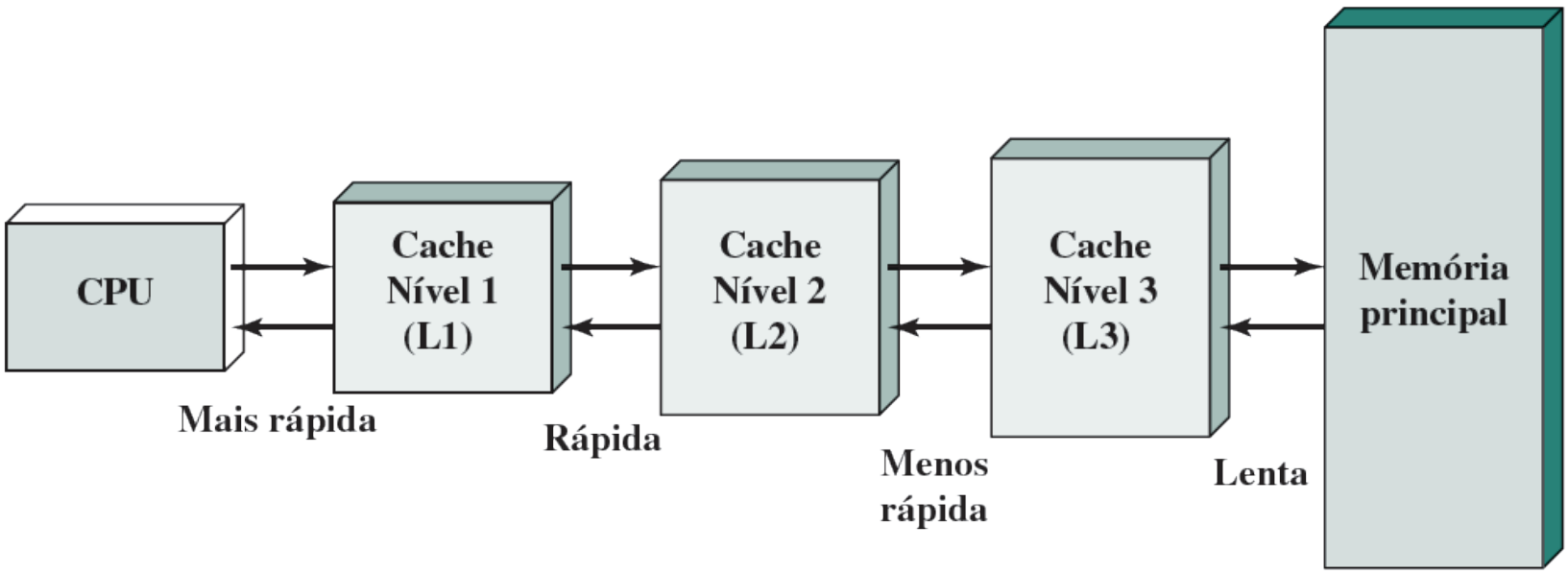


Memória cache

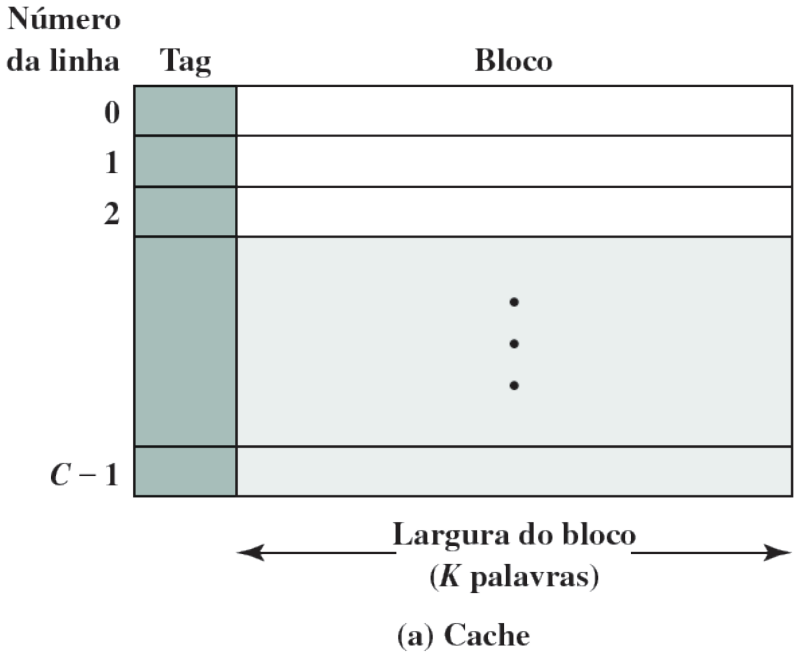
- A memória cache é desenvolvida para combinar o tempo de acesso de memórias de alto custo e alta velocidade com as memórias de menor velocidade, maior tamanho e mais baixo custo.



