Resumo

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

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 métodos de análise, em razão da grande variedade quanto às formas nas quais esses são instalados em um computador.

1.1 Aspectos característicos de diferentes memórias

A diferenciação entre os tipos de memória se faz mais prática e sólida quando observada segundo critério definidos e comparáveis, 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 duração de uma operação.

- Taxa de transferência

  Velocidade de locomoção 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 offline

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 sub-rotinas 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 hierárquica. 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 Elementos do projeto da memória cache

Endereçamento

Uma pratica comum na implementação de quase todos os processadores são tecnologias de endereçamento de memória virtual. Esta trata-se da tradução de células da memória principal em uma representação virtual que possibilite a facilidade aos programas em execução de considerar o armazenamento de um ponto de vista puramente logico sem lidar diretamente com as limitações físicas de capacidade física da memória. Essa tradução normalmente fica a cargo de uma unidade especializada, a MMU.

Existem dois métodos de endereçamento virtual: o lógico, também chamado de virtual, e o físico. Estes diferem nas interações entre cache, MMU, processador e memória principal. Sendo a cache física estacionada entre a MMU e a memória, a custo de velocidade de comunicação com o processador, enquanto a cache virtual, alocada entre a MMU e o processador tem de lidar com conflitos no endereçamento uma vez que os programas tem acesso índices de referencia semelhantes implicando em choques que tem de ser resolvidos pela MMU.

Mapeamento

O processo de transferência de dados nas memórias principais com a cache da-se por intermédio de uma função que mapeia blocos de palavras na memória principal a linhas na cache. Existem diversas funções com esta finalidade, sendo as principais:

- Mapeamento Direto

  Uma função matemática aloca fixamente todos os blocos da memória principal a linhas únicas na cache. Esse processo, apesar de simples, implica em choques entre diferentes blocos de memória alocados as mesmas linhas, o que pode ser especialmente desvantajoso, especialmente quando dois blocos alocados a um mesmo endereço na cache são repetidamente utilizados, incorrendo em trocas desnecessárias constantes entre ambos em razão da pouca dinamicidade desta função.

- Mapeamento Associativo

  Uma função matemática mapeia blocos da memória principal a quaisquer linhas na cache, tomando em conta a disponibilidade utilizando um sistemas de tags para identificação das informações transferidas, desta forma não incorrendo nos mesmos choques. Essa inteligencia na alocação dos blocos ocorre ao custo de uma complexidade comparativamente muito alta no circuito que implementa essa técnica.

- Mapeamento Associativo em Conjunto

Uma função associativa mapeia múltiplos blocos a uma mesma linha, e as organiza internamente segundo um mapeamento direto. Desta forma incorrendo, num custo relativamente intermediário comparado aos demais enquanto ainda evitando choques excessivos.

Algoritmo de Substituição

Na eventualidade da lotação da memória cache, faz-se necessário desocupar espaço para um novo bloco de memória. Evidente que para sistemas baseados em mapeamento direto, esse processo não possui complexidade significativa uma vez que existe somente uma linha possível na cache para cada bloco da memória principal. Quando tratando-se de processadores e caches baseados em mapeamento associativo ou associativo por conjunto tem de se estabelecer critério para exclusão de um conjunto de dados em favor da inserção de um novo. esse critério pode obedecer variados principio de analise, os quais são implantados a nível de hardware.

Alguns exemplos de crivos aplicados em sistemas associativos de conjuntos são:

- LRU - Least Recently Used

  Utiliza-se de um bit adicional USE em cada linha da cache de um dado conjunto relativo a múltiplos blocos de memória, e este sofre alterações de modo a demarcar o seu uso mais recente em comparação as demais linhas referentes ao mesmo bloco de memória. Desta forma, quando necessário, exclui a substituição ocorre na linha menos recentemente utilizada, demarcada pelo USE 0.

- FIFO - First In First Out

  Simplesmente elege-se a linha relativa ao bloco da memória que esta na cache ha mais tempo. Uma implementação comum dessa pratica baseia-se no algoritmo round-robin.

- LFU - Least Frequently Used

  Associa-se um contador a cada linha da memória cache, de forma quantizar os usos daquela dada linha, desta forma elege-se aquela menos frequentemente utilizada.

Politica de Escrita

No momento da substituição de um bloco de memória na cache, atenta-se a qualquer alteração que tenha sido imposta sobre a palavra na cache, essas discrepâncias precisam então serem transplantadas para memória principal. Também pode-se incorrer em circunstancias semelhantes quando lidando com múltiplas CPUs e/ou dispositivos de E/S que tenham capacidade de alterar a memória principal, necessitando uma correção de modo a lidar com as informações dessincronizadas em relação as na memória cache.

Essas operações de correção ocorrem majoritariamente seguindo uma de duas praticas mais comuns:

- Write-Through Replica-se toda operação de escrita entre a memória principal e a memória cache, implicando num fluxo constante de informações e eventuais gargalos, em razão da falta de verificação para casos redundantes, onde a correção não se faz necessária.

