Memória Cache

Prof. Filippo Valiante Filho

http://prof.Valiante.info

Memórias – Aula 3 de 4 - Versão 2

Analogia da Mesa, Gaveta e Arquivo

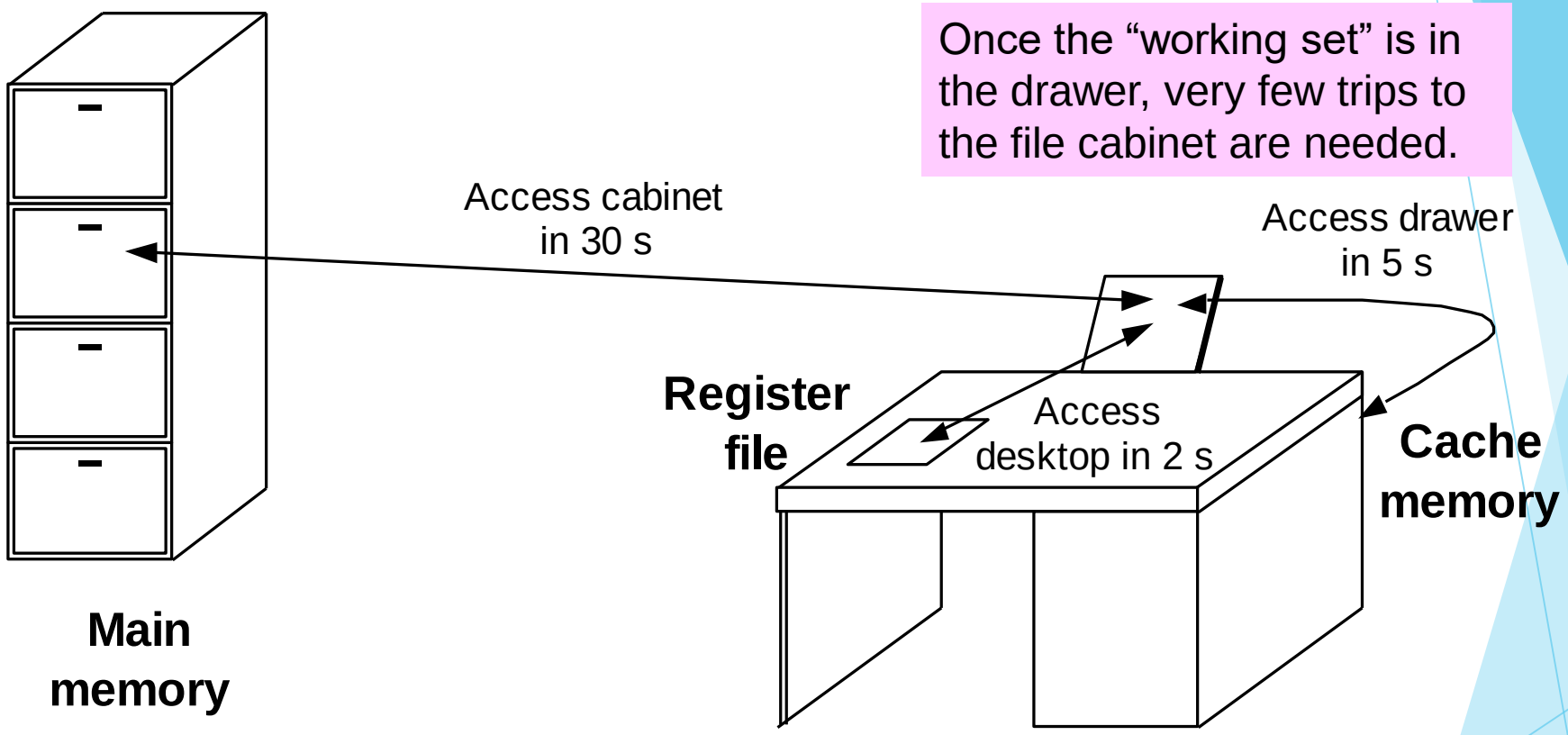


Fig. 18.3 Items on a desktop (register) or in a drawer (cache) are more readily accessible than those in a file cabinet (main memory).

Parhami (2007)

Memória Cache

- Pequena quantidade de memória rápida (SRAM).
- Fica entre a memória principal normal (DRAM lenta) e a CPU, dentro do microprocessador.
- Termo significa “escondido”. A memória cache é “transparente”, totalmente controlada pelo hardware.

Funcionamento da Cache

- CPU requisita conteúdo de uma posição de memória.
- O hardware verifica se o dado está na cache.
- Se sim:
  - Entrega a partir da cache (rápido).
- Se não:
  - Lê bloco solicitado na memória principal e copia os dados para a cache.
  - Depois entrega da cache à CPU.
  - Cache inclui tags para identificar qual bloco da memória principal está em cada slot da cache.