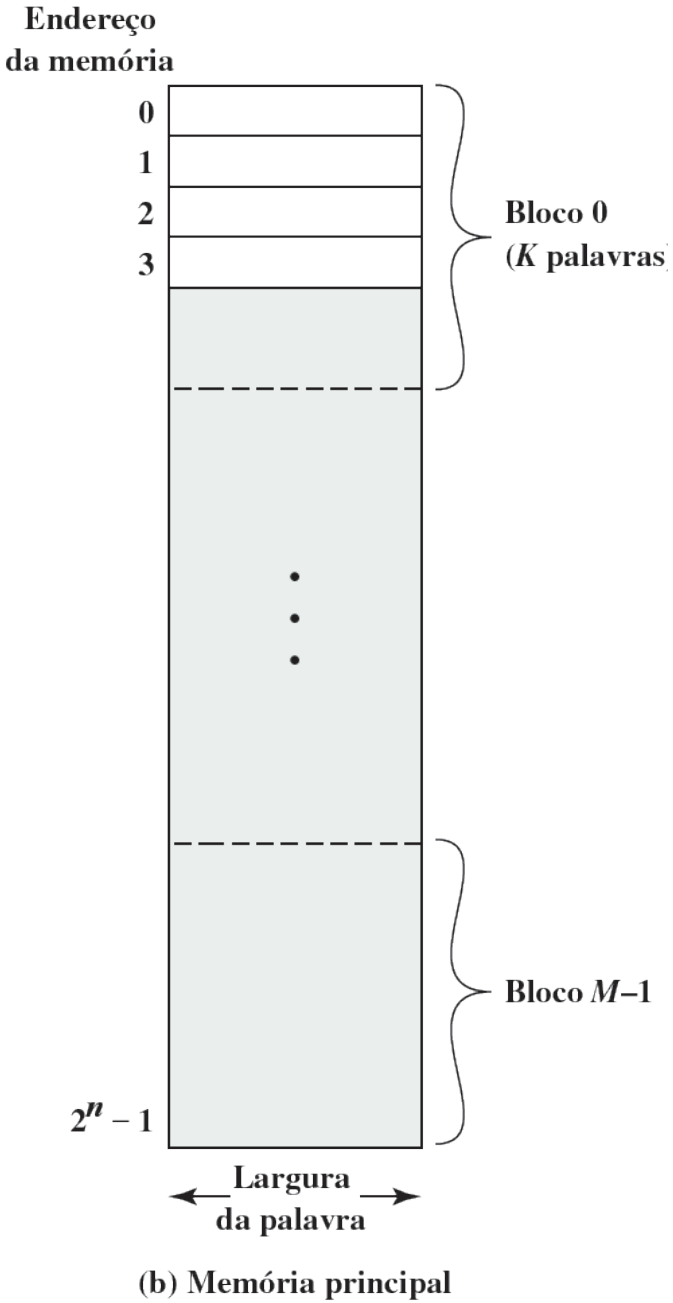
Memória cache



Memória cache



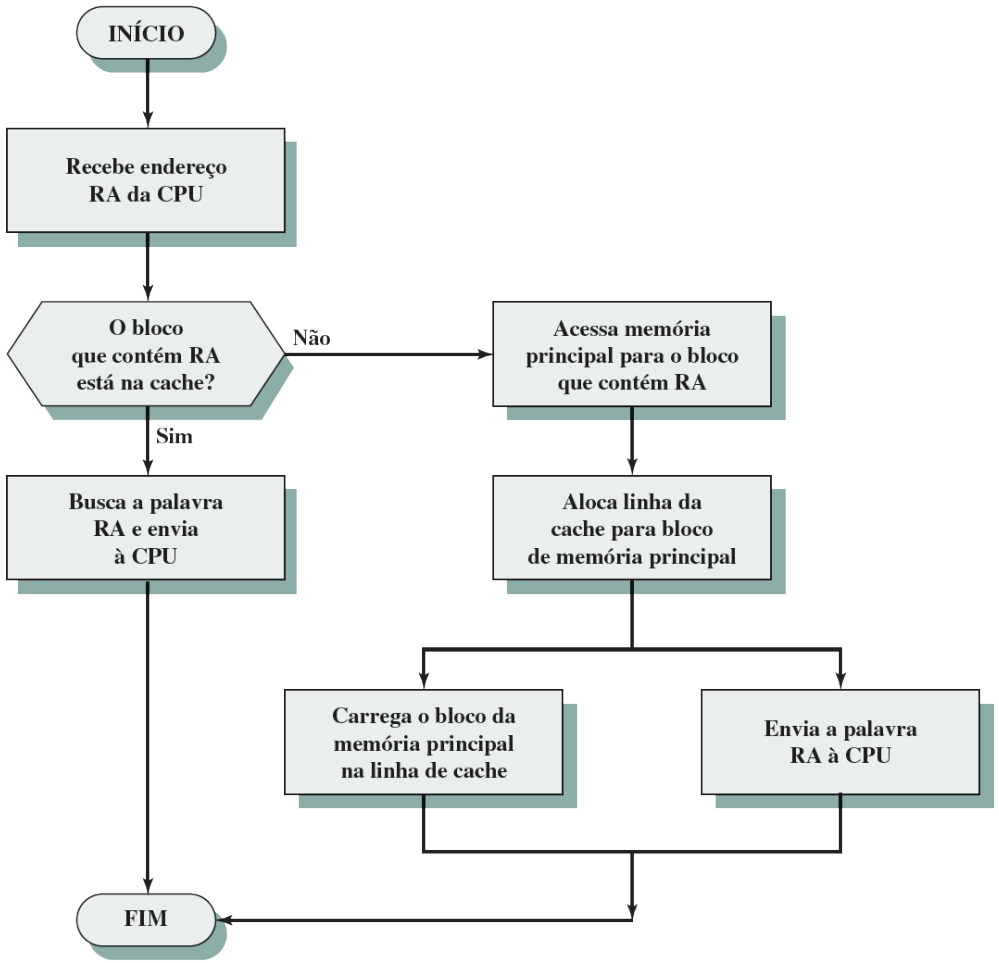
(a) Cache



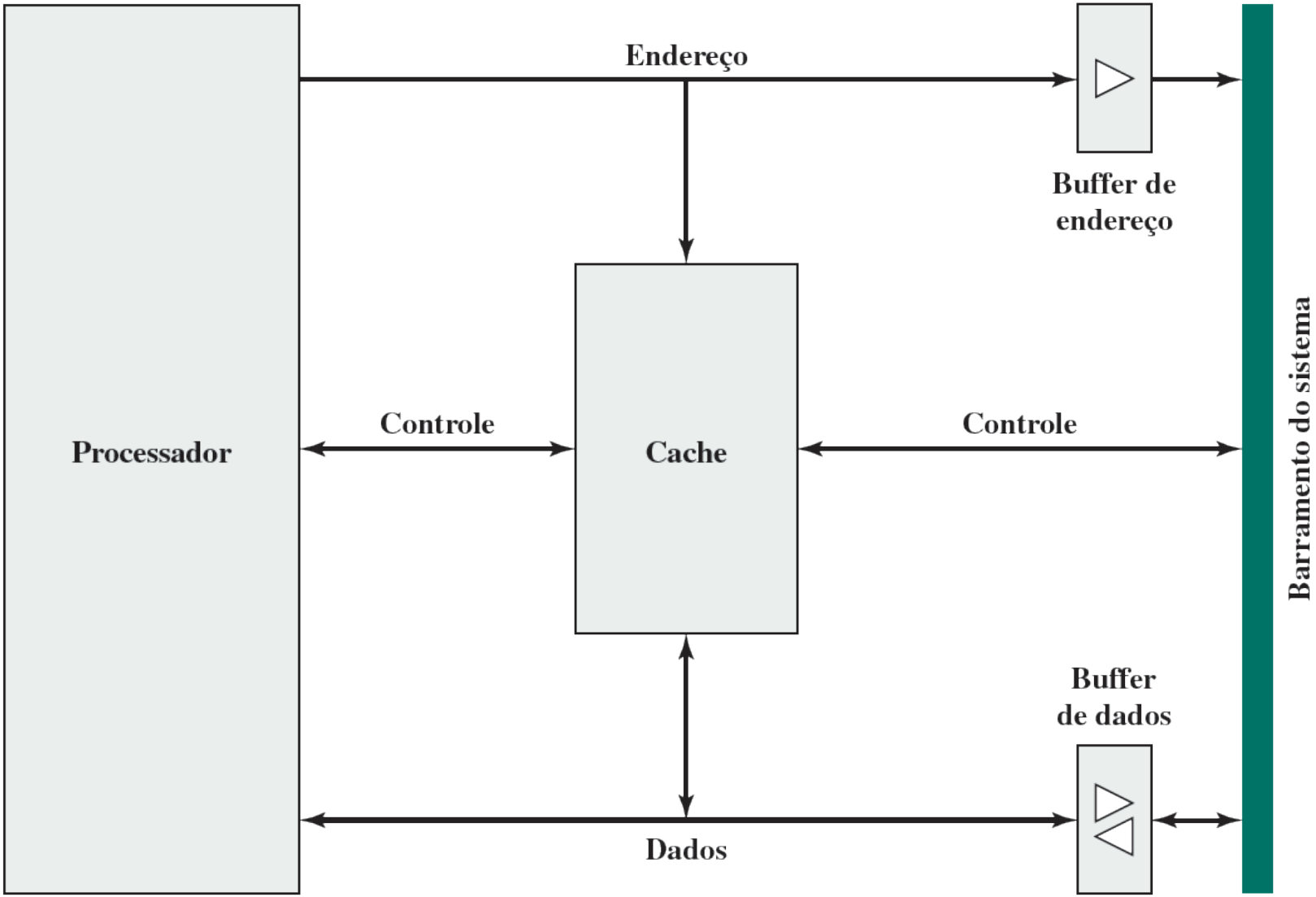
(b) Memória principal

Memória cache

- Endereço de leitura (RA - Read Adress).



