Resumo 1

João Roberto da Silva Porto

Ufba - Departamento da Computação - Profa. Mirlei Moura da Silva

Abstract

Resumo dos capítulos 4, 5 e 6 do livro Arquitetura e Organização de Computadores pelo autor William Stallings.aa

1. Capítulo 4

O capítulo 4 discute os mecanismos de implementação de memória em sistemas computadorizados. Elaborando acerca das funcionalidades desempenhadas por essas tecnologias e tanto suas respectivas capacidades quanto suas deficiências.

A construção de modelos eficientes de memória perpassa diferentes crivos e metódos de análise, em razão da grande variedade quanto às formas nas quais esses são instalados em um computador.

1.1. Aspectos Gerais e Aspectos comparativos

A diferenciação entre os tipos de memória se faz mais prática e sólida quando observada segundo critério definidos e comparaveis, sempre levando em conta o contexto prático da aplicação daquela tecnologia que ela integra.

Mais notavelmente são observados aspectos como localização relativa ao processador e placa mãe, capacidade bruta de armazenamento, a quantidade de informação acessável e inserível de uma única vez, a maneira como se acessa os dados armazenados.

Na análise do desempenho de certa unidade de memória são comumente estudados três métricas:

- Latência

  O tempo para escrita ou leitura de uma informação, desde acesso ao endereço até prontidão para leitura e/ou alteração.

- Tempo de ciclo de memória

  Relativo a memória de acesso aleatório é o intervalo entre operações na memória adicionado a duraçãp de uma operação.

- Taxa de transferência

  Velocidade de locomação de uma unidade de memória para dentro e fora do dispositivo.

As memória variam na natureza da tecnologia que as compõem até o nível das características físicas que as definem, sendo essas eleitas segundo necessidades e limitações práticas. As mais comumente empregadas são as memórias semicondutoras e magneto-óptica.

Essas variações implicam em diferentes funcionalidades, cada uma com seus prós e contras.

1.2. Hierarquia de memória

Os três principais eixos de análise para escolha de uma memória no design de um computador giram em torno da capacidade, velocidade e custo.

De maneira geral geral, essas três frentes existem em contradição, ainda que todas sejam, na maioria dos casos, essenciais para uma experiência satisfatória e, sobretudo, eficiente.

Por tanto, são empregados diferentes tipos de memória na construção de um computador, de modo a atender diferentes necessidades ainda mantendo custos monetários numa margem razoável.

Quanto mais próximo à base da hierarquia, maior o armazenamento bruto, menor frequência de uso, menor velocidade de acesso e menor custo por bit.

A pirâmide de memória subdivide-se em 3 categorias principais, cada uma com sua hierarquia interna seguindo os mesmos princípios supracitados. Estes segmentos são:

1. Memórias na placa
2. Memórias fora da placa
3. Memórias off-line

1.2.1. O princípio da Localidade de Referência

Esse modelo de segmentação tem como pilar central o princípio da localidade de referência, isto é a tendência a existência de subrotinas iterativas que implicam na repetição de referências já acessadas, desta forma durante a execução de um programa existe uma tendência natural ao "agrupamento"de operações que partem à execução de um software. Desta maneira, informações armazenadas em memória de hierarquia inferior são gradualmente aglutinadas conforme são acessadas e transferidas para as de maior posição hierarquica. Assim, reduzindo a necessidade de acessos aos segmentos inferiores com o passar do tempo executando um programa.

1.3. Memória cache: Princípios

A memória cache integra a memória principal e é a segunda de maior prioridade na hierarquia de memória. Ela exerce função de intermédio, otimizando a comunicação entre os corpos de maior capacidade da memória principal, como memórias do tipo RAM, e os registradores do processador. Esse processo só é verdadeiramente eficiente graças ao fenômeno da Localidade de Referência.

Esta integração ocorre por meio da subdivisão da cache em diferentes camadas, normalmente 3 chamadas L1, L2 e L3. Estes subníveis comunicam-se sequencialmente, e seus extremos, L1 e L3 comumente, interagem, respectivamente, com o processador e a memória principal.

O acesso e transferência de dados entre a cache e a memória principal dá-se na forma de blocos, um conjunto de palavras, que são por sua vez agrupados nas chamadas linhas dentro da memória cache. Esse processo é dinamizado graças a Localidade de Referência, em razão da tendência de dados relativos uma mesma sub-rotina estarem agrupados em um mesmo bloco. Com base nisto, sempre que uma palavra é exigida pelo processador e não está presente na cache, todo o bloco a quem essa palavra pertence e transplantado pela cache.

Já a comunicação entre a cache e os registradores do processador é conduzida na forma de palavras em uma velocidade muito maior do que as trocas entre níveis da cache, e consideravelmente mais rápida que as transferências entre cache e o resto da memória principal.

1.4.

2. Capítulo 5