- Write-Back Alterações ocorrem somente na cache e, por meio de um sistema de tags usando bits específicos, verifica-se quais blocos sofreram alterações e somente esses são alterados na memória principal. Esse modelo incorre na elaboração de circuitos progressivamente mais complexos especialmente em ambiente com múltiplos dispositivos de E/S que tem acesso a memória.

2 Capítulo 5

As implementações de memórias de acesso rápido evoluíram consideravelmente no decorrer da historia, desde implementações primitivas usando loops magnético ate o padrão atual baseado em tecnologias semicondutoras.

Também conhecidas como memórias Cores, nome herdade pela estrutura em núcleos das primeiras projeções das memória RAM, apesar de a maioria ser baseada nas mesmas tecnologias semicondutoras, elas ainda variam significantemente em suas funcionalidades e qualidades, de acordo com a suas respectivas arquiteturas internas implementadas. Seno alguns tipos de memória principal que iremos nos aprofundar as: ROM, DRAM e SRAM.

2.1 memória principal: Estrutura básica

A unidade mais básica de uma memória Core é a célula de memória, entre todas suas implementações, define-se pela capacidade de armazenar um estado binário e de ser lida e/ou escrita, uma ou mais vezes.

Esses processos são viabilizados por estruturas de comunicação, escrita e controle da memória principal. Normalmente possuindo ao menos 3 interfaces:

- Transporte

  Responsável por de fato armazenar o valor binário

- Controle

  Delegado a verificação e seleção da célula onde ocorrera alguma operação

- Escrita e Saída

  Encarregado de transformar o valor binário no terminal de transporte e/ou externar o valor armazenado por ele.

Tipos de memória Core

A memória RAM (Random Access Memory), apesar de sua descrição não se trata da única Core com acesso randomizado. Sendo, na verdade, sua característica definidora suas altas velocidades de escrita e leitura e sua volatilidade, isto 'e sua necessidade de alimentação energética constante para armazenamento de dados.

As duas principais implementações da RAM são:

- DRAM

  DRAM, ou Dynamic RAM, utiliza de um sistemas comparativamente simples de transistores cujo sinal por si só de um deles define diretamente os valores armazenados numa célula da Core. Esse processo requer um *Refresh*, uma restauração auxiliar dos sinais armazenados esporadicamente em razão da dissipação natural das cargas.

- SRAM

  Alternadamente, existem as SRAM, ou Static RAM, cujo design utiliza de circuitos *flip-flops* para conservar os sinais de um dado conjunto de celular, de forma a não necessitar de um circuito dedicado para o *Refresh* de cada uma delas.

Desta maneira, as DRAM incorrem custos inicialmente mais baixos em razão da maior simplicidade de seu circuito em comparação à uma SRAM que necessita de estruturas mais complexas para conservação de dados. Entretanto, as circunstancias se invertem, a medida que o tamanho da Core aumenta, os transistores necessários para os *Refresh* excedem significantemente mais do que os *flip-flops* precisos.

Em contraste a simetria percebida nas operações nas RAM, as ROM (Read Only Memory) permitem uma velocidade maior ou comparável as RAM e, também, não são voláteis. Normalmente, a implementação das ROM da-se atraves da gravação física dos dados nos transistores. Essa falta de dinamicidade aparente, apesar de limitante sob certas ópticas, faz-se irrelevante e até útil em muitas aplicações, como sistemas embarcados, bibliotecas e funções constantes, entre outros processos da mesma natureza.

As ROM também são encontradas em diversas variações, como:

- ROM

  ROM. ou Read Only Memory, tem como principal característica a escrita física dos dados armazenados ainda na fabricação da memória, num processo custoso e, em termos gerais, irreversível. Permitindo acesso rápido a dados sem a necessidade de consultar a memória secundaria e sem necessidade de alimentação elétrica.

- PROM

  PROM, ou Programmable ROM, uma variação mais "acessível"da ROM, ela tem a escrita de dados efetuada por um processo elétrico por maquinário especializado já nas mãos do consumidor final. Apesar do custo aparentemente mais acessível em conjunto com as vantagens das ROM comuns, os preços das PROM não escalam bem em comparação com as ROM.

- EPROM

  Uma alternativa mais prática da PROM, tem sua escrita realizada eletrônica, e ainda permite sua reutilização, por meio de um processo de apagamento utilizando raios ultravioleta. Desta forma, sendo possivelmente mais acessível que as PROM, entretanto escalam, em termos de produção em massa, pior que as PROM.

- EEPROM

Semelhante a EPROM, a escrita e o apagamento são ambos feitos eletricamente, facilitando e barateando essa etapa do processo, as custas de valores ainda maiores quando aplicadas a produção em larga escala.

- Flash

  Popularizada na forma de dispositivos de armazenamento móvel, contrasta-se das demais por servir como uma alternativa intermediaria de preço acessível entre EPROM e EEPROM.