Princípios de Localidade

- Localidade Temporal
  - posições da memória, uma vez acessadas tendem a ser acessadas outra vez no futuro próximo.
    - Dados
- Localidade Espacial
  - endereços em próximos acessos tendem a ser próximos (subsequentes) de endereços acessados anteriormente.
    - Instruções

Características da Memória Cache

- Tamanho.
- Função de mapeamento.
- Algoritmo de substituição.
- Política de escrita.
- Tamanho de bloco.
- Número de níveis de cache.

Tamanho

- Custo X Benefício
- Custo:
  - Mais cache é mais caro.
- Benefício (aumento de desempenho):
  - Mais cache implica em melhor desempenho, mas só até certo ponto.
  - Verificar dados na cache leva tempo.

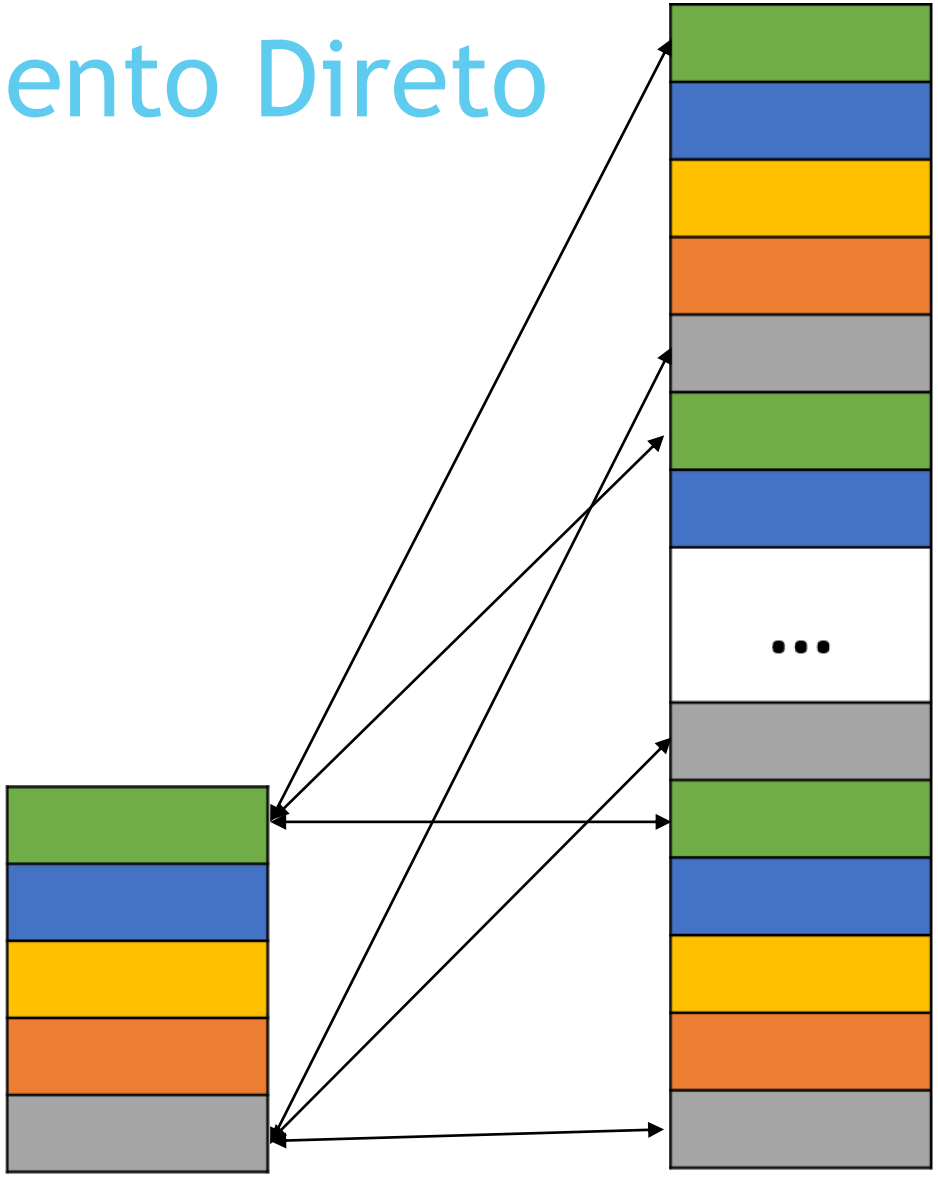
Função de Mapeamento

- A cache possui menos endereços que a Memória Principal.
- A função de mapeamento é a formar de estabelecer uma correspondência entre os endereços da memória principal e os endereços da memória cache.
  - Mapeamento direto
  - Mapeamento associativo
  - Mapeamento associativo em conjuntos

Mapeamento Direto

- Cada bloco de memória principal mapeado apenas para uma linha de cache.
  - Ou seja, se um bloco está na cache, ele deve estar em um local específico.

