FACULDADE PITÁGORAS

PRONATEC

DISCIPLINA: ARQUITETURA DE COMPUTADORES

Prof. Ms. Carlos José Giudice dos Santos

carlos@oficinadapesquisa.com.br

www.oficinadapesquisa.com.br

APOSTILA V

MEMÓRIA PRINCIPAL

Objetivos

Ao final desta apostila, o aluno deverá ser capaz de:

1. Saber a diferença entre memória volátil e não volátil
2. Saber o que é memória RAM e suas principais características.
3. Saber o que é memória ROM e suas variações.

ESQUEMA EXTERNO DE UM MICROCOMPUTADOR

Agora que já conhecemos o funcionamento básico de uma CPU, vamos conhecer um outro componente importante de um microcomputador:

Microcomputador

(Organização Externa)

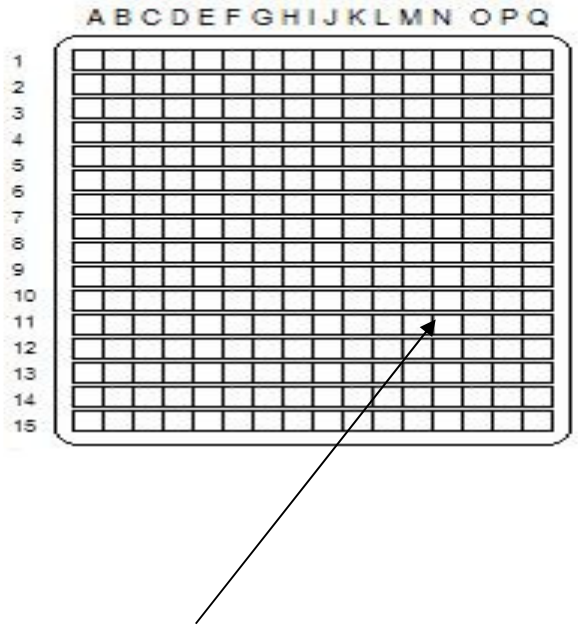
- Teclado / Mouse / Monitor de Vídeo
- Gabinete + Fonte de Alimentação
- Placa mãe (Motherboard ou Mainboard)
- Processador
- → **Memória Principal (RAM)** ←
- Placa de Vídeo
- Placas de Comunicação (Modem / Rede)
- Placa de Som
- Disco Rígido
- Outros dispositivos de armazenamento

MEMÓRIA PRINCIPAL - I

A memória principal de computador é conhecida como Random Access Memory (RAM) – Memória de Acesso Randômico (ou Aleatório). Uma memória desse tipo pode ser acessada em qualquer posição. Para isso, basta saber o endereço dessa posição.

Em oposição a esse tipo de memória, temos a memória de acesso sequencial (fitas). Por exemplo, se quisermos acessar uma determinada posição de um filme em uma fita de vídeo, temos que colocar a fita no início e avançar por todas as posições desde o zero até a posição desejada, ou seja, o acesso a este tipo de memória é sequencial e muito lento. Veja a figura a seguir:

MEMÓRIA PRINCIPAL - II



Posição N11

Na figura acima (exemplo de acesso sequencial), para acessarmos a posição 34 da fita, temos que passar por todas as posições desde a posição 00 até a posição desejada.

Na figura ao lado (exemplo de acesso aleatório), cada célula (por exemplo, posição N11) tem um endereço que pode ser acessado diretamente sem ter que passar pelos endereços anteriores.

MEMÓRIA CACHE

As primeiras CPU's faziam acesso síncrono à memória principal. A evolução tecnológica permitiu um aumento excepcional da velocidade de processamento das CPU's. A memória principal não acompanhou esse ritmo de evolução. Uma solução possível seria aumentar o número de registradores, mas o limite desta técnica é facilmente atingido.

Hoje a memória principal (RAM) de um computador tem um tempo de acesso (latência) grande em relação à velocidade de uma CPU (acesso assíncrono). Por esse motivo, as CPU's passaram a incorporar uma memória interna mais rápida chamada de CACHE (do francês cacher → esconder).

FUNCIONAMENTO DA CACHE - I

- CACHE é um trecho de memória que se interpõe entre a memória principal (mais lenta) e a requisição de dados e/ou instruções do sistema pela CPU (muito rápida).
- Quando o sistema faz um acesso à RAM, todo um trecho de memória é copiado para a memória CACHE, pois simulações demonstram que a possibilidade da próxima requisição do sistema acontecer na vizinhança da requisição anterior pode ser superior a 95% (dependendo do tamanho do trecho de RAM copiado).

