Memórias (Formatos)

A memória RAM é um componente essencial não apenas nos PCs, mas em qualquer tipo de computador. Por mais que exista espaço de armazenamento disponível, na forma de um HD ou memória flash, é sempre necessária uma certa quantidade de memória RAM e, naturalmente, quanto mais melhor.

Graças ao uso da memória swap, é possível rodar a maioria dos sistemas operacionais modernos com quantidades relativamente pequenas de memória. No caso do Linux, é possível inicializar uma instalação enxuta (em modo texto, com pouca coisa além do Kernel e o interpretador de comandos) com apenas 4 MB de memória. O problema é que com pouca memória o sistema fica extremamente lento, como qualquer um que já tentou usar o Windows XP ou uma distribuição Linux recente, com o Gnome ou KDE em um PC com menos de 128 MB de memória pode dizer. :)

A sigla "RAM" vem de "Random Access Memory", ou "memória de acesso aleatório", indicando a principal característica da memória RAM, que é o fato de permitir o acesso direto a qualquer um dos endereços disponíveis e de forma bastante rápida.

Ao carregar um programa, ele é lido no HD (ou outra mídia de armazenamento) e é transferido para a memória RAM, para só então ser executado pelo processador. A memória RAM oferece tempos de acesso brutalmente mais baixos que o HD e trabalha com taxas de transferência muito mais altas, mas possui a desvantagem de perder os dados armazenados quando o micro é desligado, daí a necessidade de salvar os arquivos periodicamente.

É também por causa disso que o processo de boot é refeito cada vez que você liga o micro. Durante o boot, o sistema operacional, drivers, bibliotecas e aplicativos são novamente copiados para a memória, junto com suas configurações e preferências.

A única forma de evitar repetir o demorado processo de boot é manter a memória RAM ativa, ou salvar seu conteúdo no HD, recuperando-o no próximo boot. Essas são as estratégias usadas pelas opções de suspender e hibernar, disponíveis tanto no Windows quanto em várias distribuições Linux.

Ao suspender, a maioria dos componentes do sistema são desligados, incluindo o HD, a placa de vídeo e a maior parte dos componentes da placa-mãe. Mesmo o processador entra em um estágio de baixo consumo, onde a maior parte dos componentes internos são desativados e o clock é reduzido. Praticamente, os únicos componentes que continuam realmente ativos são os módulos de memória. Graças a isso o PC acaba consumindo (geralmente) menos de 20 watts de energia e pode voltar ao estágio original muito rapidamente.

Ao hibernar, o conteúdo da memória RAM é copiado para uma área reservada do HD e o micro é desligado. Ao ligar novamente, o conteúdo da memória é restaurado e temos o sistema de volta, sem precisar passar pelo processo normal de boot. O problema da hibernação é que a restauração demora muito mais tempo, já que é necessário ler 512 MB, 1 GB ou mesmo 4 GB de dados (equivalentes à quantidade de memória RAM instalada) a partir do HD, o que muitas vezes demora mais do que um boot completo. :)

Além dos diferentes tipos de memória RAM, existem também outras tecnologias de memórias de acesso aleatório, como as SRAM e mais recentemente as MRAM. Temos ainda as onipresentes memórias Flash (que veremos em detalhes mais adiante), que concorrem com os HDs como mídia de armazenamento.

O tipo mais comum de memória RAM, aquela que compramos na forma de módulos e instalamos na placa-mãe, é chamada de DRAM, ou "dynamic RAM". Como vimos no capítulo 1, a memória DRAM passou a ser usada apenas a partir do final da década de 70, substituindo os chips de memória SRAM, que eram muito mais caros. Com o passar do tempo, as memória DRAM viraram o padrão, de forma que geralmente dizemos apenas "memória RAM" e não "memória DRAM".

Num chip de memória DRAM, cada bit é formado pelo conjunto de um transístor e um capacitor. O transístor controla a passagem da corrente elétrica, enquanto o capacitor a armazena por um curto período. Quando o capacitor contém um impulso elétrico, temos um bit 1 e quando ele está descarregado, temos um bit 0.