Mapeamento Direto



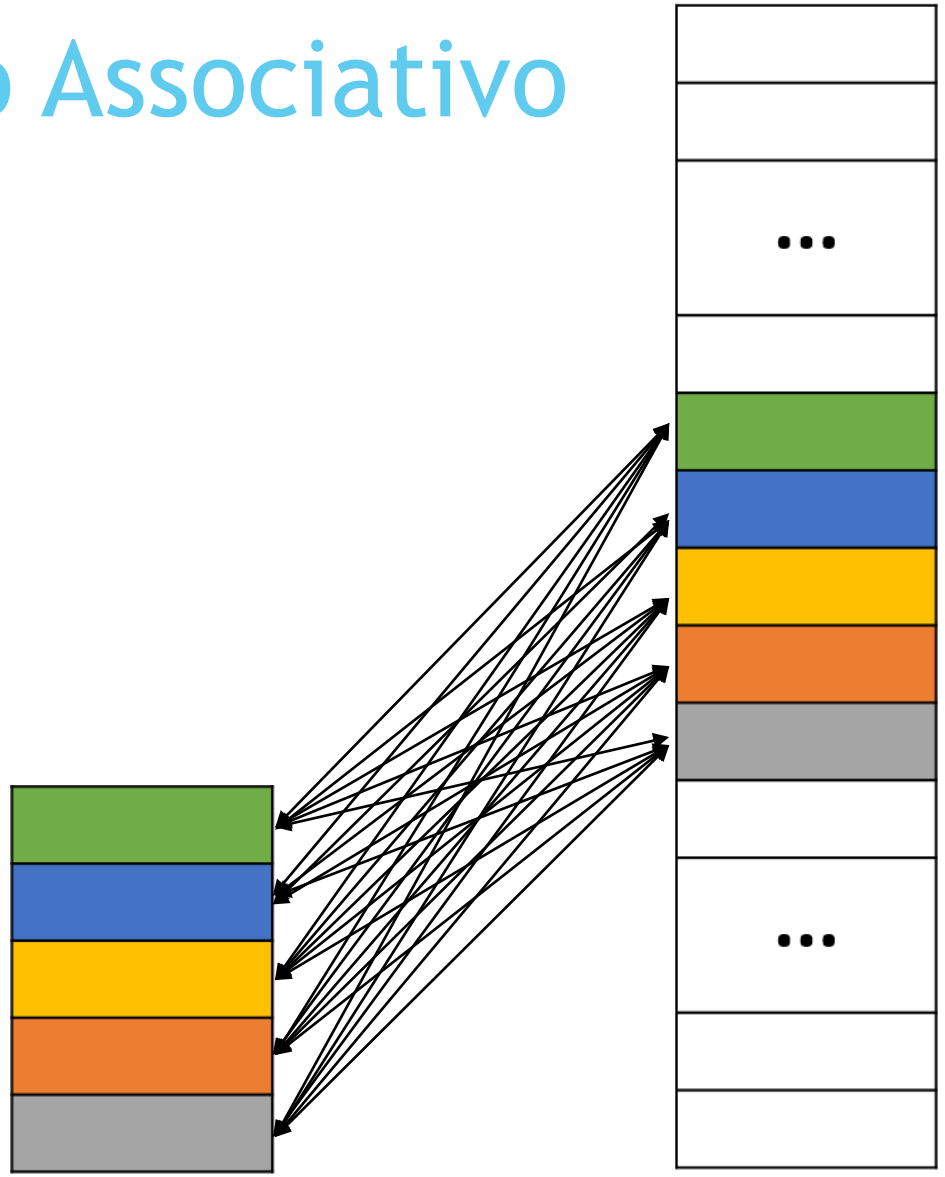
Prós e contras do mapeamento direto

- Simples.
- Barato.
- Local fixo para determinado bloco.
  - Se um programa acessa 2 blocos que mapeiam para a mesma linha repetidamente, **perdas de cache (cache miss) são muito altas.**

Mapeamento Associativo

- Um bloco de memória principal pode ser carregado em qualquer linha de cache.
- Pesquisa da cache é dispendiosa.
- Impraticável.