FUNCIONAMENTO DA CACHE - II

- Assim, quando o próxima requisição de dados ou instruções acontecer, o sistema operacional vai verificar primeiro na CACHE (que é muito mais rápida que a memória RAM).
- Caso esta requisição não encontre aquilo que ela necessita na cópia da RAM mantida em CACHE, então um novo ciclo de cache se inicia, acessando o dado ou instrução necessário na memória principal e copiando um novo trecho de RAM para a memória CACHE.

MEMÓRIA CACHE

RESUMINDO:

Quanto maior o tamanho da memória CACHE

- Maior é a velocidade de processamento (desempenho geral do sistema)
- Maior é o custo (preço do processador)
- Maior é o consumo de energia (e consequentemente, a quantidade de calor dissipado).
- Maior é o tamanho do processador

TIPOS DE CACHE - I

Existem dois tipos de cache: unificada (ou compartilhada) e dividida.

Na cache unificada há apenas uma cache para dados e instruções, ou seja, a mesma área de memória rápida é compartilhada ao mesmo tempo para dados e instruções.

Isso acaba sendo um problema, pois a cache se comunica com a UC (Unidade de Controle) e a ULA (Unidade de Lógica e Aritmética) por meio de um único barramento interno (de dados, usado tanto para dados como para instruções).

TIPOS DE CACHE - II

Assim, quando uma série de instruções estão em um pipeline, uma requisição de dados pela Unidade de Busca de Operandos impede que aconteça, ao mesmo tempo, que a Unidade de Busca de Instrução busque uma instrução, porque há um só barramento que se comunica com a CACHE. Se o barramento está ocupado buscando um operando, logo não tem como a outra unidade buscar uma instrução utilizando esse mesmo barramento simultaneamente. Em uma cache dividida, não ocorre esse problema, pois há uma cache somente para dados e outra somente para instruções, com barramentos diferentes.

TIPOS DE CACHE - III

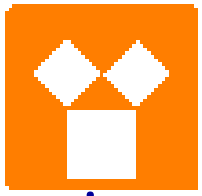
Além disso, uma instrução não se modifica durante a sua execução, o que significa que não haverá necessidade de armazenar a instrução após a sua execução. Isto significa que o tamanho de cache para instruções é significativamente menor que o tamanho de cache para dados.

O mesmo não acontece com os dados, pois uma instrução modifica esse dado (processamento) e os resultados (intermediários ou definitivos) tem que ser armazenados em algum lugar. É por este motivo que todo pipeline possui uma unidade de gravação.

NÍVEIS DE CACHE - I

O primeiro processador que se beneficiou de uma placa mãe com cache foi o Intel 80386. Era uma cache externa ao processador, mas com um tempo de acesso menor que o da RAM (algo como metade ou $\frac{1}{4}$ da velocidade do clock do processador).

A primeira evolução aconteceu com o Intel 486, primeiro processador a adotar uma cache interna de 8 KB. Essa cache interna funcionava na mesma velocidade do clock do processador. Uma cache interna com esse nível de velocidade é chamada de Cache Nível 1 (Cache Level 1, em inglês), ou simplesmente, Cache L1. Portanto, antigamente, a L1 era interna, de pequena capacidade e muito rápida.



NÍVEIS DE CACHE - II

Algumas placas mãe mantiveram uma memória cache externa, maior em capacidade de armazenamento que a cache interna do processador. Essa cache externa, mais lenta que a cache interna e mais rápida que a RAM é chamada de Cache Nível 2, ou Cache L2, ou simplesmente L2. Posteriormente, com a evolução da tecnologia de construção de processadores, a L2 migrou para dentro do processador, o que a tornou bem mais rápida. Aconteceu aqui uma segunda evolução: a cache L1, de pequena capacidade, passou a ser cache de instruções, enquanto que a cache L2, de maior capacidade, passou a ser a cache de dados. Foi deste modo que a cache deixou de ser compartilhada e passou a ser dividida.

NÍVEIS DE CACHE - III

Alguns fabricantes de placas-mãe mantiveram uma cache na placa mãe, intermediária entre as caches L1 e L2 e a memória principal. Eles chamaram esta cache de Nível 3 ou cache L3. Posteriormente (ou seja, hoje), a cache L3 (que armazena tanto dados como instruções) migrou para dentro do processador. Hoje não existe mais cache de memória principal em placas-mãe.

Resumindo, os processadores atuais (2011) possuem:

Cache L1 → interna, de pequena capacidade, muito rápida, dividida, utilizada para armazenar instruções.

NÍVEIS DE CACHE - IV

Cache L2 → interna, maior que L1, dividida, muito rápida, usada para armazenar dados.

Cache L3 → quando existe, ela é interna, compartilhada, maior que L2, armazena tanto dados como instruções.