Quando falo em "capacitor", tenha em mente que não estamos falando em nada similar aos capacitores eletrolíticos da placa-mãe. Os "capacitores" usados nos chips de memória são extremamente pequenos e simples, basicamente dois pequenos blocos de metal ligados ao transístor, que conservam o impulso elétrico por apenas uma fração de segundo.

Para evitar a perda dos dados, a placa-mãe inclui um circuito de refresh, que é responsável por regravar o conteúdo da memória várias vezes por segundo (a cada 64 milessegundos ou menos), algo similar ao que temos num monitor CRT, onde o canhão de elétrons do monitor precisa atualizar a imagem várias vezes por segundo para evitar que as células de fósforo percam seu brilho.

O processo de refresh atrapalha duplamente, pois consome energia (que acaba sendo transformada em calor, contribuindo para o aquecimento do micro) e torna o acesso à memória mais lento. Apesar disso, não existe muito o que fazer, pois a única solução seria passar a usar memória SRAM, que é absurdamente mais cara.

A principal diferença é que na memória SRAM cada célula é formada por 4 ou 6 transístores, em vez de apenas um. Dois deles controlam a leitura e gravação de dados, enquanto os demais formam a célula que armazena o impulso elétrico (a célula continua armazenando um único bit). As memórias SRAM são muito mais rápidas e não precisam de refresh, o que faz com que também consumam pouca energia. Além de ser usada como memória cache, a memória SRAM é muito usada em palmtops e celulares, onde o consumo elétrico é uma questão crítica.

Seria perfeitamente possível construir um PC que usasse memória SRAM como memória principal, mas o custo seria proibitivo. Foi por causa do custo que as memórias DRAM passaram a ser utilizadas em primeiro lugar.

Mesmo utilizando um único transístor por bit, os módulos de memória RAM são formados por um número assustador deles, muito mais que os processadores e outros componentes. Um módulo de memória de 1 GB, por exemplo, é formado geralmente por 8 chips de 1 gigabit cada um (8 gigabits = 1 gigabyte). Cada chip possui então mais de 1 bilhão de transístores e capacitores e o módulo inteiro acumula mais de 8 bilhões de conjuntos.

Apesar dessa brutal quantidade de transistores, os chips de memória são relativamente simples de se produzir, já que basta repetir a mesma estrutura indefinidamente. É muito diferente de um processador, que além de ser muito mais complexo, precisa ser capaz de operar a freqüências muito mais altas.

Com a evolução nas técnicas de fabricação, os módulos de memória foram ficando cada vez mais baratos com o passar das décadas. Na época dos micros 486, chegava-se a pagar 40 dólares por megabyte de memória, valor que hoje em dia compra um módulo de 512 MB (ou até mais). O problema é que os requisitos dos sistemas operacionais e aplicativos também aumentaram, quase que na mesma proporção. Enquanto o MS-DOS rodava bem com 2 ou 4 MB de memória, o Windows 95 já precisava de pelo menos 16 MB. O Windows XP (assim como a maioria das distribuições Linux atuais) não roda bem com menos de 256 MB, enquanto no Vista o ideal é usar 1 GB ou mais.

Na maioria das situações, ter uma quantidade suficiente de memória RAM instalada é mais importante que o desempenho do processador, pois sem memória RAM suficiente o sistema passa a utilizar memória swap, que é absurdamente mais lenta.

Enquanto uma seqüência de 4 leituras em um módulo de memória DDR2-800 demora cerca de 35 bilionésimos de segundo, um acesso a um setor qualquer do HD demora pelo menos 10 milésimos. A taxa de transferência nominal do mesmo módulo de memória é de 6.4 GB/s, enquanto mesmo um HD rápido, de 7200 RPM tem dificuldades para superar a marca de 60 MB/s, mesmo lendo setores seqüenciais. Ou seja, a memória RAM possui nesse caso um tempo de acesso quase 300.000 vezes menor e uma taxa de transferência contínua mais de 100 vezes maior que o HD.