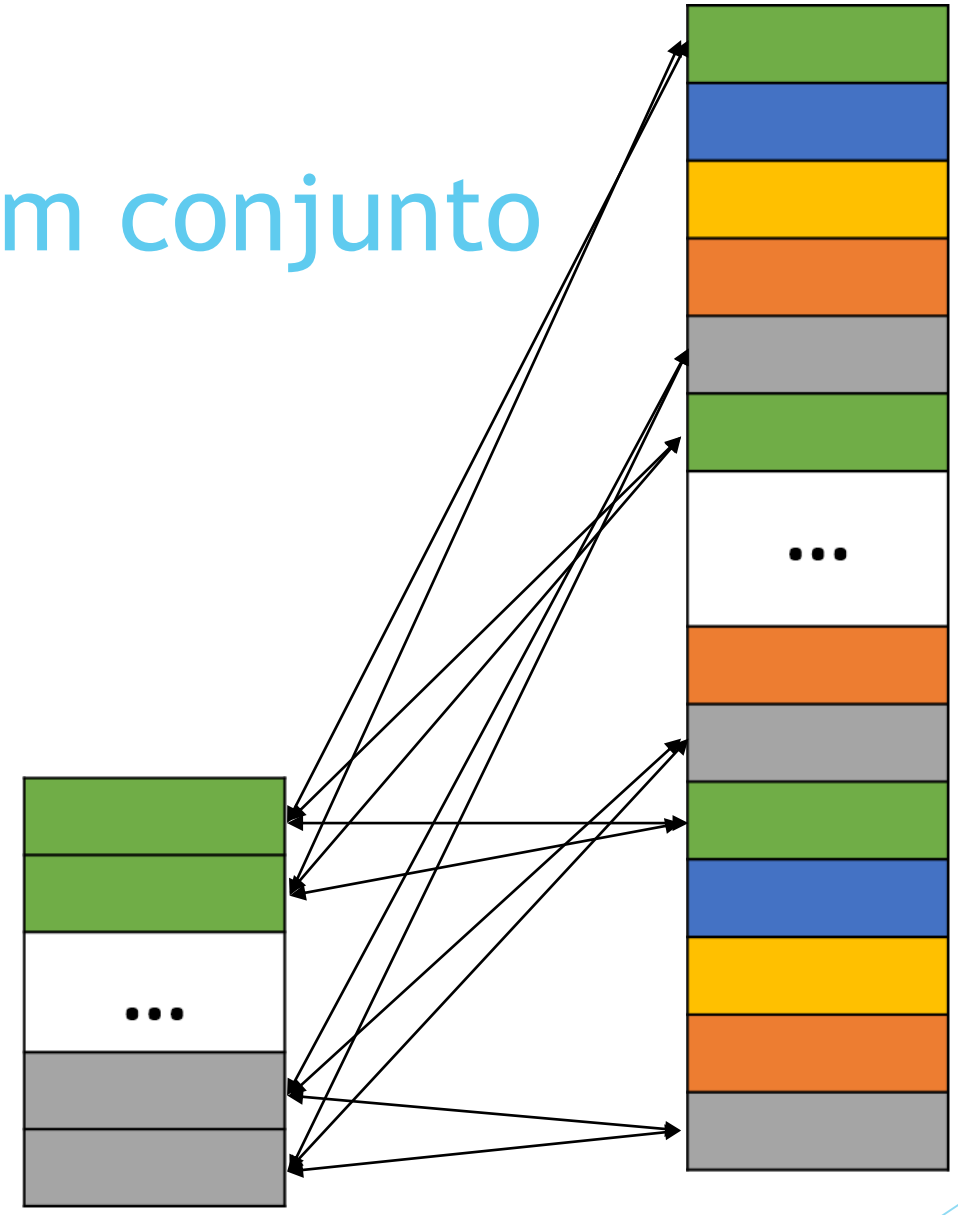
Mapeamento Associativo



Mapeamento Associativo em Conjunto

- Set Associative
- Cache é dividida em uma série de conjuntos.
- Cada conjunto contém uma série de linhas.
- Determinado bloco é mapeado a qualquer linha em determinado conjunto.
  - P.ex.,Bloco B pode estar em qualquer linha do conjunto i.
- P.ex. 2 linhas por conjunto:
  - Mapeamento associativo com 2 linhas.
  - Determinado bloco pode estar em uma de 2 linhas em apenas um conjunto.

Mapeamento associativo em conjunto



Algoritmos de substituição

- É preciso definir que bloco de dados deixa a memória cache para abrir espaço para um novo.
- Algoritmo implementado no hardware (velocidade).
- No mapeamento direto não há escolha, cada bloco é mapeado apenas a uma linha e basta substituir a linha. Mas não é usado na prática...
- Nos mapeamentos associativo e associativo por conjuntos...

