APOSTILA 4 – SUBSISTEMA DE MEMÓRIA

Prof. Murilo Parreira Leal, M.Sc.

Arquitetura e Organização de Computadores

1 – INTRODUÇÃO

A memória é o componente de um sistema de computação cuja função é armazenar as informações (por informações entendem-se os dados ou as instruções de um programa – Von Neumann) que são, foram ou serão manipuladas pelo sistema. Na prática, a memória de um computador possui tantas variedades (velocidade, capacidade de armazenamento, tecnologia, etc.) que se torna um **subsistema** de elementos hierarquicamente estruturados.

No caso de uma memória de computador, o elemento a ser manipulado é o *bit*, o qual, em grupo de *n* bits, corresponde a uma unidade de informação a ser armazenada, transferida, recuperada, etc. Para isso, realizam-se ações de *armazenamento* (transferência de bits de outro componente (UCP, HD, etc.)) ou *recuperação* (transferência de bits para outro componente (UCP, HD, etc.)). O *armazenamento* pode ser chamado de "*escrita*", *"gravação"* ou *"write"*, enquanto a *recuperação* pode ser chamada de "*leitura*" ou *"read*". A gravação é destrutiva, ou seja, os dados que estavam gravados anteriormente são substituídos pelos que estão sendo gravados. Por outro lado, a recuperação apenas copia o valor armazenado para outro local. O valor original continua sem alteração.

Para que a memória possa ser armazenada em uma memória (escrita) é necessário que seja definido um local disponível identificado de alguma forma precisa e única (um número, por exemplo). O número ou código associado ao local é o *endereço* ("*address*") e irá permitir que a informação possa ser localizada.

Observação importante: Diferentemente de uma caixa de correio ou biblioteca, o local indicado por um endereço sempre estará preenchido por sinais elétricos (0 ou 1).

2 – MÉTODOS DE ACESSO

- **Acesso Seqüencial:** Os dados são organizados na memória em unidades chamadas de registros. O acesso é feito segundo uma seqüência específica. O tempo de acesso depende da posição relativa do registro, variando significativamente. Exemplo: Fita magnética.

- **Acesso direto:** Por meio de uma pesquisa seqüencial em uma vizinhança do registro é obtido o seu endereço físico, sendo então é possível a leitura ou gravação. O tempo de acesso também é variável. Exemplo: Disco magnético (HD).

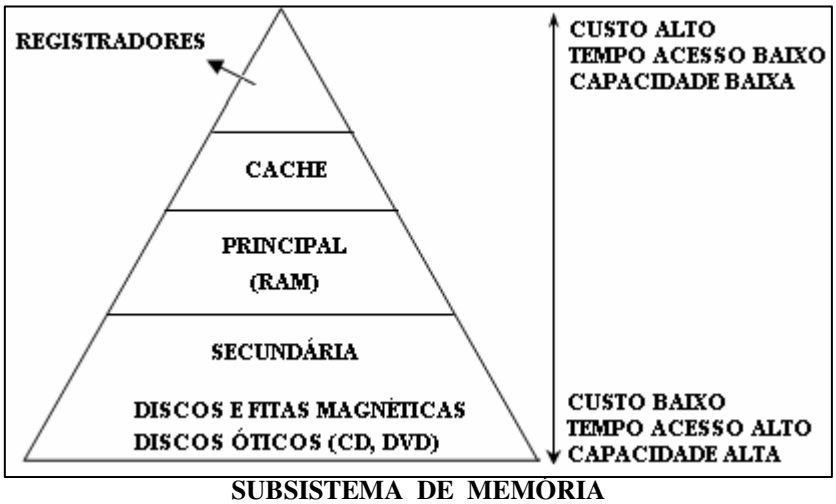
- **Acesso aleatório:** Cada posição de memória possui mecanismo de endereçamento fisicamente conectado a ela. O tempo de acesso é o mesmo para todos os endereços. Exemplo: RAM.

- **Acesso Associativo:** Um dado é buscado na memória com base em uma parte de seu conteúdo, e não de acordo com seu endereço. Exemplo: Memória CACHE.

3 – HIERARQUIA DE MEMÓRIA

Para certas atividades é fundamental que a transferência da informação seja a mais rápida possível (menor "Tempo de Acesso"), enquanto a quantidade de bits ("Capacidade") a ser manipulada pode ser pequena. Em outras situações, o "Tempo de Acesso" não é tão importante, mas sim o volume de dados gravado. A permanência da informação após o desligamento do computador é outra característica relevante em algumas situações.

Em todos os casos, o custo da memória é inversamente proporcional à "capacidade" e ao "tempo de acesso". Assim sendo, para o correto funcionamento de um computador verifica-se a necessidade de diferentes tipos de memória. Este conjunto de diferentes memórias é chamado "Subsistema de Memória".



SUBSISTEMA DE MEMÓRIA

Importante lembrar que os discos e fitas magnéticas (memória secundária) também serão considerados como memória no presente capítulo.

3.1 – Parâmetros

Obs.: Nessa apostila os parâmetros "Capacidade" e "Custo" têm sido atualizados freqüentemente, porém o "Tempo de Acesso" foi mantido constante para determinada geração de componentes. Isso ocorre porque a sua atualização exigiria novos cálculos em vários exemplos utilizados, principalmente em relação à memória CACHE. Última atualização – 10/09/2007.

- **CAPACIDADE:** É a quantidade de informação que pode ser armazenada em uma memória. A medida básica é o byte, embora também possam se usadas outras unidades como células (memória principal ou CACHE), setores (discos) e bits (registradores). Evidentemente é possível se utilizar os multiplicadores K (kilobyte), M (megabyte), G (gigabyte) ou T (terabyte).

Registradores:	100 b
CACHE L1:	16 KB
CACHE L2:	512 KB
RAM:	256 MB
HD:	80 GB

- **TEMPO DE ACESSO:** Indica quanto tempo a memória gasta para colocar uma informação na barra de dados após uma determinada posição ter sido endereçada. Varia de acordo com o tipo de memória analisado:

Registradores:	0.1 ns
CACHE L1:	1 ns
CACHE L2:	2 a 5 ns
RAM:	10 a 20 ns
HD:	10 ms

  O Tempo de Acesso para diferentes endereços em memórias eletrônicas (RAM, ROM) é igual; já em discos e fitas, que são dispositivos eletromecânicos, o Tempo de Acesso depende da localização do dado em relação ao último acesso. Após um acesso, algumas memórias podem impedir por um pequeno intervalo de tempo, o uso do sistema de memória para um novo acesso. Este intervalo de tempo somado ao Tempo de Acesso, é o Ciclo de Memória (Memory System´s Cycle Time), que é o período de tempo decorrido entre duas operações sucessivas de acesso. Se a capacidade da memória for excedida em um microcomputador, o programa será executado a partir de uma área de disco chamada "área de troca" ou "swap". A velocidade do processamento será, portanto muito prejudicada por causa dos diferentes "Tempos de Acesso" (entre 1.000.000 e 10.000.000 vezes mais lento).

- **VOLATILIDADE**: Podem ser do tipo volátil ou não-volátil. A memória não-volátil mantêm a informação armazenada quando a energia elétrica é desligada. Todo computador tem que possuir uma certa quantidade de memória não volátil (EPROM) para carregar o microprocessador quando o computador é ligado. A memória secundária também é do tipo não-volátil (discos, fitas, etc).

- **TECNOLOGIA:** Semicondutores (transistores C-MOS, HC-MOS, TTL), meio magnético (HD, fitas, etc.), meio ótico (CD-ROM, CD-RW).

- **TEMPORARIEDADE:** Indica o conceito de tempo de permanência da informação de um dado tipo de memória. Alguns dados ficam armazenados por um longo período de tempo (no HD, por exemplo), sendo chamada de memória permanente. Outros, ficam por um curto intervalo de tempo, como nos registradores, por exemplo, sendo por isso chamada de memória transitória.

- **CUSTO:** A unidade não pode ser absoluta (US$). Utiliza-se neste parâmetro o custo por capacidade das unidades de armazenamento (US$/MB, US$/GB, etc.). Quando for necessária a comparação de parâmetros de custo deve-se utilizar a mesma unidade de armazenamento (MB, GB, etc.) em todos eles. (Cotação do dólar: R$ 2,00/US$):

CACHE:	256 KB	US$ 20,90 →	85606,40 US$/GB
RAM:	512 MB	US$ 44,50 →	89,00 US$/GB
HD:	80 GB	US$ 77,50 →	0,97 US$/GB
CD-ROM:	700 MB	US$ 0,50 →	0,73 US$/GB
DVD-ROM:	4.2 GB	US$ 0,80 →	0,19 US$/GB
FITA DAT:	40 GB	US$ 30,00 →	0,75 US$/GB

4 – REGISTRADORES

Através das portas lógicas podemos fabricar memórias. As memórias podem ser estáticas **(SRAM)** ou dinâmicas **(DRAM)**. As memórias estáticas (SRAM) mantêm o dado armazenado enquanto seus "flip-flops" estiverem alimentados. As dinâmicas (DRAM) também dependem da alimentação, mas, além disso, precisam de um "pulso" de tensão de tempos em tempos para restaurar os dados **(pulso de refresh)**.

As memórias dinâmicas (DRAM) necessitam de correntes elétricas baixíssimas, gastando menos energia e, portanto, gerando menos calor. Podem por isso, ser mais compactadas (maior quantidade de bytes por mm^2 no chip) e também, mais baratas. São utilizadas na Memória Principal do computador.

As memórias estáticas (SRAM) são bem mais rápidas, porém mais caras. Não necessitam do pulso de refresh. Os registradores são memórias estáticas (SRAM), por permitirem de alta velocidade de processamento. São utilizadas apenas na Memória CACHE e na UCP (microprocessador).

Existem vários registradores especiais no microprocessador (UCP), e serão estudados no próximo capítulo. Neste momento, precisamos conhecer apenas dois:

1. **RDM–Registrador de Dados da Memória (MBR–Memory Buffer Register):** Armazena temporariamente a informação que está sendo transferida da MP para a UCP (leitura) ou da UCP para MP (escrita). Possui a mesma quantidade de bits do barramento de dados.
2. **REM–Registrador de Endereços da Memória (MAR–Memory Address Register):** Armazena temporariamente o endereço de acesso a uma posição de memória, ao se iniciar a operação de leitura ou escrita. Em seguida, o endereço é encaminhado à área de controle da MP para decodificação e localização da célula desejada. Possui a mesma quantidade de bits do barramento de endereços.