Memória cache



Memória cache

Endereços da cache	Política de escrita
Lógico	*Write through*
Físico	*Write back*
Tamanho da memória cache Função de mapeamento	**Tamanho da linha Número de caches**
Direto	Um ou dois níveis
Associativo	Unificada ou separada
Associativo em conjunto	
Algoritmo de substituição	
Usado menos recentemente (LRU — do inglês, *Least Recently Used*)	
Primeiro a entrar, primeiro a sair (FIFO — do inglês, *First In, First Out*)	
Usado menos frequentemente (LFU — do inglês, *Least Frequently Used*)	
Aleatória	

Memória cache

- Uma cache física armazena dados usando endereços físicos da memória principal.
- Uma cache lógica, também conhecida como cache virtual, armazena dados usando endereços virtuais.
- Uma vantagem da cache lógica é que a velocidade de acesso a ela é maior do que para uma cache física, pois a cache pode responder antes que a unidade de gerenciamento da memória (MMU - memory management unit) realize uma tradução de endereço.
- Uma desvantagem é que a maioria dos sistemas de memória virtual fornece, a cada aplicação, o mesmo espaço de endereços de memória virtual.

Figure: Cache Lógica

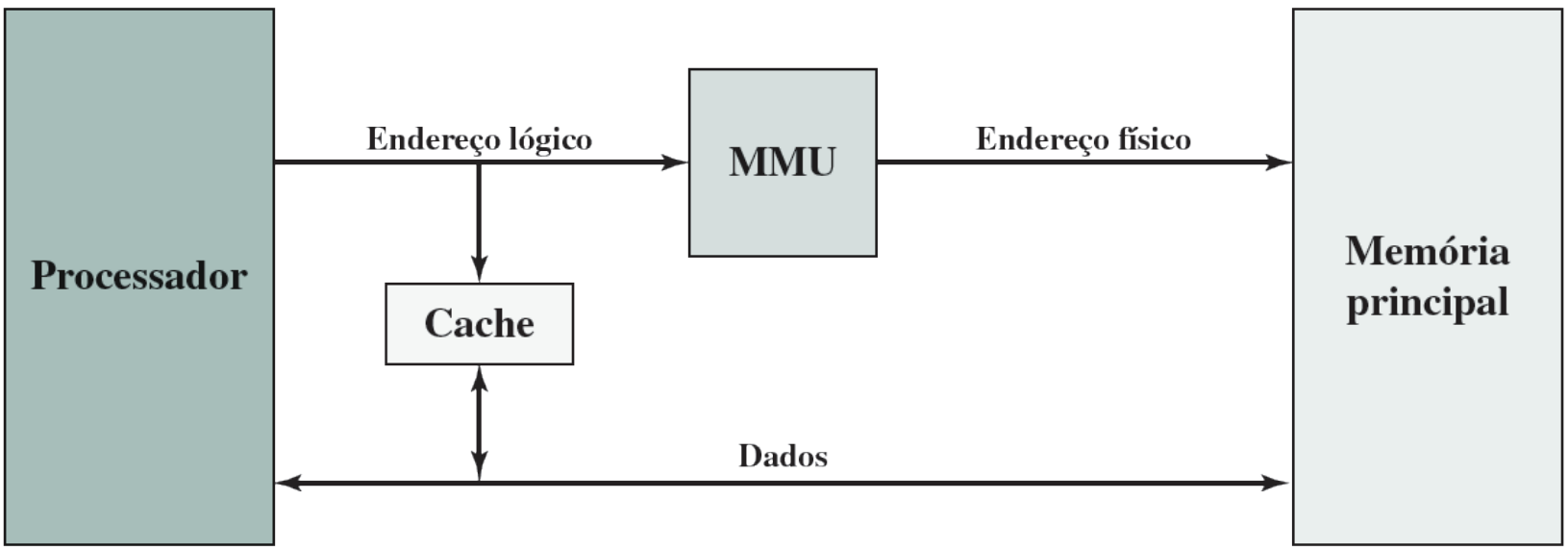
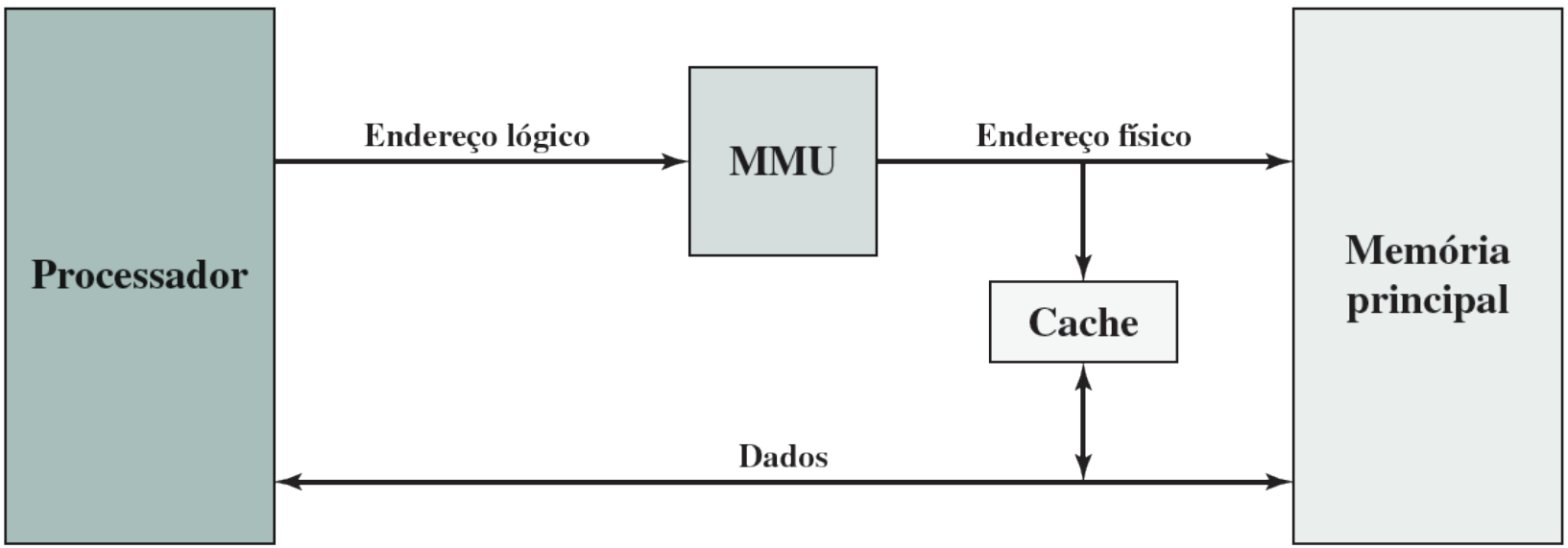