Algoritmos de Substituição

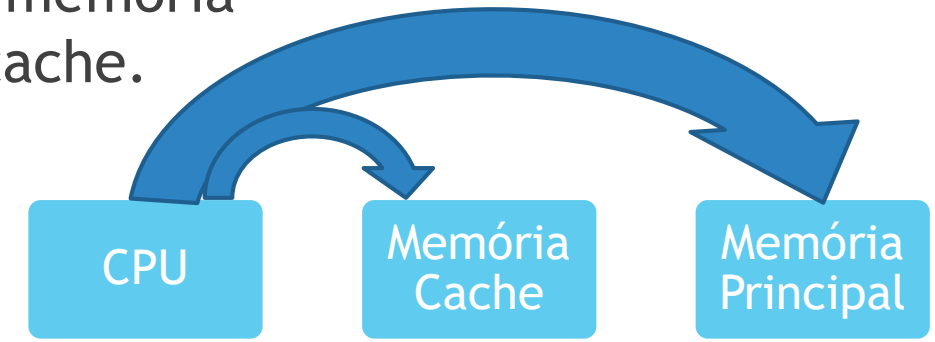
- FIFO – First In First Out.
  - Substitui bloco que está na cache há mais tempo.
- LFU – Least Frequently Used.
  - Substitui bloco menos usado.
- LRU – Least Recently Used.
  - Substitui o bloco que foi usado há mais tempo
- Aleatório.

Política de escrita

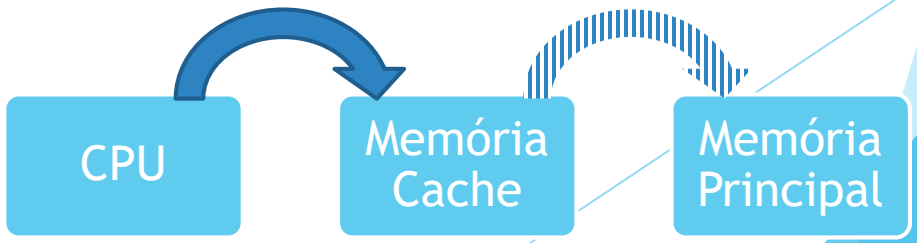
- Se há alguma mudança na cache é necessário atualizar a memória principal.
  - Não deve sobrescrever bloco de cache a menos que a memória principal esteja atualizada.
- Múltiplas CPUs podem ter caches individuais, mas na maior parte dos sistemas compartilham a mesma memória principal.
- E/S pode endereçar memória principal diretamente (DMA).

Política de escrita

- Write-through
  - Todas as escritas vão para a memória principal e também para a cache.



- Write-back
  - Atualizações feitas inicialmente apenas na cache, que por sua vez atualiza a memória principal
  - E/S deve acessar a memória principal através da cache.