4.1 – Características dos Registradores

Tempo de Acesso:
- Ciclo de Memória é igual ao Tempo de Acesso.
- 0.1 a 0.5 ns.

Capacidade:
- 50 b a 800 b (100 B).

Volatilidade:
- Voláteis.

Tecnologia:
- Memórias Estáticas (SRAM).

Temporariedade:
- Grande, ou seja, o dado é muito temporário.

Custo:
- O mais caro de todos os tipos de memória.

5 – MEMÓRIA CACHE

O Ciclo de Memória da Memória Principal (MP) é bem mais demorado que o tempo gasto pela UCP para fazer uma operação. Para resolver isso, foi desenvolvida uma técnica que é a inclusão de uma memória (CACHE – pronuncia-se CASH) entre a UCP e a MP, otimizando a transferência. Esta memória é construída com a mesma tecnologia da UCP (registradores), tendo Tempo de Acesso compatível, diminuindo o tempo de espera (WAIT STATE) da UCP para receber o dado ou instruções.

A tecnologia dos microprocessadores vem dobrando o desempenho a cada 18 ou 24 meses, o que não ocorre com as memórias DRAM (RAM Dinâmicas), que vem aumentando pouco de ano para ano. É importante ressaltar que as diversas versões de DRAM que vêm surgindo no mercado não modificam substancialmente o problema da diferença de desempenho delas em relação a UCP, ou seja, as CACHE continuam necessárias porque os microprocessadores continuam ainda mais rápidos que as memórias.

A memória CACHE está sempre repleta de dados ou instruções, visando otimizar o relacionamento entre a UCP e a MP. Uma vez que ela é muito menor do que a MP, o critério utilizado para seu preenchimento está baseado no Conceito da Localidade. Uma parte da CACHE fica constantemente preenchida com os programas mais utilizados do Sistema Operacional.

Os microprocessadores atuais já vêm com uma pequena CACHE embutida (Level 1 ou L1) e possuem outra, a principal, externa (Level 2 ou L2). Algumas vezes existe ainda uma L3. Estas CACHEs internas aos microprocessadores trabalham numa freqüência de clock muito mais alta do que a memória RAM, por exemplo: Pentium IV 3 GHz (placa-mãe 400 MHz) tem CACHE L1 operando a 3 GHZ e CACHE L2 e memória RAM operando a 400 MHz (800 MHz se essa última for DDR).

5.1 – Justificativa da Utilização da Memória CACHE

A simples existência de uma memória de alta velocidade entre a RAM e o microprocessador faria, na realidade, aumentar o tempo de acesso.

Os conceitos de **"Localidade"** justificam a existência da CACHE por prever vários acessos à mesma memória num breve espaço de tempo:

- **Localidade Temporal:** Se um programa acessa um dado (endereço), existe uma boa probabilidade que ele venha a acessá-la novamente, em breve.
- **Localidade Espacial:** Se um programa acessa um dado (endereço), existe uma boa probabilidade que ele venha a acessar os dados armazenados em sua proximidade, em breve.

Baseado nestes conceitos, a memória CACHE trabalha seguindo o seguinte algoritmo:

1. Sempre que a UCP vai buscar uma nova instrução, ela acessa a memória CACHE.
2. Se a instrução ou dado estiver na CACHE (**ACERTO ou HIT**), ela é transferida em alta velocidade para a UCP.
3. Se a instrução ou dado não estiver na CACHE (**FALTA ou MISS**), então o sistema está programado para interromper a execução do programa e transferir a instrução desejada da MP para a UCP. A UCP entra em "Estado de Espera" (**WAIT STATE**) enquanto ocorre a demorada transferência do dado vindo da MP. Simultaneamente é transferida uma cópia da instrução desejada mais o conteúdo de alguns endereços de memória subseqüentes para a memória CACHE, prevendo novo acesso baseado no princípio da localidade espacial.

5.2 – Algoritmos de Substituição de Dados na Memória CACHE

Por outro lado, os dados armazenados na CACHE devem ser substituídos de tempos em tempos, permitindo a ocorrência de mais acertos. Esta substituição obedece um dos seguintes algoritmos:

- **LRU – Least Recently Used** – O sistema escolhe o bloco de dados que está há mais tempo sem ser utilizado. (Critério mais utilizado)
- **FILA (FIFO)** – O sistema escolhe o bloco que está armazenado há mais tempo.
- **LFU – Least Frequently Used** – O sistema escolhe o bloco que tem tido menos acesso por parte da UCP.
- **Escolha aleatória** – O sistema escolhe o bloco a ser substituído aleatoriamente.

O objetivo destes algoritmos é aumentar o número de acertos, evitando as faltas. Estudos indicam que o último algoritmo (escolha aleatória) reduz muito pouco o desempenho do sistema em comparação aos demais, e é extremamente simples de implementar.

5.3 – Políticas de Escrita pela Memória CACHE

Em sistemas com memória CACHE, toda vez que a UCP realiza uma operação de escrita, esta ocorre imediatamente na CACHE. Como a CACHE é apenas uma memória intermediária, e não a principal, é necessário que em algum momento a MP seja atualizada para que o sistema mantenha sua integridade.

Problemas:

I. A MP pode ser acessada tanto pela CACHE quanto por dispositivos de E/S utilizando DMA. Neste caso, o sistema tem que perceber se o dispositivo de E/S alterou diretamente a MP (através do DMA) e por isso é a CACHE que se encontra desatualizada, ou o contrário, a CACHE foi alterada e por isso a MP tem que ser atualizada.

II. A MP pode ser acessada por várias UCP, cada uma contendo uma memória CACHE. É possível que a MP tenha sido alterada atendendo a solicitação da memória CACHE de uma UCP, enquanto as demais CACHEs permanecerão desatualizadas.

Existem "técnicas" ou "políticas" de escrita para evitar os problemas citados:

a) Escrita em ambos ("**Write Through**"): A escrita de uma palavra na CACHE acarreta escrita na MP. Existindo outros módulos UCP/CACHE, estes também serão alterados garantindo a integridade.

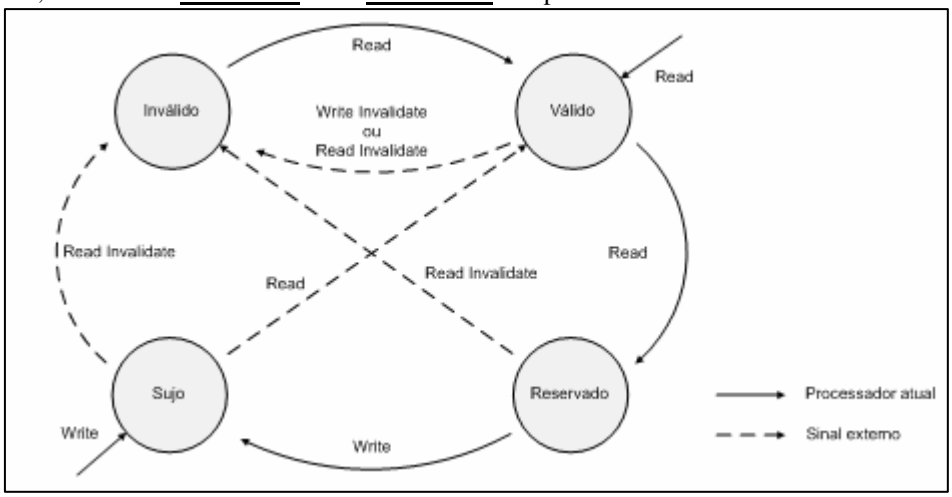
b) Escrita somente no retorno (**"Write Back"**): Sempre que há uma alteração na CACHE, é marcado um bit denominado ATUALIZA. Quando um bloco armazenado for substituído, o sistema verifica o bit ATUALIZA; caso seja 1, o bloco é armazenado na MP; se não for, é descartado.

c) Escrita uma vez (**"Write Once"**): Foi o primeiro protocolo com política de **"Write Invalidate"**, sendo uma mistura de **"Write Back"** e **"Write Through"**. São considerados quatro possíveis estados para um bloco de CACHE:

- **INVÁLIDO:** o bloco é inconsistente;
- **VÁLIDO:** a cópia do bloco é consistente com a cópia da memória;
- **RESERVADO:** o bloco foi escrito exatamente uma vez e essa cópia é consistente com a memória, que é a única outra cópia desse bloco;
- **SUJO:** o bloco foi modificado mais de uma vez e essa cópia local é a única no sistema.

Se o estado do bloco é **VÁLIDO**, a alteração da informação é realizada na CACHE local e na memória (**Write Through**). Se o estado do bloco é **SUJO**, a alteração é feita apenas localmente, atualizando a memória apenas quando o bloco é reposto ou quando outra CACHE tentar ler esse bloco (**Write Back**).

Além de ler e escrever um bloco na memória, esse protocolo requer do hardware duas outras operações: **"Write Invalidate"**, que invalida todas as outras cópias de um bloco, e **"Read Invalidate"**, que lê um bloco e invalida todas as outras cópias. Se um bloco de CACHE está **INVÁLIDO**, ele não pode ser utilizado, ou seja, caso seja feita uma leitura ou escrita nesse bloco, ocorre um **"Read Miss"** ou **"Write Miss"** respectivamente.



- **Reposição:** Se a cópia está suja, então ela deve ser escrita de volta na memória. Caso contráario o bloco pode ser simplesmente descartado para a alocação de outro bloco.
- **Read Hit:** Realiza a leitura normalmente e não ocorrem mudanças de estado.
- **Read Miss:** Se nenhuma cópia suja do bloco existir, então a leitura é realizada normalmente, pois a memória possui uma cópia consistente e essa cópia na CACHE recebe o estado de válido. Se alguma cópia suja existir, então o controlador da CACHE que possui essa cópia inibe a memória e envia uma cópia desse bloco no barramento. A memória e a CACHE requisitante recebem essa cópia, de forma que as três cópias ficam consistentes, e o estado do bloco nas duas CACHEs vira válido.