Se lembrarmos que a memória RAM já é muito mais lenta que o processador (justamente por isso temos os caches L1 e L2), fica fácil perceber o quanto o uso de memória swap por falta de memória RAM física pode prejudicar o desempenho do sistema.

É fácil monitorar o uso de swap. No Windows XP ou Vista basta pressionar Ctrl+Alt+Del e acessar o gerenciador de tarefas, enquanto no Linux você pode usar o comando "free" ou um aplicativo de gerenciamento, como o ksysguard.

No caso do Windows Vista é possível usar um pendrive como memória adicional, através do ReadyBoost. Neste caso entretanto, o pendrive é usado como uma extensão da memória swap e não como um substituto da memória RAM. Como o pendrive oferece tempos de acesso muito mais baixos, ele acaba sendo mais eficiente que o HD nessa tarefa, muito embora a taxa de leitura seja geralmente mais baixa.

Esse recurso pode ajudar em micros com pouca memória RAM e também reduzir o tempo de carregamento dos programas. É uma opção para casos em que você já tem o pendrive e procura um uso para ele, mas não espere milagres. Em se tratando de memória, não existe o que inventar: ou você procura um sistema operacional e programas mais leves, ou compra mais memória. Não dá para ficar em cima do muro. ;)

Como disse há pouco, embora seja brutalmente mais rápida que o HD e outros periféricos, a memória RAM continua sendo muito mais lenta que o processador. O uso de caches diminui a perda de desempenho, reduzindo o número de acessos à memória; mas, quando o processador não encontra a informação que procura nos caches, precisa recorrer a um doloroso acesso à memória principal, que pode demorar o equivalente a mais de 100 ciclos do processador.

Para reduzir a diferença (ou pelo menos tentar impedir que ela aumente ainda mais), os fabricantes de memória passaram a desenvolver um conjunto de novas tecnologias, a fim de otimizar o acesso aos dados. Acompanhando essas mudanças, tivemos também alterações físicas no formato dos módulos, de forma que podemos classificar os módulos de memória de duas formas:

* Quanto à tecnologia usada (EDO, SDRAM, DDR, DDR2, etc.)

* Quanto ao formato usado (SIMM, DIMM, etc.)

**Formatos**

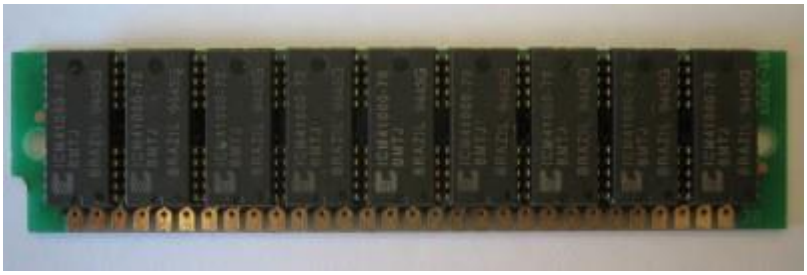
Nos micros XT, 286 e nos primeiros 386, ainda não eram utilizados módulos de memória. Em vez disso, os chips de memória eram instalados diretamente na placa-mãe, encaixados individualmente em colunas de soquetes (ou soldados), onde cada coluna formava um banco de memória.

Esse era um sistema antiquado, que trazia várias desvantagens, por dificultar upgrades de memória ou a substituição de módulos com defeito. Imagine você, fazendo um upgrade de memória numa placa como esta:

Não é só você que não achou muito atraente a idéia de ficar catando chips de memória um a um. Foi questão de tempo até que alguém aparecesse com uma alternativa mais prática, capaz de tornar a instalação fácil até mesmo para usuários inexperientes.

Os módulos de memória são pequenas placas de circuito onde os chips DIP são soldados, facilitando o manuseio e a instalação.