Figure: Cache Física



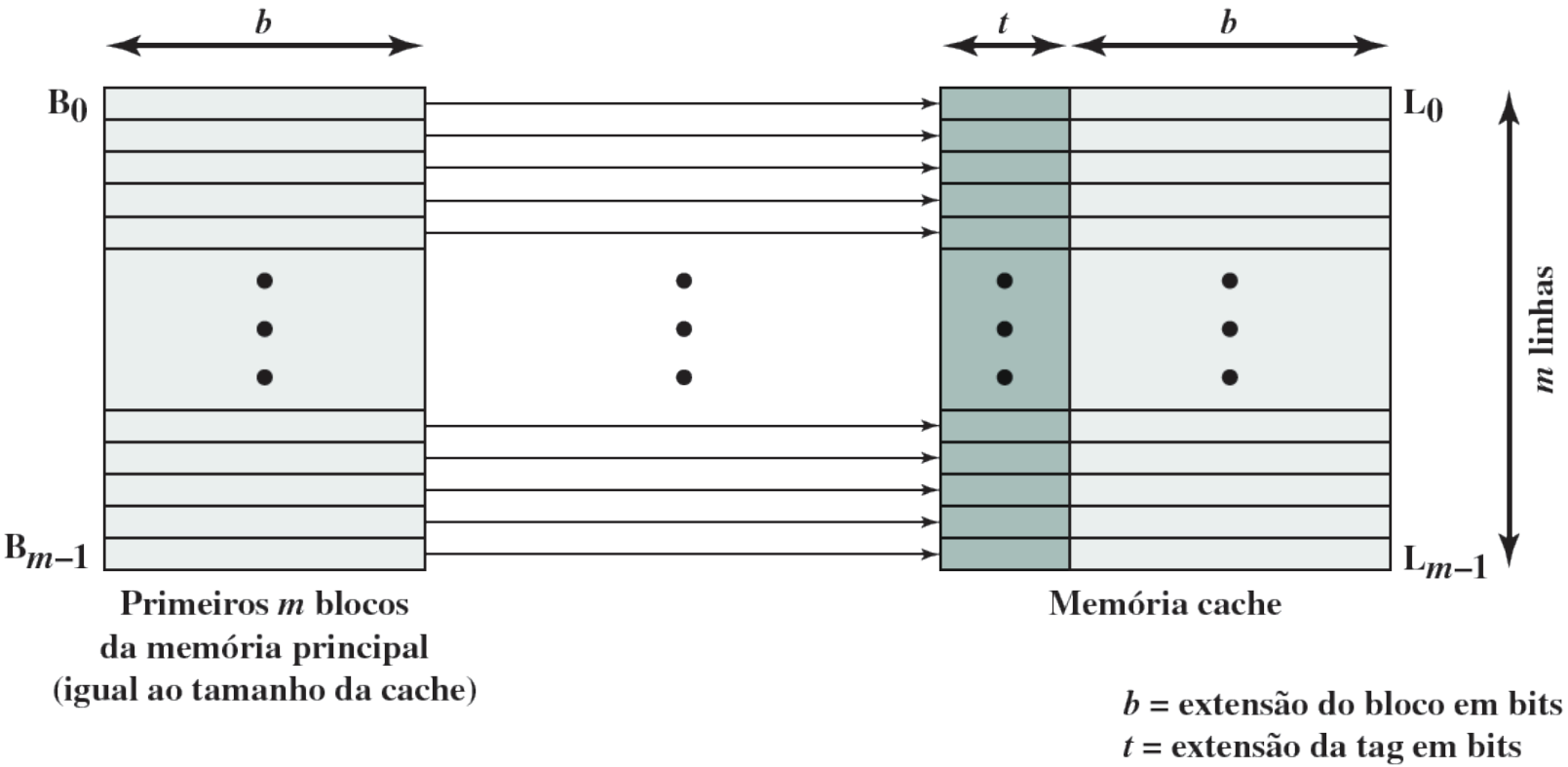
Memória cache

- Quanto maior a cache, maior o número de portas envolvidos no endereçamento da cache.
- Caches grandes tendem a ser mais lentas que as pequenas mesmo quando construídas com a mesma tecnologia de circuito integrado e colocadas no mesmo lugar no chip e na placa de circuito.
- A área disponível do chip e da placa limita o tamanho da cache.
- Como o desempenho da cache é muito sensível à natureza da carga de trabalho, é impossível chegar a um único tamanho ideal de cache.

- **4.4) Quais são as diferenças entre mapeamento direto, mapeamento associativo e mapeamento associativo em conjunto?**
- É necessário haver um algoritmo para mapear os blocos da memória principal às linhas de cache.
- É preciso haver um meio para determinar qual bloco da memória principal atualmente ocupa uma linha da cache.
- Para função de mapeamento três técnicas podem ser utilizadas:
  - **Direta**;
  - **Associativa**;
  - **Associativa por Conjunto**.

Mapeamento direto

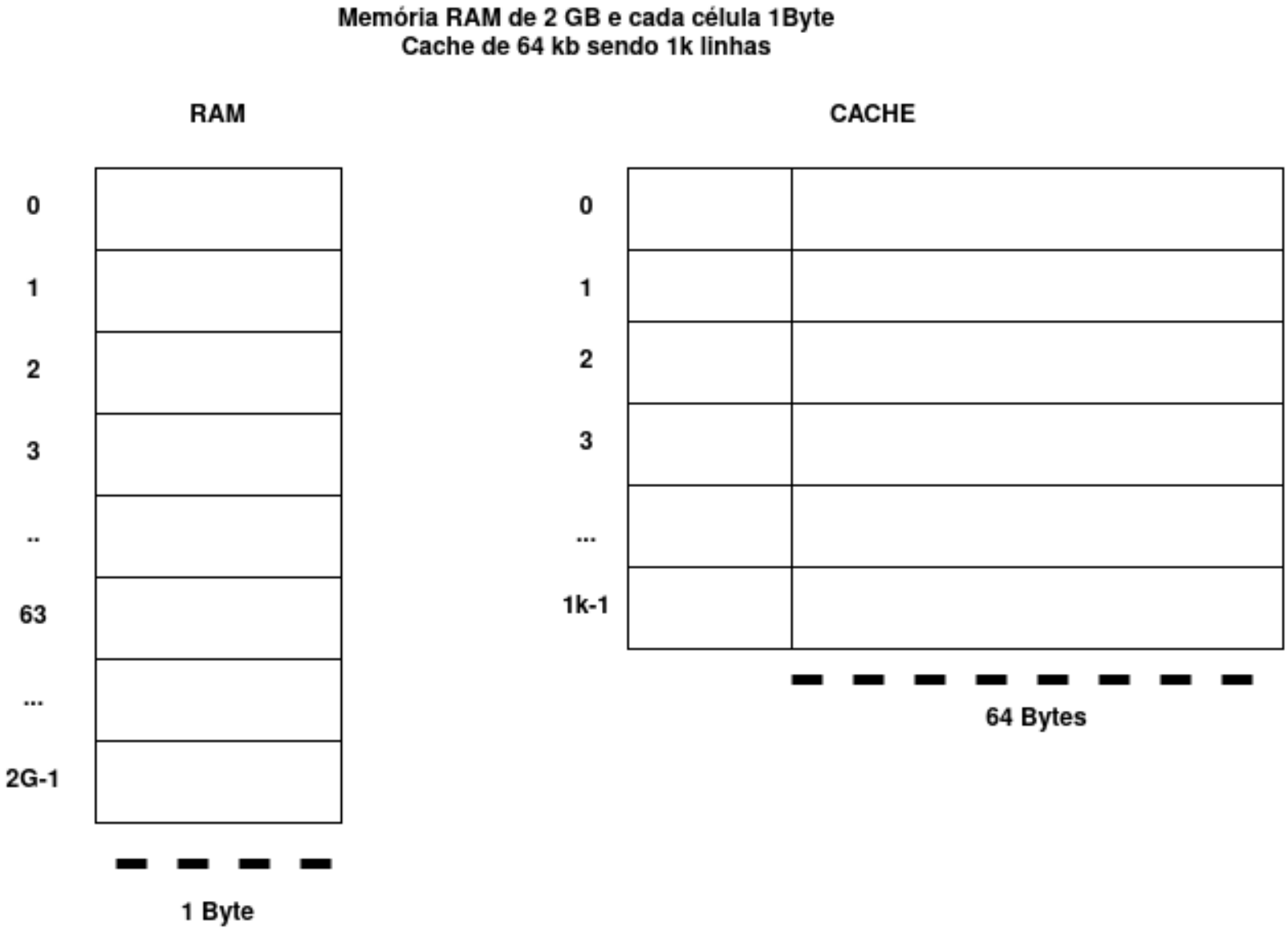
- A técnica mais simples, conhecida como mapeamento direto, mapeia cada bloco da memória principal a apenas uma linha de cache possível.