- **Write Hit:** Se a cópia está no estado reservado ou sujo, o bloco é alterado localmente e o novo estado do bloco é sujo. Se o estado for válido, então um "Write Invalidate" é enviado em broadcast para todas as CACHEs, invalidando todas as outras cópias desse bloco. Uma cópia do bloco alterado é enviada para a memória e o estado desse bloco na CACHE é reservado.
- **Write Miss:** Se existir alguma CACHE com uma cópia suja, a leitura é feita dessa CACHE, que atualiza a memória principal. Caso contrário, a leitura é feita diretamente da memória principal. Essa leitura é feita com o comando de "Read Invalidate", que além dessa leitura, invalida todas as cópias do bloco. A cópia do bloco é alterada localmente na CACHE e o estado resultado é sujo.

Conclusões:

1. A política **"Write Through"** pode provocar uma grande quantidade de escritas desnecessárias na MP, com a natural redução e desempenho do sistema.
2. A política **"Write Back"** minimiza a desvantagem anterior, mas a MP fica potencialmente desatualizada para utilização por outros dispositivos a ela ligados, como módulos de E/S, o que os obriga a acessar o dado através da CACHE, o que neste caso é um problema.
3. A política **"Write Once"** pode ser conveniente mas apenas para sistema com múltiplas UCP, não sendo ainda muito usado.
4. Estudos realizados apontam para uma percentagem de escrita baixa na memória, da ordem de 15%, apontando para a simples política do **"Write Through"** como melhor método.

5.4 – Mapeamento de Memória Cache

É preciso lembrar que a memória CACHE é muito menor do que a MP então obviamente, a primeira não possuirá a mesma quantidade de endereços da segunda. Desta forma é necessário armazenar na CACHE não só uma cópia do dado existente na MP, como também o seu endereço. Assim sendo, é necessário um algoritmo para mapear os blocos da memória principal em linhas da memória CACHE.

Existem três alternativas atualmente disponíveis:

- Mapeamento direto;
- Mapeamento associativo;
- Mapeamento associativo por conjuntos.

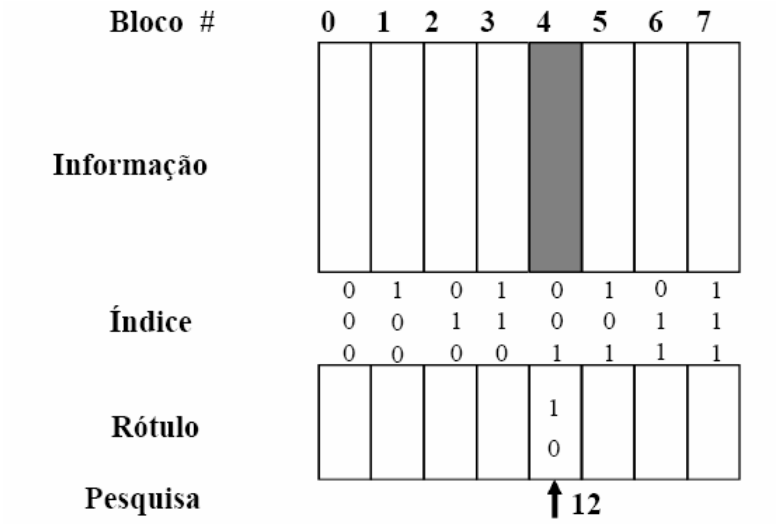
5.4.1 Mapeamento direto

Cada bloco da MP tem uma linha da CACHE previamente definida para ser armazenada. Como há mais blocos do que linhas de CACHE isso significa que muitos blocos irão ser destinados a uma mesma linha, sendo preciso definir a regra a ser seguida para a escolha da linha específica de cada bloco. A maioria das CACHEs mapeadas diretamente usa o seguinte processo de mapeamento:

(endereço do bloco) MOD (número de blocos da *CACHE*).
Onde MOD é a "divisão inteira"

Para saber se um dado está na CACHE basta comparar o rótulo. O rótulo contém informações sobre um endereço de memória, que permite identificar se a informação desejada está ou não na CACHE. Ele só precisa conter a parte superior do endereço, correspondente os bits que não estão sendo usados como índice da CACHE. Os índices são usados para selecionar a única entrada da CACHE que corresponde ao endereço fornecido.

Consideremos uma MP com 32 endereços, variando de 00000 e 11111. A memória CACHE possui apenas 8 endereços (índices), variando de 000 a 111. Neste caso, são necessários três bits para indexar a memória CACHE e os dois bits restantes são utilizados como rótulos.



Suponha que desejamos saber se o bloco com endereço de memória principal 12 (01100) está contido na memória CACHE. Assim, para o endereçamento temos 12 módulo 8 = 4. O bloco 12 só pode estar armazenado no bloco 4 da memória CACHE. Basta então comparar os dois bits mais significativos do bloco 12 (01) com o rótulo do bloco 4 da memória CACHE para saber se o conteúdo está presente. Na figura acima, no bloco 4 (índice 100) da memória CACHE só podem ficar armazenadas as palavras da memória principal com índice igual a 4 nos seus três bits menos significativos. Isto é, 4 (00100), 12 (01100), 20 (10100) e 28 (11100).

5.4.2 Mapeamento Associativo

As CACHEs mapeadas por associatividade possuem um bloco maior do que o conteúdo de uma memória. Neste esquema, ao ocorrer uma falha (MISS), várias palavras adjacentes são trazidas para a CACHE. Este fato aumenta a probabilidade de ocorrer um acerto (hit) nos próximos acessos. As CACHEs mapeadas por associatividade usam o seguinte processo de mapeamento, o conjunto que contém o bloco de memória é dado por:

(endereço do bloco) MOD (número de conjuntos da *CACHE*).
Onde MOD é a "divisão inteira"

As CACHEs mapeadas por associatividade por ser classificadas em "**associativas por conjuntos**" ou "**totalmente associativas**".

O número total de rótulos nas CACHEs mapeadas por associatividade é menor do que nas CACHEs mapeadas diretamente porque são usados para todas as palavras que estão contidas num mesmo bloco e que formam um conjunto. Estas memórias são chamadas memórias CACHE associativas por conjunto (set associative CACHE memory). Este compartilhamento contribui para um uso mais eficiente do espaço disponível para armazenamento na CACHE.

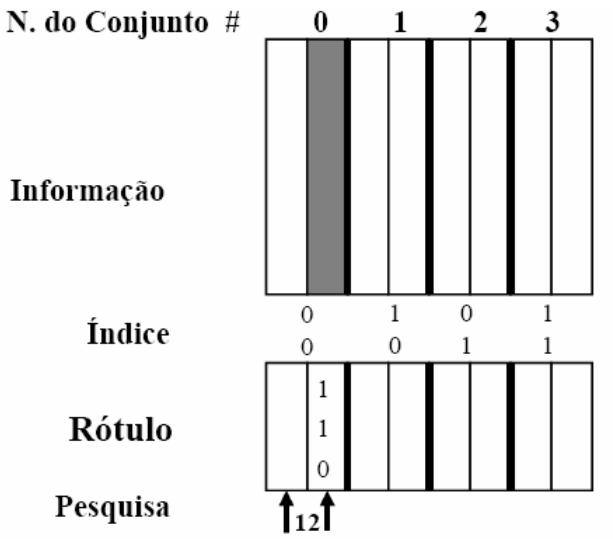
Uma memória CACHE **associativa por conjunto** com n linhas em cada conjunto é chamada CACHE n-associativa por conjunto (n-way set associative CACHE). As memórias CACHE associativas por conjunto são as encontradas na maioria dos processadores. É importante ressaltar que o tamanho de linha, o número de linhas por conjunto e o número de conjuntos podem variar, resultando em configurações diferentes.

A memória CACHE **associativa por conjunto**, ao receber o endereço de uma locação de memória, interpreta o endereço da forma descrita a seguir.

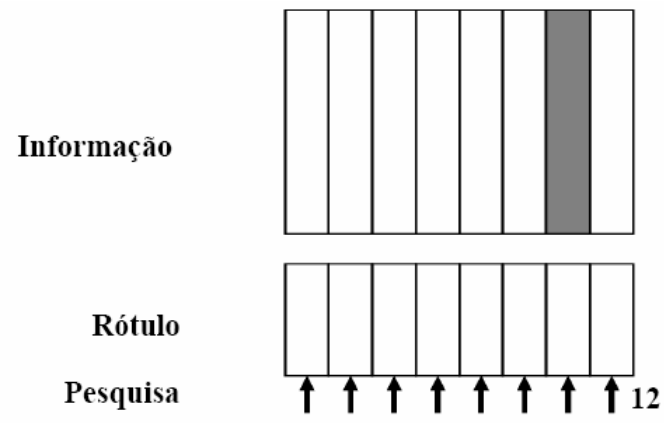
O endereço é logicamente dividido em três campos: o campo byte, formado pelos bits menos significativos; o campo conjunto, formado pelos n bits seguintes; e o campo rótulo, composto pelos m bits mais significativos do endereço. A memória CACHE usa os bits do campo byte para selecionar um byte específico dentro de uma linha. O campo conjunto é usado para selecionar um dos conjuntos, enquanto o campo rótulo é usado para verificar se o dado referenciado se encontra em alguma das linhas do conjunto selecionado.

Num acesso a memória CACHE seleciona primeiro o conjunto, e em seguida compara o rótulo do endereço recebido com os rótulos armazenados nas entradas de diretório do conjunto selecionado. O bloco pode ser colocado em qualquer elemento deste conjunto. Se o rótulo no endereço coincide com algum dos rótulos no diretório, isto significa que o bloco com o byte referenciado encontra-se na linha associada à entrada do diretório que contém o rótulo coincidente. Esta linha é então selecionada e o byte dentro desta linha é finalmente acessado.

Consideremos a mesma MP anterior com 32 endereços, variando de 00000 e 11111. A memória CACHE possui 4 conjuntos, variando de 00 a 11.



Desejamos saber se o bloco com endereço de memória principal 12 (01100) está contido na memória CACHE. A memória CACHE é formada por quatro conjuntos de duas linhas. Assim, para o endereçamento temos 12 módulo 4 = 0. O bloco 12 só pode estar armazenado no conjunto 0 da memória CACHE. Basta então comparar o rótulo do bloco 12 (011) com os rótulos do conjunto 0 da memória CACHE para saber se o conteúdo está presente.