Os primeiros módulos de memória criados são chamados de módulos SIMM, sigla que significa "Single In Line Memory Module", justamente porque existe uma única via de contatos, com 30 vias. Apesar de existirem contatos também na parte de trás do módulo, eles servem apenas como uma extensão dos contatos frontais, de forma a aumentar a área de contato com o soquete. Examinando o módulo, você verá um pequeno orifício em cada contato, que serve justamente para unificar os dois lados.



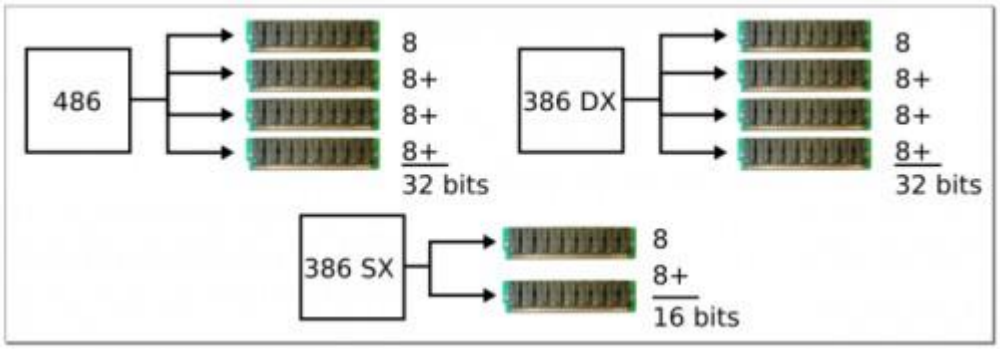
*Módulo SIMM de 30 vias*

Os módulos de 30 vias possuíam sempre 8 ou 9 chips de memória. Cada chip fornecia um único bit de dados em cada transferência, de forma que 8 deles formavam um módulo capaz de transferir 8 bits por ciclo. No caso dos módulos com 9 chips, o último era destinado a armazenar os bits de paridade, que melhoravam a confiabilidade, permitindo identificar erros. Hoje em dia os módulos de memória são mais confiáveis, de forma que a paridade não é mais usada. No lugar dela, temos o ECC, um sistema mais avançado, usado em módulos de memória destinados a servidores.

Os módulos de 30 vias foram utilizados em micros 386 e 486 e foram fabricados em várias capacidades. Os mais comuns foram os módulos de 1 MB, mas era possível encontrar também módulos de 512 KB, 2 MB e 4 MB. Existiram também módulos de 8 e 16 MB, mas eles eram muito raros devido ao custo.

Os processadores 386 e 486 utilizavam um barramento de 32 bits para o acesso à memória, o que tornava necessário combinar 4 módulos de 30 vias para formar um banco de memória. Os 4 módulos eram então acessados pelo processador como se fossem um só. Era preciso usar os módulos em quartetos: 4 módulos ou 8 módulos, mas nunca um número quebrado.

A exceção ficava por conta dos micros equipados com processadores 386SX, onde são necessários apenas 2 módulos, já que o 386SX acessa a memória usando palavras de 16 bits:

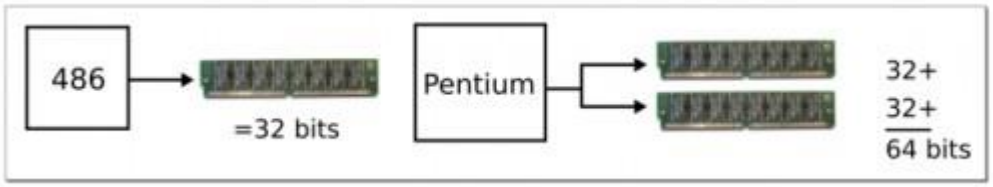



Apesar de serem muito mais práticos do que manipular diretamente os chips DIP, os módulos SIMM de 30 vias ainda eram bastante inconvenientes, já que era preciso usar 4 módulos idênticos para formar cada banco de memória. Eles foram desenvolvidos pensando mais na questão da simplicidade e economia de custos do que na praticidade.