OBS.1: Quando se fala que a cache L1 é pequena, estamos comparando com a memória RAM. Os processadores atuais chegam a possuir cache L1 de 1 MB de capacidade, ou seja, a mesma capacidade total de RAM do primeiro IBM PC 8086 / 8088.

NÍVEIS DE CACHE - V

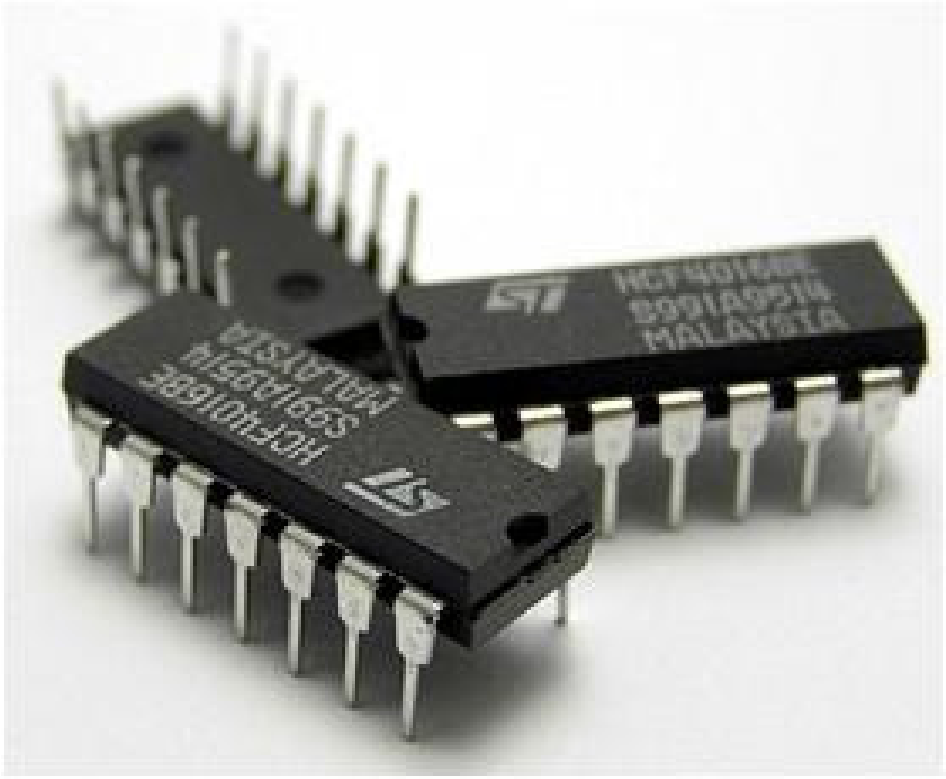
OBS.2: A capacidade da cache L2 subiu nos processadores atuais (maior que 20 MB em alguns processadores).

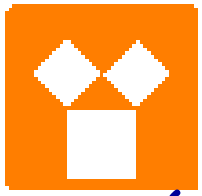
OBS.3: Quando se fala em tamanho de cache interna ao processador, estamos nos referindo à L3 (quando ela existe). Caso contrário, estamos no referindo à cache L2 (caso mais comum)

Na prática, hoje fala-se mais em nível de cache (L1, L2 ou L3) do que em tipo de cache (unificada ou dividida) pois os fabricantes já adotaram um padrão (L1 p/ instruções, L2 para dados e L3 para ambos).

MÓDULOS DE MEMÓRIA - I

Os primeiros computadores possuíam memória RAM em chips que eram espetados na placa mãe. Esse tipo de encapsulamento de memória é chamado DIP (*Dual Inline Package*). Ver figura abaixo:





MÓDULOS DE MEMÓRIA - II

A título de informação, existem diversos tipos de encapsulamento de memória (SOJ, TSOP, CSP, etc), mas vamos nos preocupar somente com os módulos de memória mais utilizados atualmente. Existem também diversas tecnologias de funcionamento das memórias: FPM (Fast Page Mode), EDO (Extended Data Output), SRAM (Static RAM) e RAMBUS. Todos esses padrões caíram em desuso. Por isso, não precisamos nos preocupar com eles. As memórias RAM atuais utilizam minúsculos capacitores para representar bits. Um capacitor carregado equivale a 1. Se estiver descarregado, equivale a zero. Este tipo de representação binária usando capacitores é muito eficiente, porque é muito rápida.