Uma memória CACHE **totalmente associativa** pode ser vista como uma CACHE onde existe um único conjunto. O bloco da memória principal pode ser colocado em qualquer um dos elementos deste conjunto.

Suponhamos ainda o mesmo exemplo de uma memória principal com 32 palavras e que desejamos saber se o bloco com endereço de memória igual a 12 está contido na memória CACHE. Neste caso, como o bloco pode ser colocado em qualquer elemento da CACHE, é necessário pesquisar todos os blocos.

A não coincidência do rótulo do endereço com os rótulos armazenados no conjunto indica que o byte referenciado não se encontra na CACHE (ou seja, ocorreu um MISS).

Neste caso, a lógica de controle da memória CACHE se encarrega de copiar o bloco apropriado a partir da memória principal. Este bloco será armazenado em uma das linhas do conjunto indicado pelo endereço. Uma vez que o bloco tenha sido carregado na memória CACHE, o acesso prossegue como descrito anteriormente.

5.5 – Características da CACHE

Tempo de Acesso:
- 1 a 5 ns.

Capacidade:
- São encontrados valores entre 64 KB e 2 MB. O valor típico é 256 KB, a partir da geração do Pentium IV.
- Não são muito grandes porque são muito caras. Alguns processadores são construídos com menos CACHE para diminuir o seu preço (CELERON/DURON).

Volatilidade:
- Voláteis.

Tecnologia:
- Memórias Estáticas (SRAM).

Temporariedade:
- Varia, pois ao mesmo tempo em que pode estar servindo apenas de ponte para a MP (grande temporariedade), pode estar armazenando os endereços mais utilizados (pequena temporariedade).

Custo:
- As CACHE L1 são mais caras que os registradores.

Observação importante: O conceito de CACHE é utilizado sempre que existe necessidade de aperfeiçoar a troca de dados entre dispositivos com velocidades diferentes. No estudo de Bancos de Dados é aplicado entre o DBMS e o disco físico (HD). No acesso à Internet é utilizado entre a conexão real (WEB) e a cópia temporária local (HD). Os próprios discos magnéticos (HD) já estão vindo com uma CACHE própria (FLASH EPROM) para esta otimização.

6 – MEMÓRIA PRINCIPAL (MP)

6.1 – Memória RAM

A arquitetura definida por Von Neumann se baseava em dois princípios básicos:
- Base binária (0 e 1);
- Armazenamento dos programas, assim como dos dados, na memória do computador.

Para tal é importante uma memória de bom tamanho para executarmos os passos dos programas assim como armazenarmos os dados necessários. É importante diferenciarmos a Memória Principal (MP), que é elétrica, da Memória Secundária, composta por discos, fitas, CD-ROM e outros dispositivos de armazenamentos de massa.

A memória principal dos computadores modernos é fabricada com tecnologia de semicondutores, o que lhes permite elevada velocidade de acesso e transferência de bits, já que são circuitos apenas elétricos em funcionamento (não há partes mecânicas ou magnéticas). A velocidade de propagação de um sinal elétrico é cerca de 300.000 km/s (velocidade de luz).

A memória principal (RAM - Random Access Memory) é a memória de trabalho da UCP, seu grande "bloco de rascunho", onde os programas e os dados e sucedem em execução, uns após os outros. A memória RAM permite o acesso a qualquer uma das células de memória a qualquer tempo, se diferenciando das memórias chamadas seqüenciais (como as fitas magnéticas), nas quais é necessário acesso a todos os registros até a identificação da célula desejada.

Uma vez que esta memória é volátil há necessidade deste programa e seus dados estarem armazenados em alguma forma de memória secundária (HD, CD-ROM) antes de serem chamados para a memória principal (RAM).

Na realidade, o programa não precisa mais estar inteiro na memória RAM, bastando que o mesmo seja dividido em "pedaços" chamados de páginas, executadas em seqüência.

Assim sendo, a MP é composta por "locais" onde podem ser armazenados dados na forma de "bits". Estes "locais" podem ser acessados pelo seu "endereço".

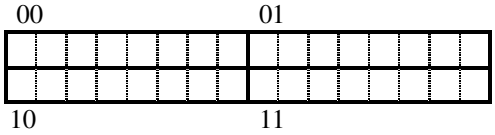
O microprocessador gasta normalmente dois pulsos de clock ("externo" – placa-mãe) para acessar a memória RAM. A memória RAM tem que possuir um Tempo de Acesso (TA) inferior a este valor:

Freqüência de Operação Máxima (MHz) = $\frac{2}{\text{TA (ns)}}$

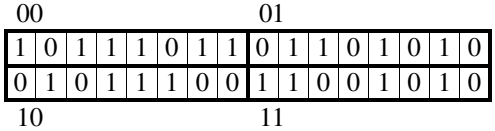
CLOCK EXTERNO	TEMPO DE UM CLOCK	TEMPO DE ACESSO MÍNIMO
66 MHz	15 ns	30 ns
100 MHz	10 ns	20 ns
133 MHz	7.5 ns	15 ns
200 MHz	5 ns	10 ns
400 MHz	2.5 ns	5 ns

Uma memória de 60 ns num microcomputador de 66 MHz tem que esperar dois "Wait States" de 15 ns, que somados ao T.A. mínimo de 30 ns vão permitir a comunicação.

Imaginemos uma memória elementar composta por apenas 4 "locais". Cada "local" deve possuir seu próprio endereço, portanto serão necessários 4 endereços (binários, naturalmente):



Por outro lado, cada "local" deve ser capaz de armazenar um grupo de bits. Cada bit deve possuir um caminho (trilha) desde o microprocessador até a memória (na verdade, entre quaisquer dispositivos que devam transmitir ou receber dados). Se a capacidade de armazenamento do "local" for 8 bits, seriam necessárias 8 trilhas para receber ou transmitir estes dados.



Para endereçarmos 4 "células" de memória utilizamos 2 bits, porque $2^2 = 4$. Desta forma, para endereçarmos 32 "células" de memória utilizaremos 5 bits ($2^5 = 32$). Para endereçarmos 256M endereços, utilizaremos 28 bits ($2^8 * 2^{20} = 2^{28}$). Assim, se **E** for o número de bits de um endereço e **N** for o número de endereços, temos:

$$\mathbf{N = 2^E}$$

Esta é a equação que será utilizada na resolução dos exercícios. Na realidade o endereçamento da memória utiliza o sistema "Linha/Coluna", o qual permite localizar um endereço a partir das coordenadas "Linha/Coluna" de uma matriz. Assim se os 256M (2^{28}) endereços de uma memória estiverem distribuídos em uma matriz 2^{14} X 2^{14}, precisaremos de apenas 14 bits (a metade dos 28 vistos anteriormente) para o endereçamento.

Nesta sistemática o endereço é divido em dois: **RAS (Row Address Strobe)** e **CAS (Column Address Strobe)**, responsáveis pelos endereços "linha" e "coluna", respectivamente.

A memória é organizada como um conjunto de N células com M bits cada. Se possui N células, necessita de N endereços. Por outro lado se a célula possui M bits, podemos armazenar nela um valor entre 0 até 2^{N-1}. O valor N representa a "capacidade da memória", ou seja, a quantidade de células ou de endereços, enquanto M indica a quantidade de bits que podem ser armazenados.

O total de bits de uma memória é portanto T = M X N

A MP é o "depósito" de trabalho da UCP, isto é, a UCP e a MP trabalham íntima e diretamente na execução de um programa. As instruções e os dados do programa ficam armazenados na MP e vai "buscando-os" um a um à medida que a execução vai se desenrolando.

Os programas são organizados de modo que os comandos são descritos seqüencialmente e o armazenamento das instruções se faz da mesma maneira, fisicamente seqüencial (células contíguas de memória).

A palavra é a unidade de informação que o sistema UCP/MP processa em um ciclo de clock, devendo ser do tamanho do Registrador de Instruções (RI). Os programas são armazenados em células contíguas de memória, logicamente a MP e os registradores da UCP deveriam ser constituídos de células com tamanho compatível com a palavra, porém os fabricantes seguem idéias próprias quanto a isso, não tendo sido adotado um padrão.

O microprocessador PENTIUM (INTEL) possui barramento de 64 bits (trilhas) mas utiliza palavra de 32 bits. Os processadores INTEL 80486, MOTOROLA 68000, IBM 4381, VAX 11 possuem palavras definidas como de 32 bits, mas utilizam MP organizada em grupos de 8 bits (1 byte).

Em nossa disciplina idealizaremos o computador de uma forma primária, ou seja, se o dado (palavra) precisa transitar por um barramento, este barramento necessita ter a mesma quantidade de trilhas que ele. Por outro lado, a célula de memória também deve ser do mesmo tamanho para permitir o seu armazenamento.

Um microcomputador deve possuir três barramentos:

1. **Barramento de dados:** Interliga a RDM à MP. É bidirecional, ou seja, os dados trafegam da UCP para a MP ou da MP para a UCP.
2. **Barramento de endereços:** Interliga a REM à MP. É unidirecional, uma vez que a UCP sempre informa à MP qual o endereço pretende ler ou escrever.
3. **Barramento de controle:** Interliga a UCP à MP, permitindo a passagem de sinais de controle durante uma operação de leitura ou escrita. É bidirecional, pois a UCP pode enviar sinais de controle para a MP (comando de leitura ou escrita), e a MP pode enviar sinais do tipo WAIT (ordena que a UCP fique em estado de espera – WAIT STATE).

6.2 – Considerações Sobre Memória RAM

6.2.1 Tipos de memória RAM

A memória RAM passou por evoluções tecnológicas buscando, dentre outras características, o menor Tempo de Acesso, recebendo ao longo do tempo várias denominações específicas. Alguns exemplos são:

- **DRAM original - (Dynamic Random Access Memory)** – Foi o primeiro tipo de memória usado em micros PC. Neste tipo antigo de memória, o acesso é feito enviando primeiro o endereço RAS e em seguida o endereço CAS, da forma mais simples possível. Este tipo de memória foi fabricado com o tempo de acesso a partir de 150 ns, mais do que suficientes para suportar o clock de 4,77 MHz do PC original. Foram desenvolvidas posteriormente versões de 80, 100 e 120 ns para serem utilizadas em micros 286 e 386.