Para solucionar o problema, os fabricantes criaram um novo tipo de módulo de memória SIMM, de 32 bits, que possui 72 vias. Os módulos de 72 vias substituíram rapidamente os antigos nas placas para 486 e se tornaram o padrão nos micros Pentium, sendo em seguida substituídos pelos módulos de 168 vias.

*Módulo SIMM de 72 vias*

Em vez de quatro módulos, é preciso apenas um módulo SIMM de 72 vias para formar cada banco de memória nos micros 486. Como o Pentium acessa a memória usando palavras de 64 bits, são necessários 2 módulos em cada banco. É por isso que nos micros Pentium 1 precisamos sempre usar os módulos de memória em pares:




O acesso de 64 bits à memória foi introduzido para permitir que o processador conseguisse acessar grandes quantidades de dados mais rapidamente. O processador é tão mais rápido que a memória RAM, que depois de esperar vários ciclos para poder acessá-la, o melhor a fazer é pegar a maior quantidade de dados possível e guardar tudo no cache. Naturalmente os dados serão processados em blocos de 32 bits, mas a poupança ajuda bastante.

Dentro de um banco, todos os módulos são acessados ao mesmo tempo, como se fossem um só, por isso era sempre recomendável usar dois módulos iguais. Ao usar quatro módulos, o importante era que cada par fosse composto por dois módulos iguais. Não existia problema em usar dois pares de módulos diferentes, como ao usar dois de 16 MB e mais dois de 8 MB para totalizar 48 MB, por exemplo.

Uma curiosidade é que algumas placas-mãe para Pentium podem trabalhar com apenas um módulo de 72 vias. Nesse caso, a placa engana o processador, fazendo dois acessos de 32 bits consecutivos, entregando os dados de uma só vez para o processador. Apesar de funcionar, esse esquema reduz bastante a velocidade do micro, pois a taxa de transferência ao ler dados a partir da memória é efetivamente reduzida à metade.

Finalmente, temos os módulos **DIMM**, usados atualmente. Ao contrário dos módulos SIMM de 30 e 72 vias, os módulos DIMM possuem contatos em ambos os lados do módulo, o que justifica seu nome, "Double In Line Memory Module" ou "módulo de memória com dupla linha de contato".

Todos os módulos DIMM são módulos de 64 bits, o que eliminou a necessidade de usar 2 ou 4 módulos para formar um banco de memória. Muitas placas-mãe oferecem a opção de usar dois módulos (acessados simultaneamente) para melhorar a velocidade de acesso. Esse recurso é chamado de **dual-channel** e melhora consideravelmente o desempenho, sobretudo nas placas-mãe com vídeo onboard, onde a placa de vídeo disputa o acesso à memória RAM com o processador principal. De qualquer forma, mesmo nas placas dual-channel, usar os módulos em pares é opcional; você pode perfeitamente usar um único módulo, mas neste caso o suporte a dual-channel fica desativado.

Existem três formatos de memória DIMM. Os mais antigos são os módulos de memória **SDR**, de 168 vias, que eram utilizados há até poucos anos. Em seguida, temos os módulos de memória **DDR**, que possuem 184 contatos e os módulos **DDR2**, que possuem 240.

Apesar do maior número de contatos, os módulos DDR e DDR2 são exatamente do mesmo tamanho que os módulos SDR de 168 vias, por isso foram introduzidas mudanças na posição dos chanfros de encaixe, de forma que você não consiga encaixar os módulos em placas incompatíveis.