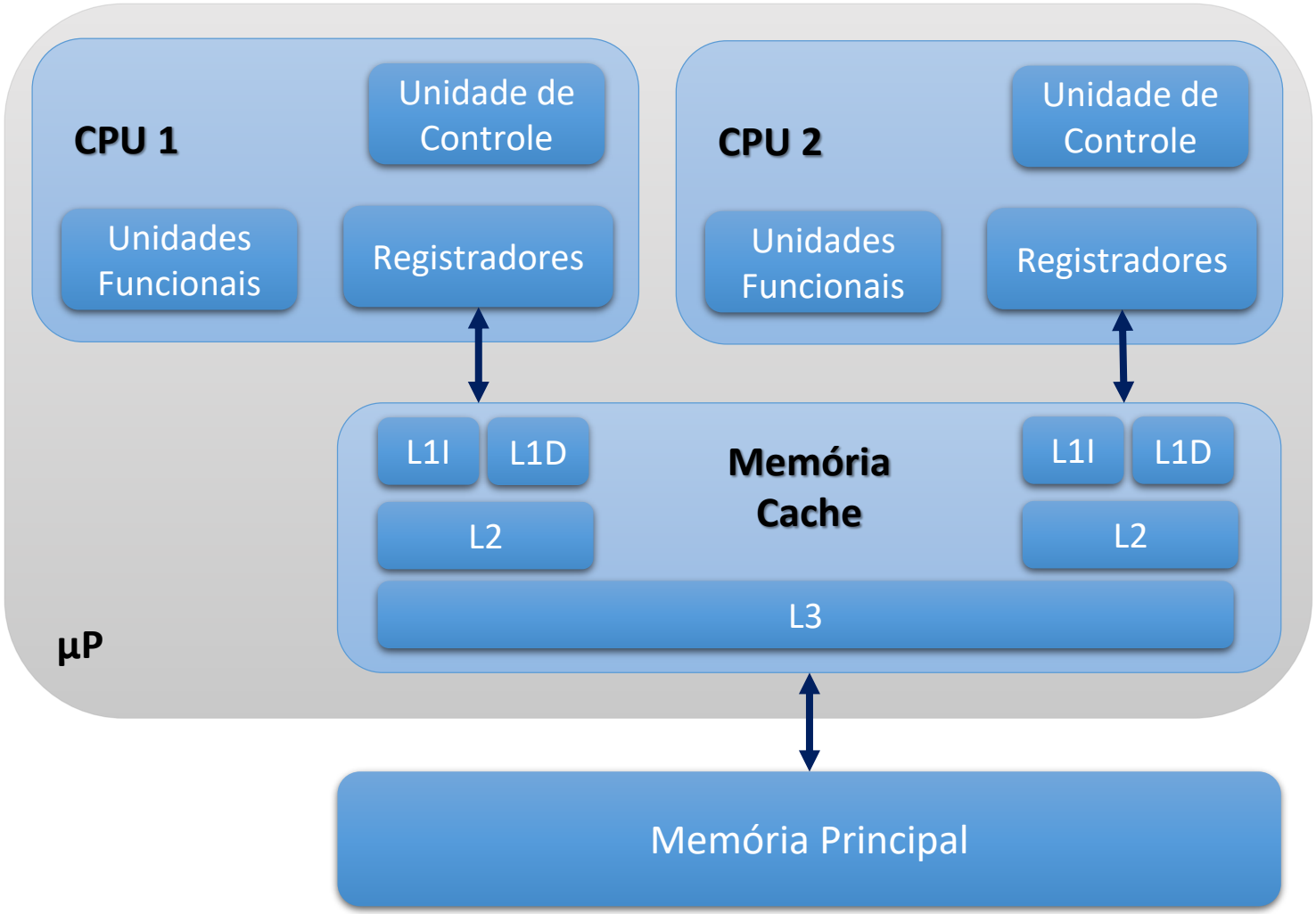
Tamanho de Linha/Bloco

- Ao invés de transferir para a memória cache apenas a posição solicitada pela CPU, transfere-se um bloco de posições de cada vez.
- Qual o tamanho ideal?
  - Muito pequeno não há tanta melhora de desempenho
  - Muito grande transfere posições que não serão acessadas
- Usa-se de 8 a 64 bytes na prática
  - 64 bytes são 8 palavras em um computador de 64 bits...