- **FPM DRAM - (Fast Page Mode)** – Ao ler um arquivo qualquer gravado na memória, os dados estão na maioria das vezes gravados seqüencialmente. Não é preciso enviar o endereço RAS e CAS para cada memória a ser lida, mas simplesmente enviar o endereço RAS (linha) uma vez e em seguida enviar vários endereços CAS (coluna) **(bursting)**. Devido ao novo método de acesso, as memórias FPM conseguem ser cerca de 30% mais rápidas que as DRAM originais e foram utilizadas em micros 386, 486 e nos primeiros micros Pentium, com tempos de acesso de 60, 70 e 80 ns.

- **EDO DRAM - (Extended Data Output) –** Além de ser mantido o modo de acesso rápido das memórias FPM, foram feitas algumas modificações para permitir que um acesso a dados possa ser iniciado antes que o anterior termine, permitindo aumentar perceptivelmente a velocidade dos acessos. Este tipo de memória foi fabricado em velocidades de 50, 60 e 70 ns, conseguindo ser 25% mais rápidas que as FPM.

- **BEDO DRAM (Burst Extended Data Output) –** As memórias BEDO utilizam uma espécie de pipeline para permitir acessos mais rápidos. Eram quase 30% mais rápido que as memórias EDO e custavam o mesmo preço. Não eram, entretanto, compatíveis com os chipsets Intel e por isso foram pouco utilizadas.

- **SDRAM - (Synchronous DRAM)** – As memórias FPM, EDO e BEDO são assíncronas, isto significa que elas trabalham em seu próprio ritmo, independentemente dos ciclos da placa mãe. Isso explica por que memórias FPM que foram projetadas para funcionar em placas para processadores 386 ou 486 (clocks de 25 ou 33 MHz), funcionam sem problemas em placas para processadores Pentium, que funcionam a 66 MHz. Na verdade, as memórias continuam funcionando na mesma velocidade, o que muda são os tempos de espera que passam a ser mais altos. Assim, ao invés de responder a cada 2 ciclos da placa mãe, por exemplo, elas podem passar a responder a cada 4 ciclos. As memórias SDRAM por sua vez, são capazes de trabalhar sincronizadas com os ciclos da placa mãe, sem tempos de espera. Isto significa que a temporização de uma memória SDRAM é sempre de uma leitura por ciclo, independentemente da velocidade de barramento utilizada.

- **DDR-SDRAM (Double Data Rate SDRAM) –** Memórias SDRAM que transferem dados tanto na subida quanto na descida do pulso de clock.

- **DR-SDRAM – (Direct Rambus SDRAM)** – As requisições de leitura e gravação da memória podem ser transmitidas a qualquer momento por meio do sinal de sincronismo, sem a necessidade de esperar que pedidos anteriores sejam atendidos. Internamente é uma memória que transmite os dados de forma serial. Externamente, o caminho do barramento de dados é formado por quatro vias com quatro bytes de largura, permitindo que até quatro requisições independentes possam ser solicitadas ao mesmo tempo. Tecnologia proprietária da Rambus Inc.

6.2.2 Taxa de transferência

O Front Side Bus (FSB) ou Barramento Externo é o caminho de comunicação do processador com o chipset da placa-mãe, e, portanto com a memória RAM (Circuito Ponte Norte ou North Bridge). O termo "Clock Externo" é utilizado para representar a velocidade de comunicação nesse barramento. Um "FSB de 100 MHz" significa "clock externo de 100 MHz".

Todos os processadores a partir do 486DX2 passaram a utilizar a "Multiplicação de Clock", onde o clock interno do processador é maior do que o seu clock externo (FSB). Um Pentium IV de 3,2 GHz trabalha internamente a 3,2 GHz, porém externamente ele opera a 200 MHz (ou seja, seu FSB é de 200 MHz).

Os processadores de 7ª geração da Intel (Pentium IV e Celeron soquete 478) trabalham transferindo quatro dados por pulso de clock (QDR ou Quad Data Rate - Taxa de Transferência Quadruplicada). Com isto, muitas vezes é dito que o barramento externo (FSB) do Pentium IV é de 400 MHz, 533 MHz ou 800 MHz, enquanto na realidade este é de 100 MHz, 133 MHz ou 200 MHz, respectivamente.

As memórias DDR-SDRAM (Double Data Rate Synchronous Dynamic Random Access Memory) são memórias SDRAM que em vez de transferirem um dado por pulso de clock, como é o usual, transferem dois dados por pulso de clock. Com isso, elas conseguem obter um desempenho que é o dobro do desempenho das memórias SDRAM tradicionais operando a um mesmo clock.

No caso das memórias DDR os fabricantes informam valores que não correspondem ao clock verdadeiro, mas sim ao seu "desempenho" duplicado. Por exemplo, memórias DDR-SDRAM de "400 MHz" na verdade trabalham a 200 MHz. O barramento externo do processador Athlon XP é dito como sendo de 266 MHz, 333 MHz e 400 MHz, enquanto na verdade são de 133 MHz, 166 MHz e 200 MHz, respectivamente.

Para evitar comparação utilizando o simples clock, o qual pode conduzir a conclusões erradas, devemos comparar o desempenho de um barramento através da sua taxa de transferência máxima dada em MB/s (megabyte por segundo).

A fórmula para calcular a "taxa de transferência" é a seguinte:

CLOCK x DADOS x N / 8

- **CLOCK:** Clock real da placa mãe;
- **DADOS:** Corresponde à quantidade de dados transferidos por pulso de clock. Depende do sistema utilizado, sendo:
  - 1 para as memórias SDRAM;
  - 2 para as DDR-SDRAM;
  - 4 quando utilizamos processadores Intel configurados com QUAD;
- **N:** É o número de bits que o processador utiliza para se comunicar com a memória RAM.
  - Processadores comuns (32 bits): N = 64;
  - Processadores 64 bits: N = 128.
- A divisão por 8 é para obter o resultado em bytes por segundo (B/s)

A tabela a seguir lista os tipos de memória e suas Taxas de Transferência máxima em função do clock real (externo) e do anunciado:

MEMÓRIA	TECNOLOGIA	CLOCK REAL	CLOCK ANUNCIADO	TAXA DE TRANSFERÊNCIA MÁXIMA
PC-66	SDRAM	66 MHz	66 MHz	533 MB/s
PC-100	SDRAM	100 MHz	100 MHz	800 MB/s
PC-133	SDRAM	133 MHz	133 MHz	1.066 MB/s
PC-1700 ou DDR200	DDR-SDRAM	100 MHz	200 MHz	1.600 MB/s
PC-2100 ou DDR266	DDR-SDRAM	133 MHz	266 MHz	2.100 MB/s
PC-2700 ou DDR333	DDR-SDRAM	166 MHz	333 MHz	2.700 MB/s
PC-3200 ou DDR400	DDR-SDRAM	200 MHz	400 MHz	3.200 MB/s
PC2-3200 DDR2-400	DDR2-SDRAM	200 MHz	400 MHz	3.200 MB/s
PC2-4300 DDR2-533	DDR2-SDRAM	266 MHz	533 MHz	4.264 MB/s
PC2-5300 DDR2-667	DDR2-SDRAM	333 MHz	667 MHz	5.336 MB/s
PC2-6400 DDR2-800	DDR2-SDRAM	400 MHz	800 MHz	6.400 MB/s

A situação ideal para obter a máxima performance de um microcomputador é quando a Taxa de Transferência do barramento externo do processador é igual à da memória. Não existe problema quando a memória é mais rápida do que o barramento externo do processador.

Exemplo: Um Pentium IV trabalha externamente a 100 MHz transferindo quatro dados por pulso de clock (QUAD). Isso faz com que ele trabalhe "como se" estivesse a 400 MHz, embora fisicamente falando isso não ocorra. A taxa de transferência pode atingir, portanto, picos de 3.200 MB/s (4 x 800 MB/s). Observando a tabela acima percebemos que devemos optar pela memória DDR400 ou DDR2-400, pois são capazes de "conversar" com o Pentium IV à mesma velocidade.

No caso da memória Direct Rambus (DR-SDRAM), ela utiliza dois canais Rambus de 1.600 MB/s cada. Pela maneira que a tecnologia Rambus funciona, esse desempenho é multiplicado pelo número de canais usados. Assim, a taxa de transferência máxima dessa plataforma é de 3.200 MB/s, conseguindo fazer com que o processador comunique-se com a memória na taxa máxima possível.

Não é uma boa escolha porque a tecnologia é proprietária da empresa Rambus e todos têm de pagar direitos autorais à esta empresa: o fabricante da memória, o fabricante dos chips para placas-mãe, etc. Isso faz com que os preços da memória Rambus e das placas-mãe que aceitam esse tipo de memória fiquem bem mais caros. Um módulo de memória Rambus de 256 MB custa atualmente exatamente o dobro de um módulo de memória DDR de 256 MB. A melhor relação custo/benefício para micros é realmente a memória DDR-SDRAM.

No exemplo acima compatibilizamos o processador com a memória. No caso de processadores INTEL de 533 MHz e 800 MHz de FSB (133 e 200 MHz reais, respectivamente), teríamos que ter memórias com 4.256 MB/s e 6.400 MB/s. Neste caso as DDR400 não conseguem fornecer os dados com a velocidade planejada.

Uma opção interessante seria utilizar a configuração DDR Dual Channel, que dobra a taxa de transferência do chipset com a memória RAM, por acessar dois módulos de memória ao mesmo tempo. Usando dois módulos DDR400 simultaneamente a taxa de transferência máxima seria do dobro de 3.200 MB/s, ou seja, para 6.400 MB/s.

A placa-mãe deve possibilitar essa configuração, sendo necessária a instalação de pelo menos um par de módulos com as seguintes características idênticas:

- **Capacidade:** 128MB, 256 MB, 512MB, por exemplo;
- **Taxa de transferência:** Mesmo tempo de acesso;
- **Número de linhas**: Normalmente face simples ou face dupla

No Dual Channel, além da controladora de memória normal (integrada à Ponte Norte) de 64 bits, existe uma segunda também de 64 bits independente, totalizando 128 bits.