Os módulos SDR possuem dois chanfros, enquanto os DDR possuem apenas um chanfro, que ainda por cima é colocado em uma posição diferente:

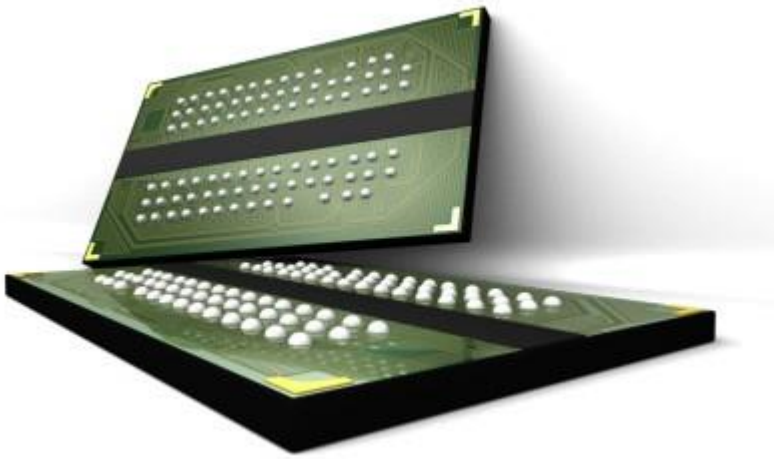
*Módulo DIMM SDR (em cima) e módulo DDR*

Os módulos DDR2 também utilizam um único chanfro, mas ele está posicionado mais próximo do canto do módulo que o usado nos módulos DDR, de forma que é novamente impossível encaixar um módulo DDR2 numa placa antiga:

*Módulo DIMM DDR2*

Isso é necessário, pois além das mudanças na forma de acesso, os módulos DDR2 utilizam tensão de 1.8V, enquanto os módulos DDR usam 2.5V. Se fosse possível instalar um módulo DDR2 em uma placa antiga, a maior tensão queimaria o módulo rapidamente.

Outra diferença é que os chips DDR2 utilizam o encapsulamento BGA (Ball Grid Array), no lugar do encapsulamento TSOP (Thin Small-Outline Package), usado nos chips SDR e DDR. A grande diferença é que no BGA os pontos de solda são posicionados diretamente na parte inferior dos chips, em vez de serem usadas as "perninhas" laterais. Isso reduz a distância que o sinal elétrico precisa percorrer, além de reduzir o nível de interferências, permitindo que os módulos sejam capazes de operar a freqüências mais altas. Esta imagem ilustrativa da Micron mostra bem como os chips se parecem:



*Chips BGA de memória*

Mais recentemente surgiram no mercado alguns módulos de memória DDR que também utilizam chips BGA, mas eles são menos comuns.

Como os módulos DDR2 trabalham a freqüências mais altas, o uso de dissipadores se tornou mais comum. Eles não são realmente necessários, mas a melhor dissipação do calor permite que o módulo trabalhe a freqüências mais altas, por isso eles se tornaram norma nos módulos DDR2 de alto desempenho e, principalmente, nos módulos "premium", destinados a overclock. Alguns fabricantes chegam a utilizar heat-pipes ou a oferecer coolers ativos, que podem ser instalados sobre os módulos.


*Módulos DDR2 com dissipadores*

Outra característica que torna os módulos DDR2 diferentes é a presença de um terminador resistivo dentro de cada chip de memória. O terminador é necessário para "fechar o circuito", evitando que os sinais elétricos retornem na forma de interferência ao chegarem ao final do barramento. Nos módulos DDR os terminadores são instalados na placa-mãe, o que torna a terminação menos eficiente. Como os módulos DDR2 operam a freqüências muito mais altas, a presença do terminador dentro dos próprios chips se tornou uma necessidade, já que torna o sinal mais estável e livre de ruídos.

Existem também os módulos SODIMM (Small Outline DIMM), destinados a notebooks. Eles são basicamente versões miniaturizadas dos módulos destinados a desktops, que utilizam os mesmos tipos de chips de memória.

Os módulos SODIMM SDR possuem 144 pinos, enquanto os módulos DDR e DDR2 possuem 200 pinos. Nos módulos SDR o chanfro fica próximo ao centro do módulo, enquanto nos DDR e DDR2 ele fica à esquerda. Assim como nos módulos para desktops, existe uma pequena diferença no posicionamento do chanfro entre os módulos DDR e DDR2, que impede o encaixe incorreto, já que ambos são incompatíveis.


*Módulo SODIMM DDR2*