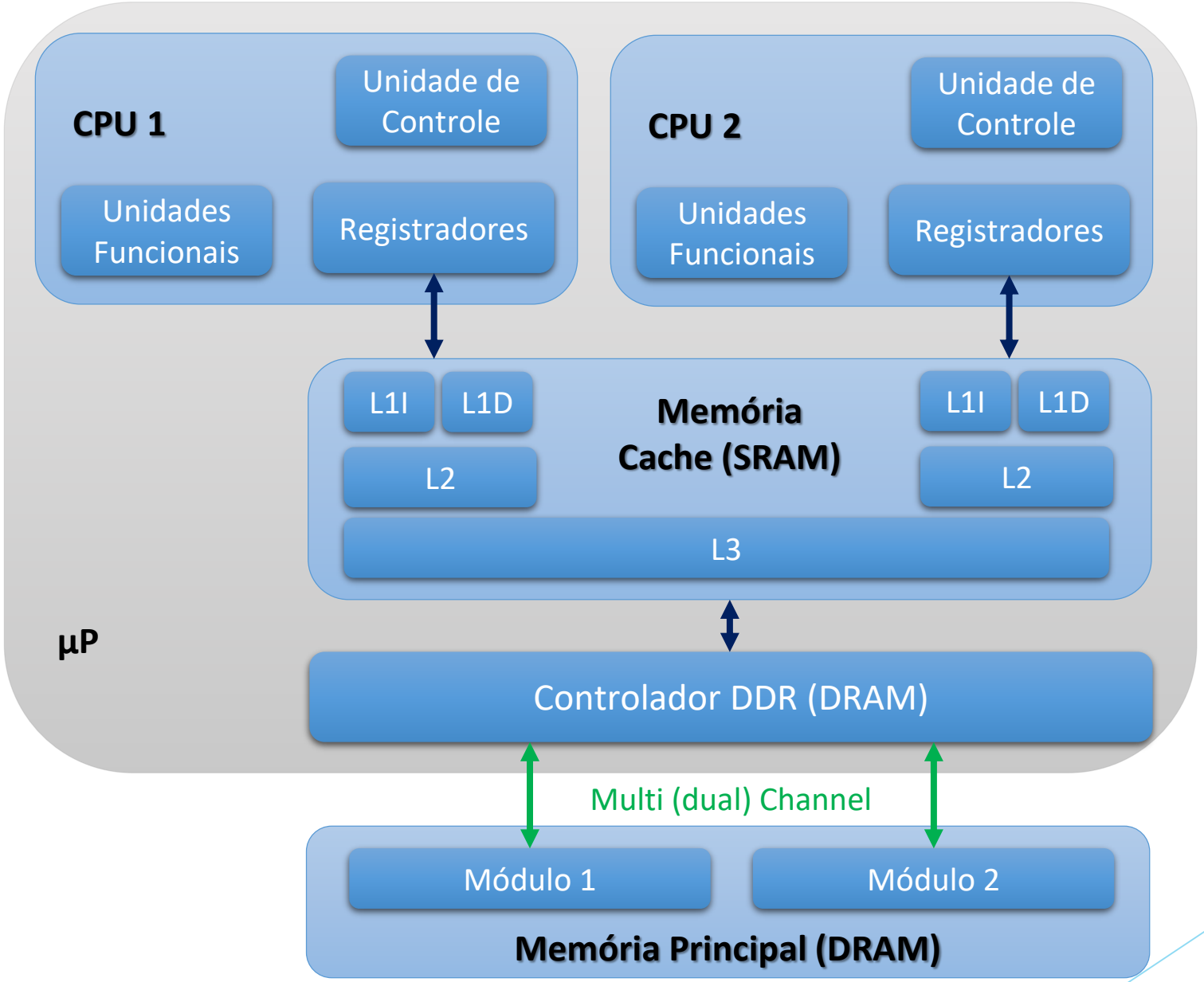
Número de Níveis de Cache

- Em geral até 3 níveis de cache (cache multinível)
- Cache nível 1 geralmente exclusiva para cada CPU e dividida em Instruções e Dados
  - Princípios de localidade distintos
  - Arquitetura Harvard
- Cache de maior nível geralmente compartilhada por todas as CPUs (na atualidade é normal a cache estar em um microprocessador multicore)