Os processadores AMD de 32 bits tem FSB é baixo (entre 266 MHz e 400 MHz), um simples módulo de memória em uma controladora dão conta do recado. A vantagem do Dual Channel nesse caso está no custo, pois é possível a instalação de dois módulos de 200 MHz DDR ao invés de um módulo de 400 MHz, mais caro. Além disso, o Dual Channel não só duplica a largura de banda para a CPU, mas pode diminuir a latência do sistema, por deixar uma controladora de memória atendendo ao HD, AGP e PCI enquanto a outra controladora atende somente a CPU.

O processador Athlon 64 bits da AMD possui o controlador de memória embutido no próprio processador, e não no chipset. Nesse caso, a configuração da técnica DDR Dual Channel depende do processador, e não da placa-mãe.

Note que o fato de a memória DDR-SDRAM possuir o dobro do desempenho da memória SDRAM não significa que um micro equipado com memória DDR-SDRAM terá o dobro do desempenho de um micro com mesma configuração usando memórias SDRAM tradicionais. A tecnologia de memória RAM utilizada é somente um dos fatores que influenciam no desempenho geral da máquina.

Existem ainda as novas memórias DDR2 e DDR3 que são simplesmente a evolução normal da tecnologia DDR. Não significa que o seu desempenho seja duas e três vezes maiores que as DDR.

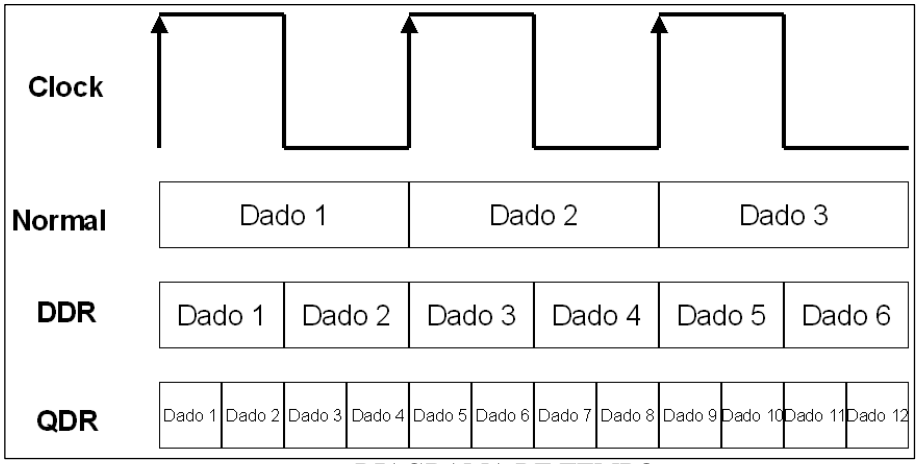


DIAGRAMA DE TEMPO

6.2.3 Interleaving

O termo "Interleaving" (ou intercalamento) se refere ao processo em que o processador (CPU) se comunica alternadamente com as células presentes em duas ou mais tabelas (bancos) de memória. Esta técnica é normalmente usada em servidores ou Workstations e funciona da seguinte forma: toda vez que a CPU seleciona uma tabela (banco) de memória, a tabela precisa de um ciclo de clock para ser inicializada.

A CPU pode ganhar um tempo precioso se selecionar a próxima (tabela) de memória enquanto a primeira tabela está sendo inicializada. O uso da técnica de Interleaving pode melhorar o desempenho da memória, pois duas ou mais tabelas independentes podem ser selecionadas produzindo um fluxo de dados mais rápido.

6.2.4 Encapsulamento

As pastilhas de memória também podem ser classificadas pelo encapsulamento (forma física dos pentes de memória), que define onde elas podem ser instaladas na placa-mãe:

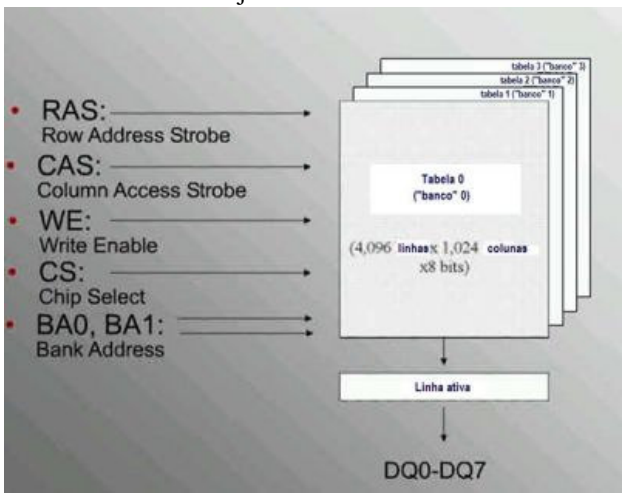
- **SIMM (Single In Line Memory Module):** Pode ser SIMM-30 (trinta terminais para endereços, dados e controle), permitindo a passagem de 8 bits em cada ciclo, ou SIMM-72 (72 terminais), que permite a passagem de 32 bits em cada ciclo.
- **DIMM (Double In Line Memory Module):** Possui 168 terminais, possibilitando a transferência de 64 bits em cada ciclo, sendo utilizado nas memórias SDRAM.
- **SO DIMM (Small Outline DIMM):** São pentes DIMM, de tamanho reduzido, com 72 ou 144 pinos, geralmente utilizados em notebooks. Os SO DIMM de 72 pinos trabalham com 32 bits e os SO DIMM de 144 pinos possuem 64 bits.
- **DDR-DIMM (Double Data Rate Double In Line Memory Module):** Possui 184 terminais, possibilitando a transferência de 2 grupos de 64 bits em cada ciclo.
- **DDR2-DIMM (Double Data Rate 2 Double In Line Memory Module):** Possui 240 terminais, possibilitando também a transferência de 2 grupos de 64 bits em cada ciclo. A sua tecnologia de construção é mais avançada, diminuindo o consumo e o aquecimento e permitindo velocidades mais elevadas.
- **RIMM (Rambus In Line Memory Module):** Possui 184 terminais, possibilitando a transferência de apenas 16 bits, porém em velocidades mais elevadas.

Apesar do tamanho físico dos módulos DIMM (usado por memórias SDRAM), DDR-DIMM e DDR2-DIMM ser o mesmo, o encaixe desses dois módulos é diferente.

6.2.5 Sinais de Controle Usados nos Chips de Memória

Os chips de memória RAM podem ser considerados um conjunto de tabelas de linhas e colunas. Estes chips recebem sinais de endereçamento e controle e fornecem ou armazenam um dado. Como uma tabela, um chip de RAM pode ser organizado em linhas e colunas.

Na figura ao lado temos um chip de memória com 4 tabelas de linhas e colunas, cada uma contendo 4096 linhas e 1024 colunas. Cada célula de memória possui 8 bits de dados. Este exemplo em particular representa um chip de 128 Mbit (4 x 4096 x 1024 x 8 bits).



Existem vários sinais usados para controlar a RAM, são eles:

- **RAS (Row Address Strobe):** Ativa o endereço da linha selecionada;
- **CAS (Column Address Strobe):** Ativa o endereço da coluna selecionada;
- **WE (Write Enable):** Sinal que controla a escrita de dados na memória;
- **CS (Chips Select):** Ativa a memória para leitura ou escrita;
- **BA0, BA1, …, BAn:** Determina que tabela (banco) de memória será acessado;
- **DQ0, …, DQ7:** Dados que saem/entram no chip de RAM

6.2.6 Temporização e Latência das Memórias

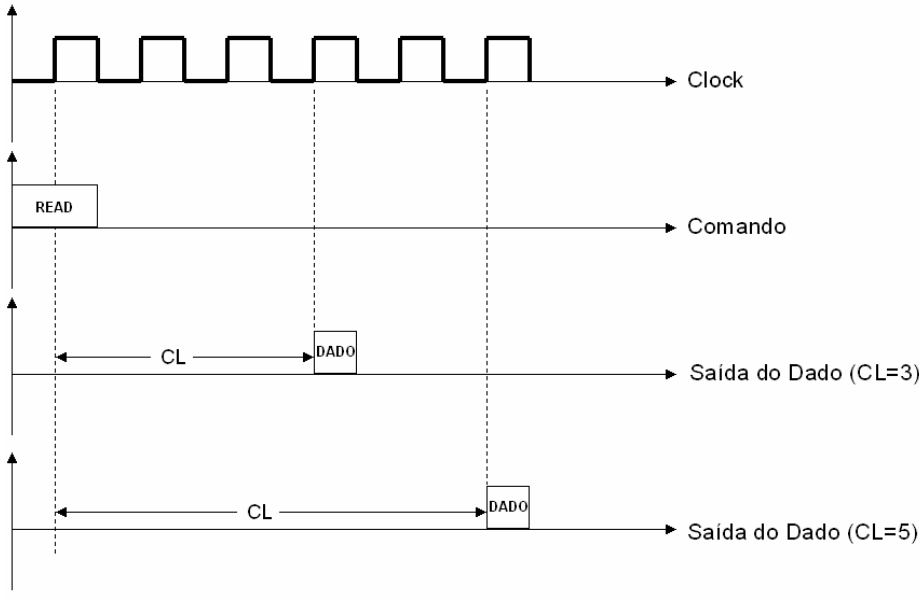
Por causa das temporizações dois módulos de memória com mesma taxa de transferência máxima teórica podem apresentar desempenhos diferentes. As temporizações medem o tempo em que o chip de memória demora para fazer algo internamente.

A temporização da memória é dada através de uma série de números, como, por exemplo 2-3-2-6-T1, 3-4-4-8 ou 2-2-2-5. Estes números indicam a quantidade de pulsos de clock que a memória demora para fazer uma determinada operação. Quanto menor o número, mais rápida é a memória. Na figura ao lado a temporização é 5-5-5-15.

Normalmente as temporizações são padronizadas ("Auto" no SETUP), porém é possível a configuração manual de temporizações menores para aumentar o desempenho do seu micro (OVERCLOCK). Algumas placas-mãe podem não funcionar com temporizações muito baixas.