Detecção de erros

Dispositivos de memória Core são sujeitos a erros, seja por razoes de erro na fabricação por exemplo, onde o reparo desse dano implicaria em um processo muito dispendioso e excessivamente caro, ou por motivos da eventualidade físico-eletrônica que podem vir ocorrer naturalmente, temporariamente forçando um dado transistor em um sinal ou variando entre eles.

Nestes casos, existe uma serie de praticas comum entre as memórias principais que visam verificar a existência, diagnosticar a natureza e, caso possível, reparar as discrepâncias indevidas na memória.

Soluções do tipo DRAM

A fim de solucionar os maiores tempos de espera das memórias do tipo DRAM, ainda tentando aproveitar seus valores de produção mais acessíveis, novas implementações desse tipo de memória foram elaboras.

- DRAM Síncrona

  Por meio de um sinal externo de clock, a DRAM síncrona comunica-se com o processador transmitindo informações a cada clock, utilizando o intervalo dessas interações para operações de controle e organização dos dados.

- DRAM RamBus

  Atraves de um barramento especial, memórias do tipo RDRAM, comunicam-se, ainda que de forma assíncrona, em altas velocidades com o processador.

- DDR-SDRAM

  A mais utilizada entre computadores pessoais, trata-se de uma versão melhorada da SDRAM que efetuar operações de escrita e leitura para o processador duas vezes por clock, tanto na alta, quanto na baixa do sinal. Efetuando a execução de processos de controle no intervalo entre essas variações e utilizando de um buffer próprio para agilizar o processo.

3 Capitulo 6

Mecanismos de memória externa existem em resposta ao problema ao problema da volatilidade das memórias principais e as limitações de capacidade destas, tal qual ditado pela ordem da hierarquia de memória, permitindo um armazenamento bruto mais significativo ainda sob um custo por bit acessível.

As implementações das memórias secundarias variam significantemente quanto a natureza das tecnologias que as compõem, desde maquinas baseadas em processos magnéticos, eletrônicos e ópticos, cada uma com suas respectivas utilidades, historias e limitações.

3.1 Disco magnético: Funcionamento

Uma estrutura análoga ao vinil, entretanto constituído de superfícies, em sua maioria, magnéticas, baseia no surgimento de campos eletromagnéticos a partir da rotação de um disco magnético que entra um contato com uma "cabeça"formada por ao menos uma bobina condutora. Desta forma o contato entre a bobina e o disco em meio ao campo magnético produz pulsos elétricos característicos e identificáveis que serão interpretados pelo computador num processo de leitura de dados, e, num contexto de escrita, a bobina produz o pulso em questão que sobrescreve os padrões magnéticos demarcando pulsos no disco.

Organização e capacidade

O Disco magnético subdivide-se em circunferências concêntricas conhecidas como trilhas, separadas entre si por lacunas progressivamente menores a medida que aproximam-se dos extremos do disco, e, por sua vez, essas trilhas são separadas em setores.

Essas subdivisões e espaços vazios tem como função prover uma velocidade de acesso uniforme a informações em todo disco, independente da distancia dos bits em questão do centro. Ancorando-se em uma distribuição que vise separe um mesmo bloco de dado pela máxima extensão do raio e, simultaneamente, preservando igualdade entre a velocidade angular relativa a quaisquer trilhas no disco.

A capacidade do disco é então determinada pela densidade de bits em relação ao raio do disco.

Aspectos físicos e parâmetros de desempenho

Os discos magnéticos podem ainda diferir entre si, de acordo com uma serie de condições. Dizendo respeito ao numero de cabeças, portabilidade e reutilização.

Os principiais crivos de analise entre discos magnéticos comparáveis são relativos as velocidades para busca de uma trilha, inversão do sentido da rotação e a extensão da transferência de dados.

RAID

RAID, ou *Redundant array of independent disks* diz respeito ao processo de utilização de múltiplos discos magnéticos como uma única memória secundária, aproveitando os discos para fins de redundância de dados, aumentando consideravelmente a segurança da informação

e a velocidade de acesso e escrita dessas informações. Esse processo faz-se possível por meio, não de uma ordem hierarquia incorporada aos múltiplos níveis de uma RAID, mas sim pela implementação de diversos modelos de redundância baseada em diferentes princípios arquitetônicos, de forma a trazer as qualidades respectivas de cada um dos modelos.

3.2 Dispositivos ópticos

O CD alcançou proeminência como central ao processo de massificação do consumo de musica em sua época, permitindo um compartilhamento e leitura baratos e de alta qualidade. Tal processo fez-se possível graças a implementação de uma memória baseada em tecnologia de gravação a laser, a qual gerava sulcos no corpo do disco óptico no processo de gravação, e utilizava informações recebidas a partir da reflexão nos sulcos da memória secundaria em questão.

Vale relembrar também o advento do DVD, que implementou uma série de melhorias no método de organização das subdivisões de dados no corpo do disco, possibilitando o compartilhamento também de mídias visuais em alta qualidade.