Exemplo de arquitetura de cache



Exemplo de arquitetura de cache

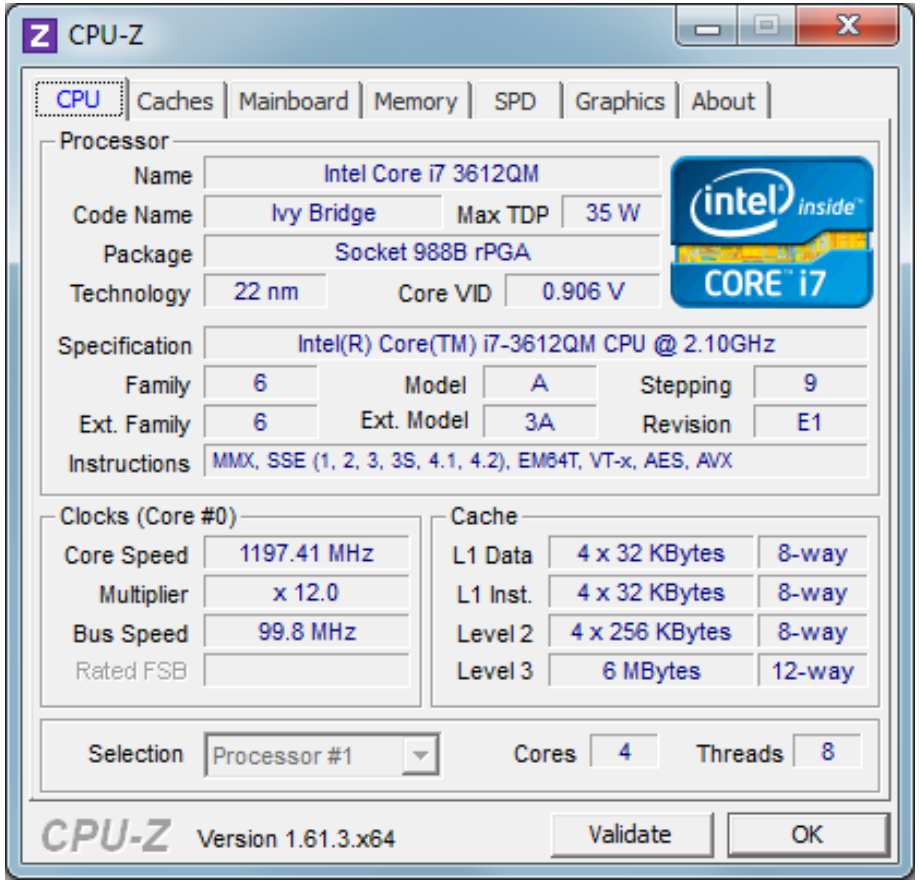


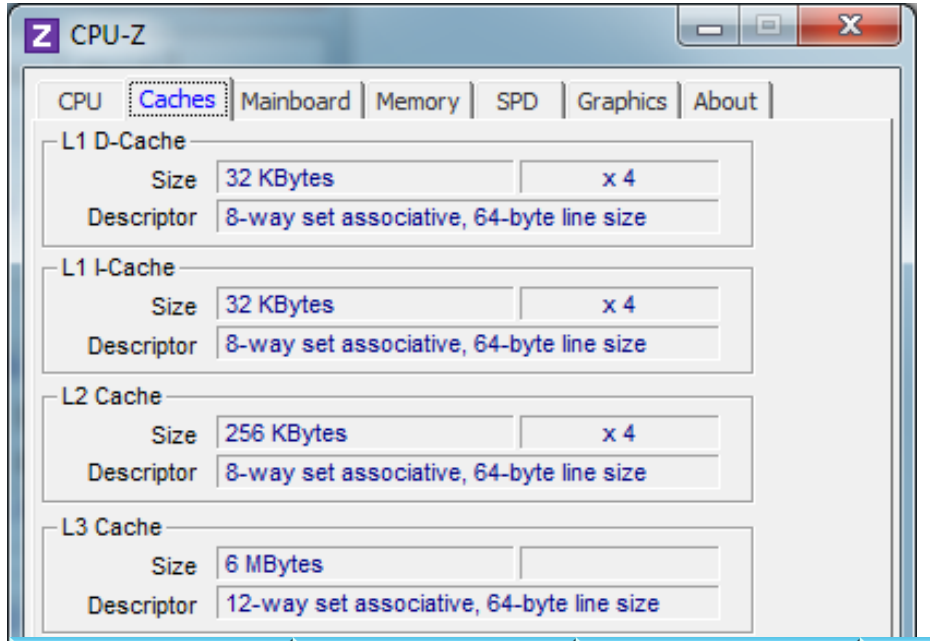
Evolução da memória cache da Intel

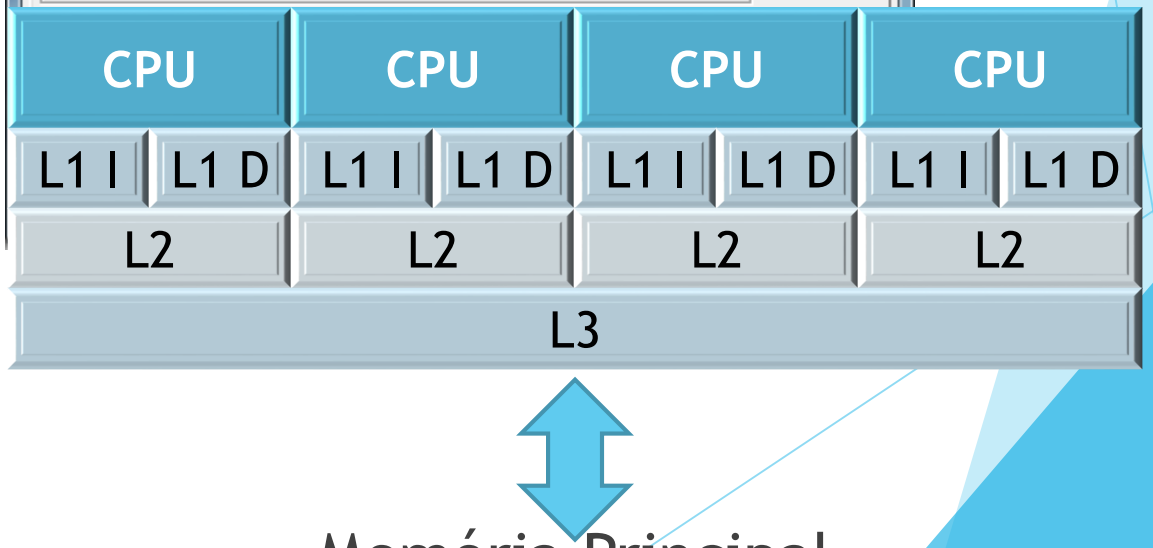
Problema	Solução	Processador em que o recurso apareceu inicialmente
Memória externa mais lenta que o barramento do sistema.	Acrescentar cache externa usando tecnologia de memória mais rápida	386
O aumento da velocidade di processador torna o barramento externo um gargalo para o acesso à memória cache..	Mover a cache externa para o chip, trabalhando na mesma velocidade do processador	486
Cache interna um tanto pequena, devido ao 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

Stallings (2010)

Conhecendo a cache do seu PC







Baixe o CPU-Z no site:

www.cpuid.com

Referências Bibliográficas

- Arquitetura e Organização de Computadores – 8ª edição
  William Stallings - Editora Pearson Education – 2010
- Organização Estruturada de Computadores - 5ª Edição
  Andrew S. Tanenbaum - Editora Pearson Education – 2007
- Introdução à Arquitetura de Computadores
  Miles J. Murdocca e Vincent P. Heuring - Editora Campus – 2000
- Arquitetura de Computadores - De Microprocessadores a Supercomputadores
  Behrooz Parhami - Editora McGraW-Hill – 2007
- Arquitetura de Computadores - Coleção Schaum
  Nicholas Carter - Editora Bookman – 2003