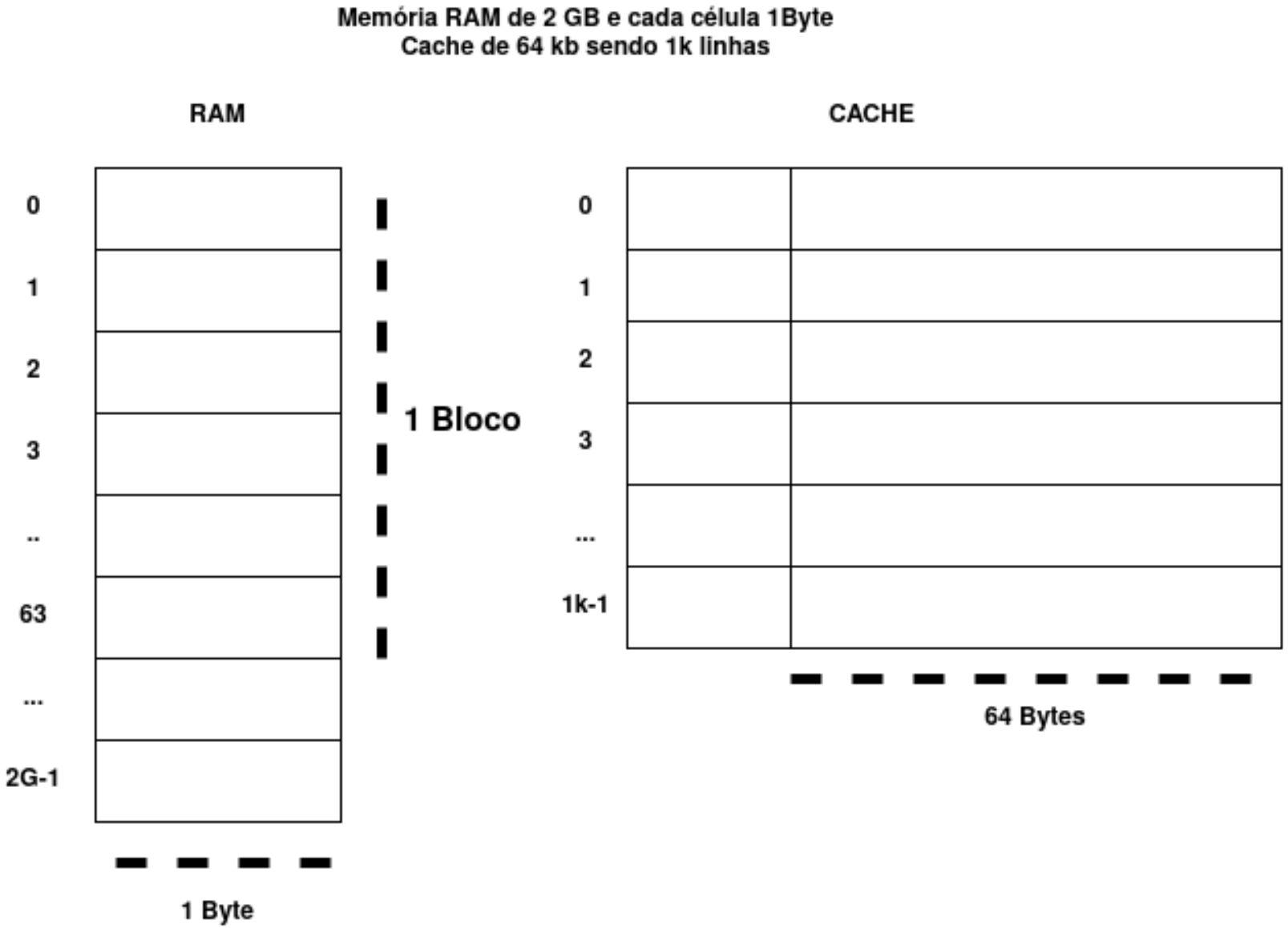
Mapeamento direto

- Vantagens:
  - Simplicidade e Velocidade.
  - Hardware barato.
  - Procura simples (posição fixa).
- Desvantagens:
  - Pode ter mau aproveitamento das posições da cache.

