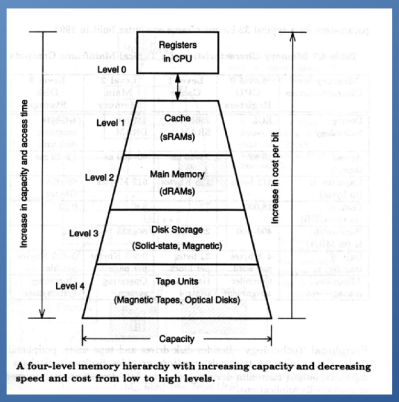
Memória Cache
Arquitetura Intel
Operações com Matrizes

William Stallings
Computer Organization and Architecture
8th Edition

Hierarquia de Memória

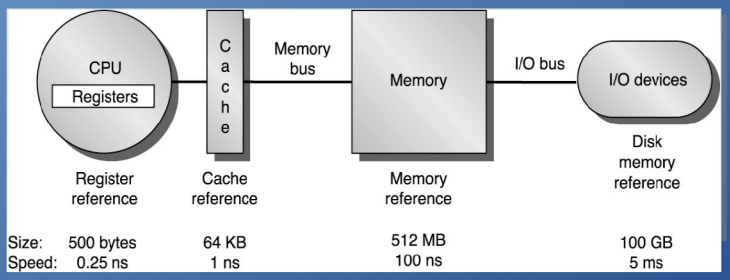
- Tecnologias



A four-level memory hierarchy with increasing capacity and decreasing speed and cost from low to high levels.

Esta imagem foi retirada de um livro antigo, mas nela fica claro as diferenças entre as memórias que podem compor a hierarquia de memória

Hierarquia de Memória



Características da Hierarquia de Memória

Nível	1	2	3	4
Nome	registos	*cache*	memória principal	memória de massa
Tamanho típico	< 1 KB	< 16 MB	< 16 GB	> 100 GB
Tecnologia	memória multi-porto, CMOS	CMOS SRAM	CMOS DRAM	disco magnético
Acesso (ns)	0.25–0.5	0.5–2.5	80–250	5×10^6
Largura de banda (MB/s)	2×10^3 – 10^5	5000–10000	1000–5000	20–150
Gerido por	compilador	*hardware*	OS	OS/admin
Backup	*cache*	memória principal	memória de massa	CD/banda

Memória Cache

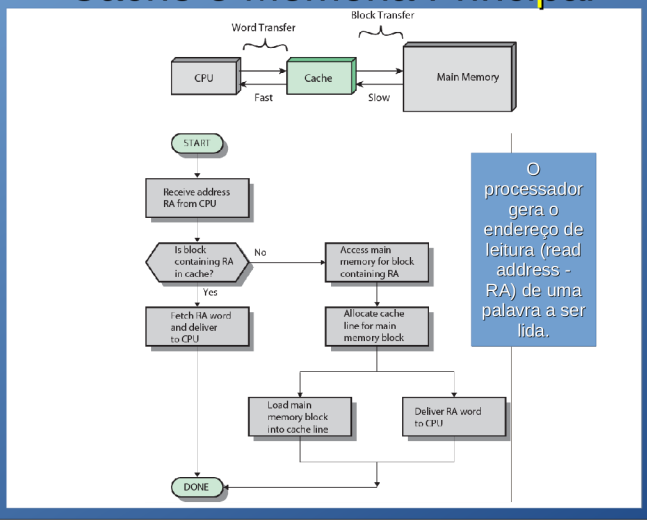
- Princípio da proximidade
  - Programas tendem a reutilizar os dados e as instruções recentemente usados ou aqueles que estão mais próximos na memória
- Referência de Proximidade ou Localidade
  - Temporal - instruções ou dados que serão referenciados outra vez num futuro próximo. (ex. laços, subrotinas, etc..)
  - Espacial - tendência de um processo fazer acesso a itens de endereços próximos. (ex. operações em tabelas ou arranjos que envolvem acessos de uma certa área agrupada no espaço de endereço)
  - Sequencial - programas típicos, a execução de instruções segue uma ordem sequencial.

Funcionamento da Cache

- Quando a CPU necessita de um dado armazenado na memória
  - Se o dado está copiado na cache (***cache hit***),
    - O dado é recuperado da cache e usado
  - Se o dado **não** está na cache (***cache miss***)
    - Isso leva a CPU a buscar este dado na memoria principal.
    - Então este dado é copiado para a memória cache e disponibilizados à CPU
    - Os novos dados que chegam da memória principal ocupam o lugar de outros dados menos usados
- Os dados na cache podem ser alterados pela CPU e assim precisarão ser atrualizados na memória principal

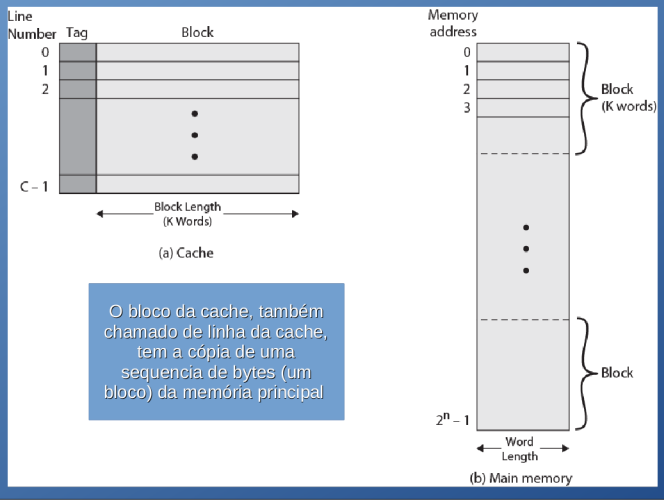
Mesmo se apenas um byte for necessário, um **bloco de muitos bytes contendo o byte** é copiado da memória principal para a cache.

Cache e Memória Principal



O processador gera o endereço de leitura (read address - RA) de uma palavra a ser lida.

Cache/Main Memory Structure



O bloco da cache, também chamado de linha da cache, tem a cópia de uma sequencia de bytes (um bloco) da memória principal

Cache Miss

- Quando a CPU necessita de um dado e o dado **não** está na cache, ocorre uma falha da cache
  1) A CPU buscar este dado na memoria principal;
  2) Um bloco de dados, contendo este dado, é copiado da memória principal para a memória cache;
  3) O dado é disponibilizado à CPU
- Uma falha penaliza o processamento, pois o sistema precisa copiar o bloco do dado para a cache
  - A latência determina o tempo necessário para obter o primeiro elemento do bloco.
  - A largura de banda determina o tempo necessário para obter (transferir) o resto do bloco.

Hierarquia da Cache Intel

- Primeiro nível (L1)
  - Parte para instruções
  - Parte para dados (L1 DCache)
- Segundo nível (L2)
  - Usada para instruções e dados
  - Compartilhada entre processadores lógicos se o processador suporta HyperThreading
- Cache de último nível (LLC)
  - Compartilhada entre os núcleos físicos