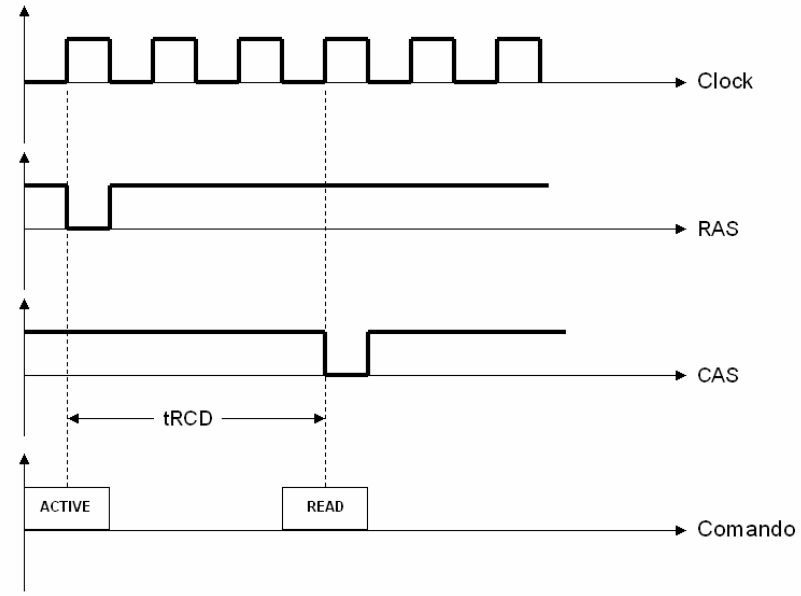
As temporizações que estes números indicam são as seguintes: **CL-tRCD-tRP-tRAS-CMD**

- **CL (CAS Latency ou Latência do CAS):** Indica a quantidade de pulsos de clock que a memória leva para retornar um dado solicitado. Uma memória com CL=3 demora três pulsos de clock para entregar um dado, enquanto que uma memória com CL=5 demora cinco pulsos de clock para realizar a mesma operação. Dessa maneira dois módulos trabalhando com o mesmo clock o que tiver a menor latência do CAS será o mais rápido. Na figura temos a comparação entre um módulo com CL = 3 e outro com CL = 5. O clock considerado aqui é o clock real, ou seja, metade do clock rotulado.

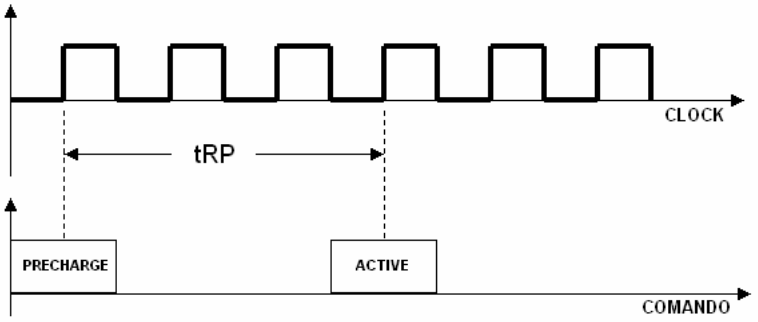


Uma memória DDR2-533 recebe pulsos de clock a cada 3.75 ns (1/(266 x 10^6)) , se CL = 3, a demora seria de 11.28 ns. As memórias SDRAM, DDR e DDR2 implementam o modo burst (rajada), onde um dado após o primeiro solicitado demora apenas um pulso de clock para ser entregue pela memória, desde que este dado esteja localizado em um endereço logo após o endereço do dado atual. Com isso, enquanto o primeiro dado demoraria a quantidade de pulsos de clock da latência do CAS para ser entregue, o próximo dado seria entregue logo após o dado que acabou de sair da memória, não tendo de esperar um outro ciclo de latência do CAS.

- **tRCD (RAS to CAS Delay):** Cada chip de memória é organizado internamente como uma matriz. Internamente o processo de acessar um dado é feito por dois sinais de controle chamados RAS (Row Address Strobe) e CAS (Column Address Strobe). Quanto menor for o tempo entre esses dois sinais, melhor, já que o dado será lido mais rapidamente. O parâmetro "RAS to CAS Delay" ou (tRCD) mede este tempo. Na figura ao lado podemos ver uma memória com tRCD=3. O parâmetro "RAS to CAS Delay" é também o número de pulsos de clock entre o comando "Active" ("Ativar" ou CS) e um comando "read" ("leitura") ou "write" ("escrita").

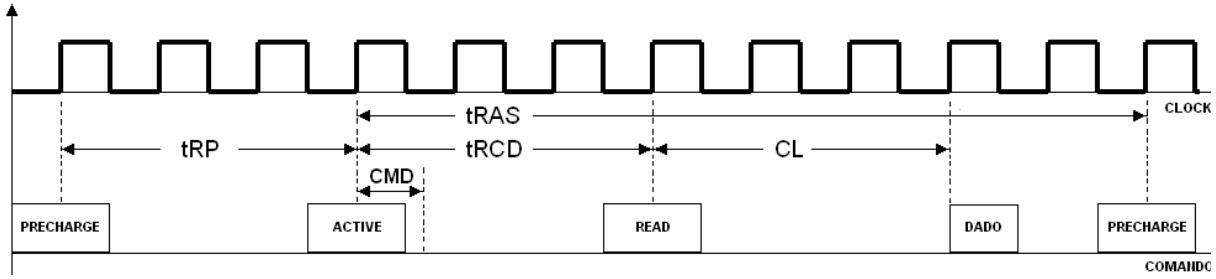


- **tRP (RAS Precharge):** Tempo demorado entre desativar o acesso a uma linha de dados e iniciar o acesso a outra linha de dados. Após o dado ter sido entregue pela memória, um comando chamado "Precharge" precisa ser executado para desativar a linha da memória que estava sendo usada e para permitir que uma nova linha seja ativada. O tempo "RAS Precharge" (tRP) é o tempo entre o comando "Precharge" e o próximo comando "Active" ("Ativar" ou CS). Na figura acima temos um exemplo de uma memória com tRP=3.



- **tRAS (Active to Precharge Delay):** Após um comando "Active" (CS) ter sido enviado, um outro comando "Precharge" não pode ser iniciado até que o tempo tRAS tenha decorrido. Em outras palavras, este parâmetro limita quando a memória pode iniciar a leitura (ou escrita) em uma linha diferente.

- **CMD (Command Rate):** Tempo demorado entre o chip de memória ter sido ativado e o primeiro comando poder ser enviado para a memória. Algumas vezes este valor não é informado. Normalmente possui o valor T1 (1 clock) ou T2 (2 clocks).



6.2.7 Identificação de Módulos de Memória

A única forma segura de identificar uma memória é através da etiqueta do fabricante (código numérico existente no chip de memória.)

GM **manufacturer code**	GM - Lucky Goldstar LGS	
72 **product family**	71 - DRAM EDO / FastPage 72 - SDRAM	
V **voltage**	C - 5.0V V - 3.3V VL - 2.35V	
Refresh	[blank] - standard S - self refresh	
16 **Density & refresh cycle**	DRAM 16 - 16M, 4k refresh 17 - 16M, 2k refresh 18 - 16M. 1k refresh 64 - 64M, 8k refresh 65 - 64M, 4k refresh	SDRAM
82 **organization**	DRAM 10 - x1 16 - x16 (2CAS) 17 - x16 (2WE) 32 - x32 (2CAS) 33 - x32 (2WE) 34 - x32 (4CAS 40 - x4 41 - x4 (4CAS) 80 - x8 82 - x8 (4CAS) 16 - x16	SDRAM 6162 = 1Meg x 16 (16MBit) 642 = 4Meg x 4 (16MBit) 682 = 2Meg x 2 (16MBit) 8164 = 8Meg x 16 (128MBit) 844 = 32Meg x 4 (128MBit) 884 = 16Meg x 8 (128MBit) 6164 = 16Meg x 16(256MBit) 644 = 64Meg x 4 (256MBit) 684 = 32Meg x 8 (256MBit) 6164 = 4Meg x 16 (64MBit) 644 = 16Meg x 4 (64MBit) 684 = 8Meg x 8 (64MBit)
1 **access mode**	0 - FastPage 1 - SDRAM 3 - EDO 5 - EDO	
C **revision**	[blank] / A / B / C / D ...	
T **package**	[blank] - plastic DIP J - SOJ T - TSOP R - TSOP reverse	
10K cycle time	5 - 50ns 6 - 60ns 7 - 70ns 8 - 80ns 10 - 100ns 10K - 66MHz (15ns/9ns/2-2-2) 7K - 100MHz (10ns/6ns/2-2-2) 7J - 100MHz (10ns/6ns/3-2-2) 8 - 125MHz (8ns/6ns/3-3-3) 75 - 133MHz (7.5ns/5.4ns/3-3-3) 7 - 143MHz (7ns/5.4ns/3-3-3)	

Por exemplo, **GM72V16821CT10K**

- **GM:** Fabricante LG Semiconductors
- **72:** Tipo SDRAM
- **V:** Voltagem de 5 volts
- **[vazio]:** Refresh de memória "standard" (padrão)
- **16:** Densidade da memória de 16 Mb com refresh de 4k
- **82:** x8 (16 Mb x 8 = 16 MB)
- **1:** Memória do tipo SDRAM
- **C:** A versão do chip é a C
- **T:** Encapsulamento TSOP
- **10K:** 66 MHz (PC66)

6.2.8 SPD (Serial Presence Detect)

O SPD é uma pequena memória EEPROM existente no módulo e que contém as informações de configuração do módulo escritas pelo fabricante do mesmo.

O JEDEC (Joint Electronic Device Engineering Council) é um organismo que define os valores padrões que devem ser colocados no SPD.

As principais informações armazenadas no SPD são:

- Número de bancos de memória
- Tensão (Voltagem)
- RAS Precharge (tRP)
- RAS to CAS Delay
- Código JEDEC do fabricante
- Tipo de memória (SDRAM, DDR, etc.)
- Número de bits (normalmente 64, mas memórias com paridade ou ECC trabalham com 72 bits)
- ECC/não ECC (Correção de Erros)
- Tempo do ciclo de clock da memória
- Latências CAS suportadas
- Fabricante do módulo

6.3 – Memória ROM

Uma vez que o processador nada realiza sem as instruções, é óbvio que ele deve ter acesso a uma certa quantidade de memória não-volátil, isto é, a um local onde estejam permanentemente armazenadas instruções que automaticamente iniciam a operação e a inicialização do sistema, tão logo a alimentação elétrica seja ligada.