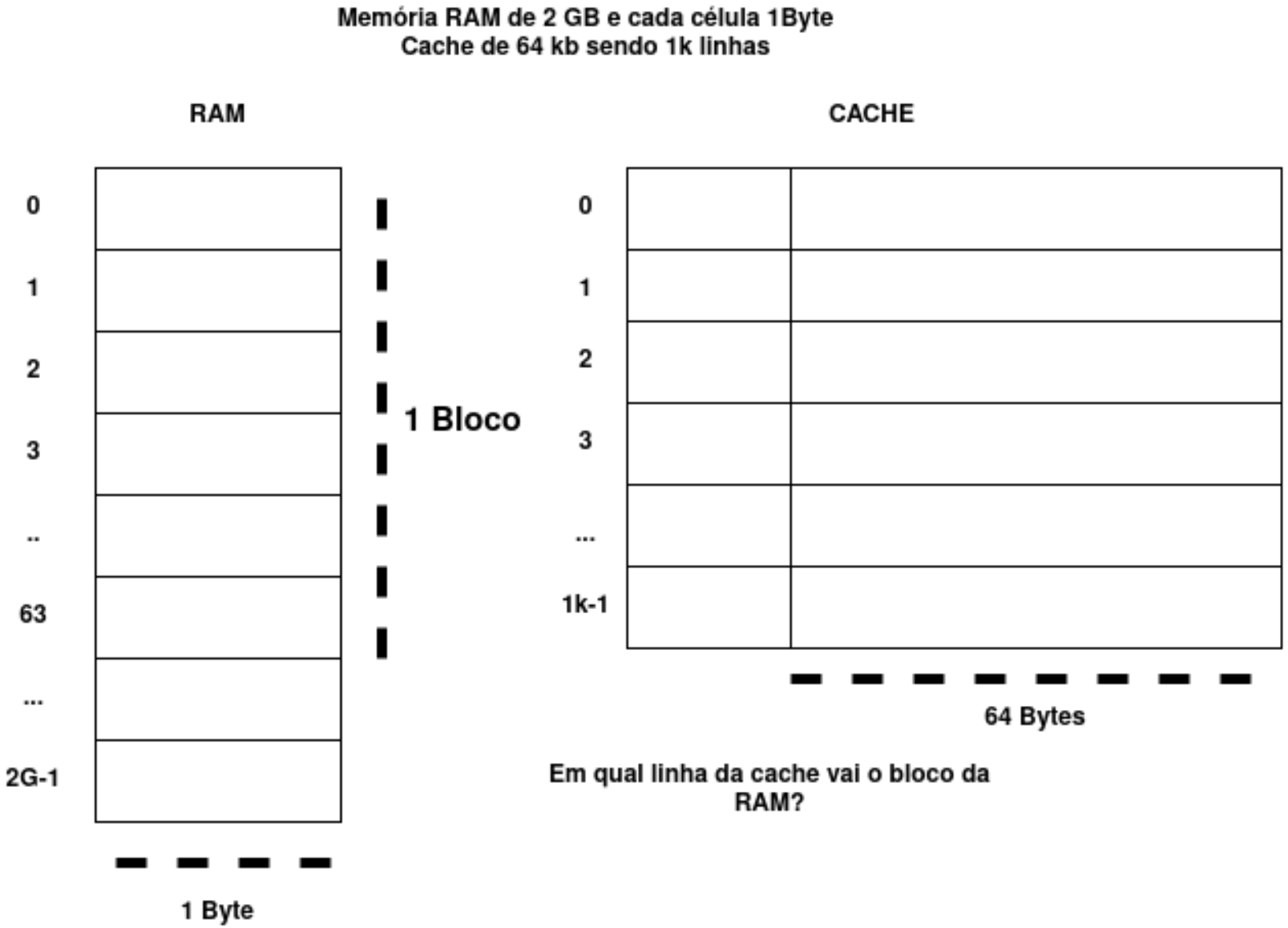
Mapeamento direto



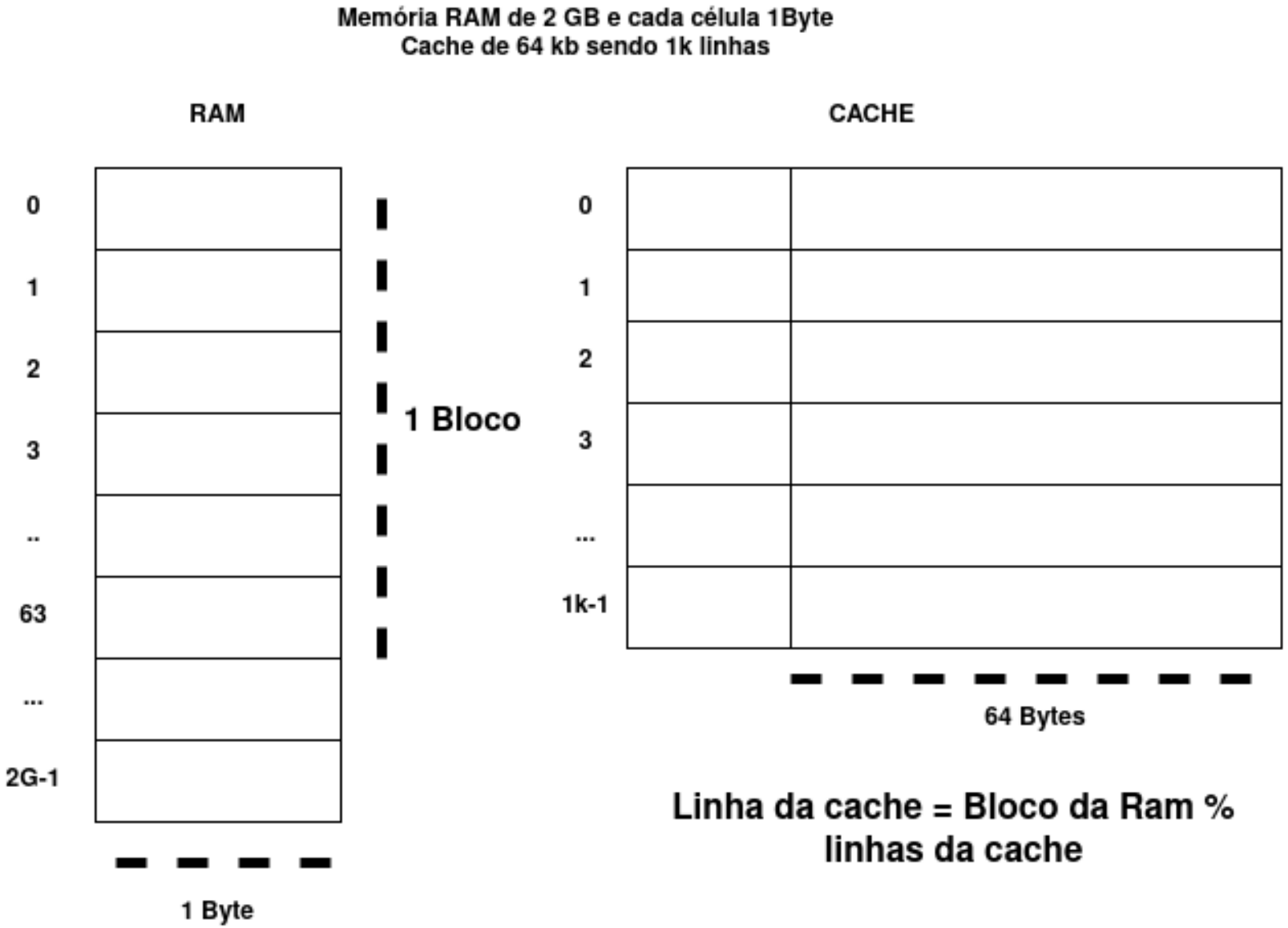
Mapeamento direto



Mapeamento direto

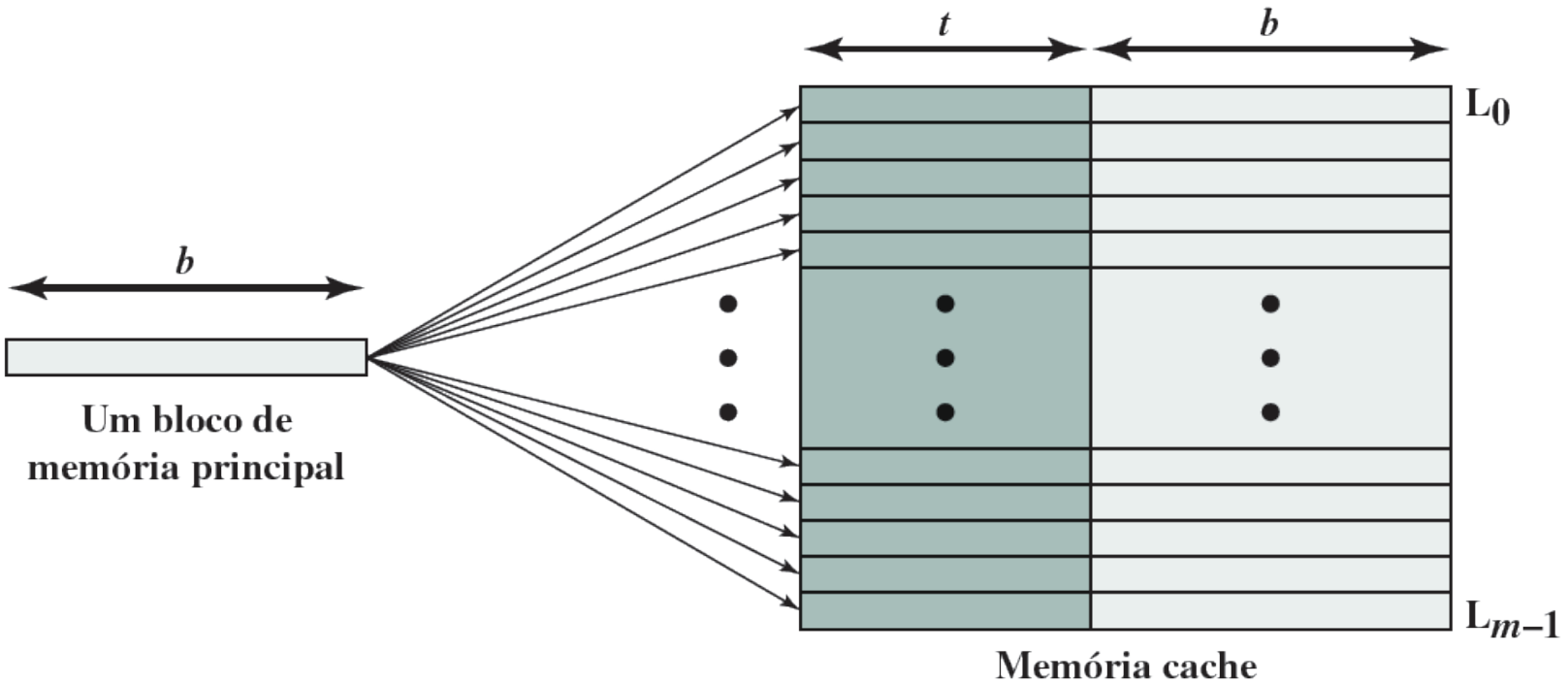


Mapeamento direto



Mapeamento associativo

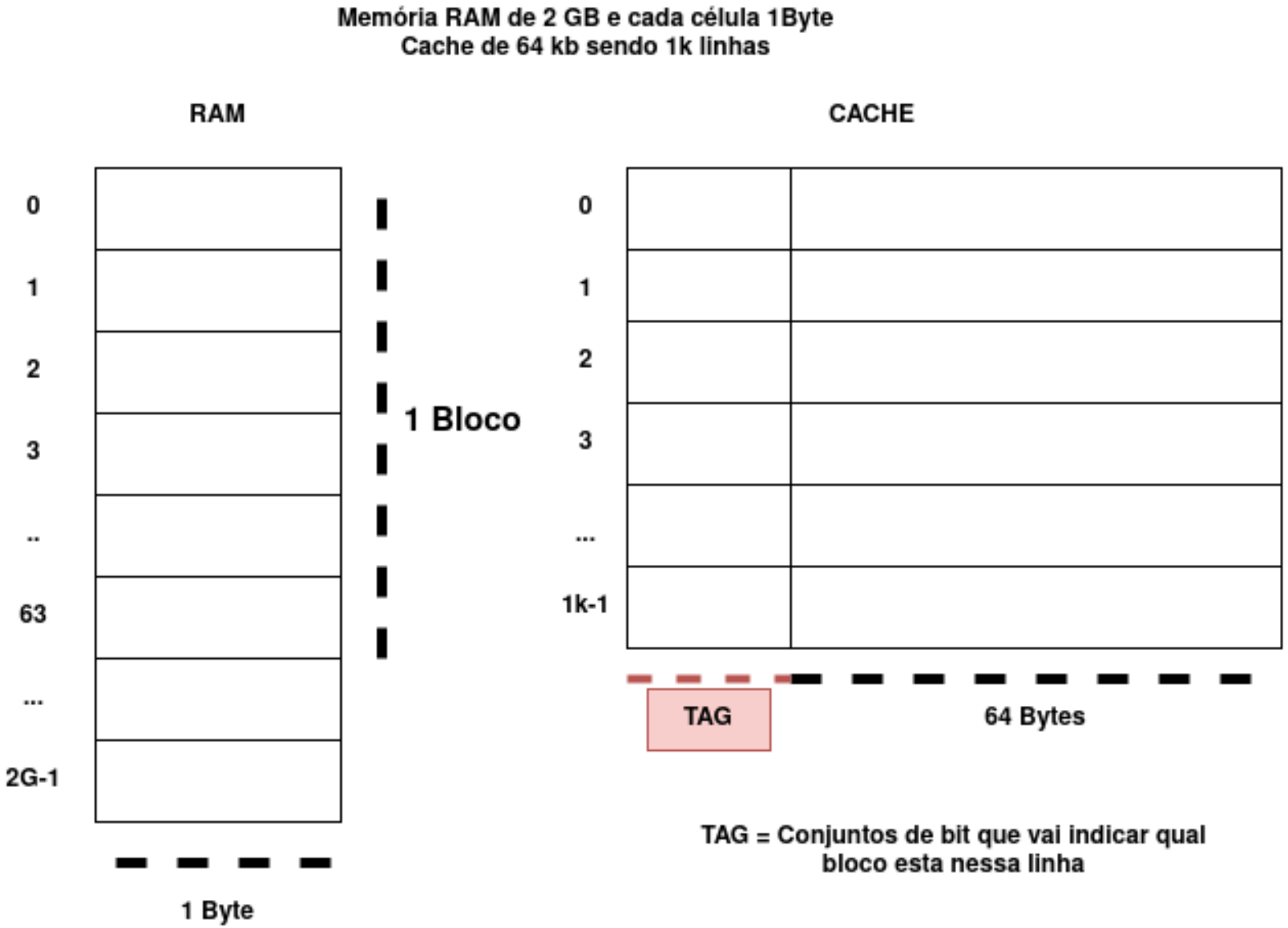
- Permite que cada bloco da memória principal seja carregado em qualquer linha da cache.



Mapeamento associativo

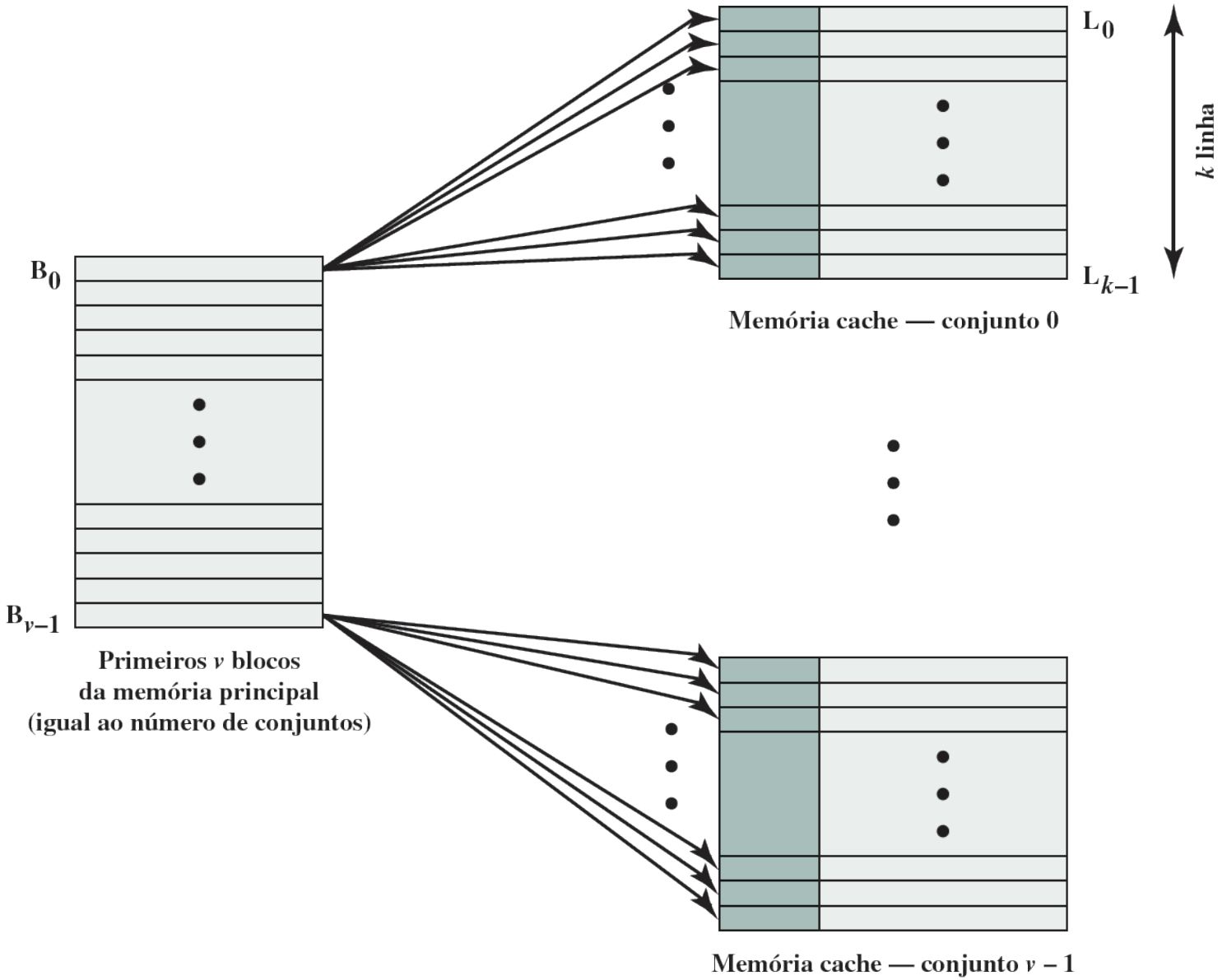
- Vantagens:
  - Melhor distribuição da informação na cache.
- Desvantagens:
  - Memória associativa tem alto custo e tamanho limitado.
  - Necessita política de substituição.