Além de não perder seus dados no caso de falta de energia, esta memória não deve sofrer alteração por parte dos usuários ou dos próprios programas sendo então uma Memória de Apenas Leitura (Read Only memory – ROM).

Na verdade a memória ROM de um computador pode ser composta por mais de um chip, dividindo três tarefas básicas: BIOS, POST e SETUP.

A BIOS (Basic Input Output System) carrega um programa inicial no processador que é chamado de "bootstrap", "boot" ou IPL (Initial Program Load – Carregamento do Programa Inicial). Não deve ser confundido com o "boot" do sistema operacional ("Iniciando o Windows-98"). Cada modelo de placa-mãe tem um IPL próprio e exclusivo, em função do CHIPSET, ou seja, do conjunto de chips que formam a placa-mãe. Após o IPL ter sido carregado, inicia-se o POST (Power-On Self Test), que testa a memória, teclado, vídeo e os dispositivos interligados que compõem o microcomputador.

O SETUP simplesmente registra a configuração atual do hardware, podendo ser modificado em função das alterações efetuadas (troca de HD, aumento de memória, etc). Existe um dispositivo conectado diretamente ao SETUP chamado de REAL-TIME, cuja função é manter a hora e data atualizada, mesmo sem energia. Diferentemente do IPL, estas configurações (SETUP e REAL-TIME) podem ser alteradas, pois estão gravadas em uma memória EEPROM (C-MOS), que é uma ROM eletricamente regravável.

Importante ressaltar que a ROM é uma RAM (Random Access memory) que não aceita modificações em seus dados armazenados!!! As memórias ROM são mais lentas do que as RAM, razão porque muitos sistemas transferem seu conteúdo da ROM BIOS para a RAM (MP), melhorando o desempenho (SHADOW ROM BIOS).

A memória ROM vem gravada de fábrica ("Mask Programmed" – programada pela máscara) e não pode ser alterada, pois já é fabricada com o programa. Nessa ROM pura, o conjunto de bits é inserido no interior dos elementos da pastilha durante o processo de fabricação. Em inglês chama-se **hardwired**, pois cada bit (seja 0 ou 1) é criado já na célula apropriada.

Após o término da fabricação, a pastilha ROM está completa, com o programa armazenado, e nada poderá alterar o valor de qualquer de seus bits.

O custo de fabricação da "máscara" (SiO_2) é elevado, sendo que a memória ROM pura só é economicamente viável para grandes remessas. Para atenuar o custo da máscara, desenvolveu-se uma variação da memória ROM chamada PROM.

Uma memória PROM (Programmable Read Only Memory) pode ser gravada **apenas uma vez** após a fabricação da pastilha (chip), utilizando equipamento especial (gravador). Após a transferência dos dados para a PROM, um "pulso" elétrico os fixa ("queima") definitivamente.

A vantagem da PROM sobre a ROM é que, apesar de mais cara, seu preço não depende da quantidade fabricada. Foram desenvolvidas ainda:

- **EPROM (Erasable PROM):** PROM apagável (regravável), utilizando luz ultravioleta;
- **EEPROM, EAROM, E2PROM (Electrically EPROM):** EPROM eletricamente regravável;
- **FLASH EEPROM:** EEPROM de alta velocidade, eletricamente regravável.

6.4 – Características da MP (RAM)

Tempo de Acesso:
- 10 a 20 ns.

Capacidade:
- O valor típico está entre 256 MB e 1 GB.

Volatilidade:
- As memórias RAM são voláteis, porém todo computador precisa de uma memória não volátil (ROM, EPROM, EEPROM) para armazenar os programas, dados e configurações básicas do computador, permitindo que ele se inicialize ao ser ligado.

Tecnologia:
- Normalmente a memória de menor custo, maior capacidade de integração e menor aquecimento são as memórias dinâmicas: DRAM. Possui tempo de acesso maior que os registradores e CACHE.

Temporariedade:
- Varia, pois depende do tamanho do programa que está sendo executado ou da quantidade de dados que estão armazenados num determinado instante.

Custo:
- A mais barata das memórias elétricas.

7 – MEMÓRIA SECUNDÁRIA (MS)

A Memória Secundária é utilizada para armazenar toda a estrutura de dados e programas de uma forma mais permanente. Se dividem em dois tipos básicos de dispositivos:
- **Acesso Imediato:** O dado é acessado imediatamente;
- **Acesso quando desejado:** É necessário colocarmos a mídia no dispositivo.

7.1 - Características da Memória Secundária

Tempo de Acesso:
- 5 a 10 ms para os HD. Depende da controladora (IDE, SCSI, “Firewire”, etc.);
- 20 ms para pendrive;
- 120 a 300 ms para CD-ROM;

Capacidade:
- HD: 20 a 60 GB. CD-ROM: 700 MB. Fitas: De 2 a 200 GB.
- Pendrive de 512 a 2 GB;

Volatilidade:
- Não voláteis.

Tecnologia:
- Cada tipo de memória secundária tem sua própria tecnologia, normalmente baseada em campos magnéticos (HD, fita ou disquete), óticos (CD-ROM) e elétricos (pendrive).

Temporariedade:
- Pequena – quase definitiva.

Custo:
- Possuem o menor custo US$/MB.

LISTA DE EXERCÍCIOS

1. O que você entende por acesso á memória? Caracterize o tempo de acesso nos diversos tipos de memória.

2. Quais são as possíveis operações que podem ser realizadas em uma memória?

3. Quais os tipos de memória existentes em um computador. Porque são necessários tipos diferentes em um computador?

4. Faça o diagrama mostrando a relação Custo x Quantidade de Memória.

5. Qual a função dos registradores REM e RDM?

6. Qual a diferença conceitual entre uma memória SRAM e uma DRAM? Cite vantagens e desvantagens de cada uma.

7. Qual a função da memória CACHE? Como ela funciona?

8. A memória CACHE foi idealizada a parir de dois conceitos teóricos. Cite-os e explique-os.

9. Cite os algoritmos de substituição de dados na CACHE e como cada um deles funciona:

10. Uma memória RAM tem um espaço máximo de endereçamento de 2K. Cada célula pode armazenar 16 bits. Qual o valor total de bits que podem ser armazenados nesta memória? Qual o tamanho do RDM e REM? Qual o tamanho dos barramentos de endereço (BE) e de dados (BD)? Qual o maior endereço desta memória?
    Respostas:
    N = 2K endereços — E = REM = 11 bits BE = 11 trilhas
    M = 16 bits — RDM = 16 bits BD = 16 trilhas
    T = 32K bits — Maior endereço = 11111111111 = 7FF

11. Uma memória RAM é fabricada com a possibilidade de armazenar um máximo de 256K bits. Cada célula pode armazenar 8 bits. Qual o tamanho de cada endereço, da REM e do BE? Qual o tamanho do RDM e BD? Qual é o total de células que podem ser utilizadas na RAM? Qual o maior endereço desta memória?
    Respostas:
    T = 256K bits — E = REM = 15 bits BE = 15 trilhas
    M = 8 bits — RDM = 8 bits BD = 8 trilhas
    N = 32K endereços — Maior endereço = 111111111111111 = 7FFF

12. Um computador, cuja memória RAM tem uma capacidade máxima de 2K palavras de 16 bits cada, possui um REM e um RDM. Qual o tamanho destes registradores; qual o valor do maior endereço dessa MP e qual a quantidade total de bits que nela podem ser armazenados? Qual o tamanho do BE e do BD?
    Respostas:
    N = 2K endereços — T = 32K bits
    M = 16 bits — RDM = 16 bits BD = 16 trilhas
    E = REM = BE = 11 bits — Maior endereço = 11111111111 = 7FF

13. Um computador possui uma memória principal com capacidade para armazenar palavras de 16 bits em cada uma de suas N células. O barramento de endereços tem 12 trilhas de tamanho. Quantos bits e bytes poderão ser armazenados nessa memória? Qual o tamanho do BD? Qual o tamanho do RDM e do REM? Qual o maior endereço desta memória?
    Respostas:
    M = 16 bits — T = 64K bits
    E = REM = 12 bits BE = 12 trilhas — RDM =16 bits BD = 16 trilhas
    N = 4K endereços — Maior endereço = 111111111111 = FFF

14. Um computador possui um RDM com 16 bits de tamanho e um REM com capacidade de armazenar números com 20 bits. Sabe-se que ele possui uma quantidade N de células, igual à sua capacidade máxima de armazenamento. Qual o tamanho do BE e do BD? Quantos endereços esta memória possui? Qual o total de bits que ela é capaz de armazenar? Qual o maior endereço desta memória?
    Respostas:
    E = REM = 20 bits — N = 1M endereços
    BE = 20 trilhas — T = 16M bits
    M = 16 bits — Maior endereço = 11111111111111111111 = FFFFF
    RDM = 16 bits
    BD = 16 trilhas

15. Um microcomputador possui uma memória principal com 32K células, cada uma com 8 bits. Quantos bits poderão ser armazenados nessa memória? Qual o tamanho do BD e do BE? Qual o tamanho do RDM e do REM? Qual o maior endereço desta memória?
    Respostas:
    N = 32K endereços — E = REM = 15 bits BE = 15 trilhas
    M = 8 bits — RDM = 8 bits BD = 8 trilhas
    T = 256K bits — Maior endereço = 111111111111111 = 7FFF

16. Considere uma célula de uma MP cujo endereço é 2C81 e que tem armazenado em seu conteúdo uma valor igual a F5A. Qual deve o tamanho mínimo do REM e do RDM nesse sistema? Qual deve ser a máxima quantidade de bits que podem ser implementados nessa memória?
    Resposta:
    REM = 14 bits M = 12 bits T = 192K bits
    RDM = 12 bits N = 16K endereços

17. Descreva os barramentos que interligam a UCP à MP, indicando a função e direção do fluxo de sinais de cada um.

18. Descreva passo a passo uma operação de leitura.

19. Descreva passo a passo uma operação de escrita.

20. Qual a diferença entre as memórias ROM, PROM, EPROM e EEPROM?

21. Uma memória ROM é uma RAM? Explique.