Mapeamento associativo



Mapeamento associativo por Conjunto

- Mapeamento associativo e direto.



Mapeamento associativo por Conjunto

- Vantagens:
  - Aumenta tamanho da cache mantendo tamanho da memória associativa.
- Desvantagens:
  - Memória associativa tem alto custo e tamanho limitado.
  - Somente faz substituição dentro do conjunto.
  - Necessita política de substituição.

- **4.5) Explique rapidamente os algoritmos de substituição LRU, FIFO, LFU e Aleatório.**
- Uma vez que a cache esteja cheia, e um novo bloco seja trazido para a cache, um dos blocos existentes precisa ser substituído.
- Para as técnicas associativa e associativa em conjunto, um algoritmo de substituição é necessário.
- Usado menos recentemente (**LRU**).
- Primeiro a entrar, primeiro a sair (**FIFO**).
- Usado menos frequentemente (**LFU**).

Memória cache

- Quando um bloco que está residente na cache estiver para ser substituído, existem dois casos a serem considerados.
  - Se o bloco antigo na cache não tiver sido alterado, ele pode ser substituído por novo bloco sem primeiro atualizar o bloco antigo.
  - Se pelo menos uma operação de escrita tiver sido realizada em uma palavra nessa linha da cache, então a memória principal precisa ser atualizada escrevendo a linha de cache no bloco de memória antes de trazer o novo bloco.
- Diversas políticas de escrita são possíveis, com escolhas econômicas e de desempenho.

Memória cache

- A técnica mais simples é **write through**.
- As operações de escrita são feitas na memória principal e também na cache, garantindo que a memória principal sempre seja válida.
- Na técnica conhecida como **write back**, as atualizações são feitas apenas na cache e somente na modificação do bloco que ele é salvo na memória principal.
- Em uma organização de barramento em que mais de um dispositivo tem uma cache e a memória principal é compartilhada, um novo problema é introduzido.

Memória cache

- Se os dados em uma cache forem alterados, isso invalida não apenas a palavra correspondente na memória principal, mas também essa mesma palavra em outras caches.
- Mesmo que uma política write through seja usada, as outras caches podem conter dados inválidos.
- Diz-se que um sistema que impede esse problema mantém coerência de cache.

- Algumas das técnicas possíveis para a coerência de cache são:
  - **Observação do barramento com write through**: cada controlador de cache monitora as linhas de endereço para detectar as operações de escrita para a memória por outros mestres de barramento.
  - **Transparência do hardware:** um hardware adicional é usado para garantir que todas as atualizações na memória principal por meio da cache sejam refletidas em todas as caches.
  - **Memória não cacheável:** somente uma parte da memória principal é compartilhada por mais de um processador, e esta é designada como não cacheável.

Memória cache

- Caches multinível.
- À medida que a densidade lógica aumenta, torna-se possível ter uma cache no mesmo chip que o processador: a cache no chip.
- A cache no chip reduz a atividade do barramento externo do processador e, portanto, agiliza o tempo de execução e aumenta o desempenho geral do sistema.
- A organização mais simples desse tipo é conhecida como uma cache de dois níveis, com a cache interna designada como nível 1 (L1) e a cache externa designada como nível 2 (L2).

Caches unificadas versus separadas

- Mais recentemente, tornou-se comum dividir a cache em duas: uma dedicada a instruções e uma dedicada a dados.
- Essas duas caches existem no mesmo nível, normalmente como duas caches L1.
- Quando o processador tenta buscar uma instrução da memória principal, ele primeiro consulta a cache L1 de instrução.
- Quando o processador tenta buscar dados da memória principal, ele primeiro consulta a cache L1 de dados.

Memória cache

Problema	Solução	Processador em que o recurso apareceu inicialmente
Memória externa mais lenta que o barramento do sistema	Acrescentar cache externa usando tecnologia de memória mais rápida	386
Maior velocidade do processador torna o barramento externo um gargalo para o acesso à cache L2	Mover a cache externa para o chip, trabalhando na mesma velocidade do processador	486
Cache interna um tanto pequena, por conta do espaço limitado no chip	Acrescentar cache L2 externa usando tecnologia mais rápida que a memória principal	486
Quando ocorre uma disputa entre o mecanismo de pré-busca de instruções e a unidade de execução no acesso simultâneo à memória cache. Nesse caso, a busca antecipada é adiada até o término do acesso da unidade de execução aos dados	Criar caches separadas para dados e instruções	Pentium
Maior velocidade do processador torna o barramento externo um gargalo para o acesso à cache L2	Criar barramento *back-side* separado, que trabalha com velocidade mais alta que o barramento externo principal (*front-side*). O barramento *back-side* é dedicado à cache L2	Pentium Pro
	Mover cache L2 para o chip do processador	Pentium II
Algumas aplicações lidam com bancos de dados enormes, e precisam ter acesso rápido a grandes quantidades de dados. As caches no chip são muito pequenas	Acrescentar cache L3 externa	Pentium III
	Mover cache L3 para o chip	Pentium 4

Memória cache

- Exemplo Simulador.
- Bit de validade (V): Se a posição na cache esta ocupada ou contém lixo.
- Tag: Identifica qual das palavras está na cache.
- https://simuladorcache.leandrogabriel.net/.

Bibliografia Básica

- STALLINGS, W. **Arquitetura e Organização de Computadores**. 10 ed. São Paulo: Pearson, 2017;
- TANENBAUM, A. S. **Organização Estruturada de Computadores**. 5 ed. Pearson 2007;
- HENNESY, J. PATTERSON, D. **Organização e Projeto de Computadores**. 3 ed. Editora Campus, 2005.

Obrigado!
Dúvidas?

Guilherme Henrique de Souza Nakahata

guilhermenakahata@gmail.com

https://github.com/GuilhermeNakahata/UNESPAR-2023