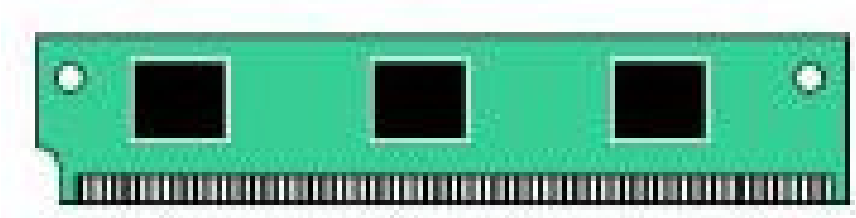
MÓDULOS DE MEMÓRIA - III

O problema desta tecnologia é que os capacitores se descarregam facilmente, sendo necessário um circuito que atualize o conteúdo destas memórias (refresh) periodicamente. Este tipo de memória é chamada SDRAM (Syncronous Dinamic RAM), porque a memória RAM é dinâmica (atualizada periodicamente) em sincronia com o clock do barramento. A maioria dos módulos atuais usam essa tecnologia.

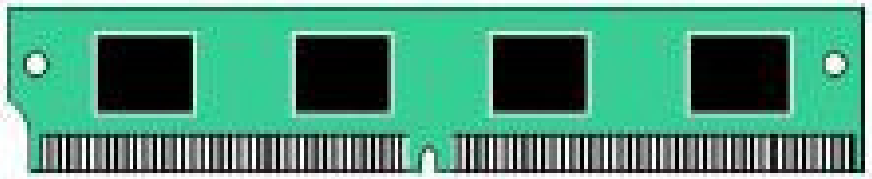
Os primeiros módulos de memória SDRAM usados em computadores foi o SIMM (Single Inline Memory Module – módulo de memória em linha única). Eles tem esse nome porque os contatos (ou pinos) estão apenas de um lado do módulo.

MÓDULOS DE MEMÓRIA - IV

Existem duas versões: o SIMM-30 (30 pinos - palavras de 16 bits) e o SIMM-72 (72 pinos - palavras de 32 bits).



SIMM-30



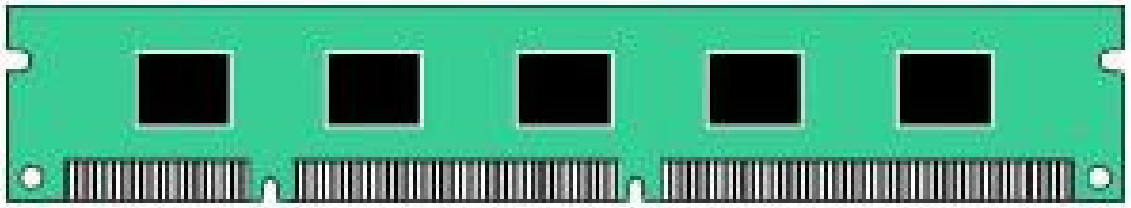
SIMM-72

MÓDULOS DE MEMÓRIA - V

A próxima evolução dos módulos de memória aconteceu no tamanho das palavras de memória (que passaram a ser de 64 bits), além de um consumo menor de energia e um pouco mais de velocidade no acesso às informações. Para ter tanta capacidade de memória em um espaço tão pequeno foi necessário colocar contatos dos dois lados do módulo. Nasce assim o módulo SDRAM DIMM. A sigla DIMM significa Dual Inline Memory Module – módulo de memória em linha dupla), ou seja, possui contatos dos dois lados do módulo. O primeiro módulo DIMM (com palavras de 64 bits) possuía 84 contatos (ou pinos) de cada lado do módulo. Veja figura no próximo slide.

MÓDULOS DE MEMÓRIA - VI

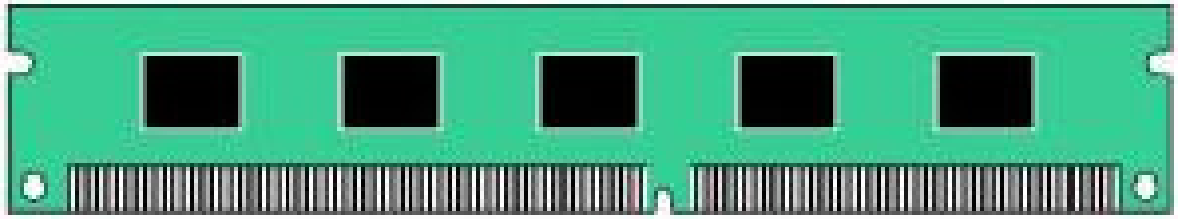
DIMM-168



Com a necessidade de memórias cada vez mais rápidas, surgiu uma nova tecnologia que permitiu taxas de transferência de dados muito mais rápidas entre a memória e o barramento. Essa nova tecnologia atende pelo nome de DDR (Double Data Rate – taxa dupla de transferência). Os novos módulos SDRAM DIMM DDR possuem 92 contatos de cada lado do módulo, totalizando 184 pinos. Como o próprio nome diz, esse tipo de memória consegue transferir o dobro (2^1 vezes mais rápida que uma DIMM comum) de dados em um mesmo ciclo de clock do barramento. Veja a figura no próximo slide.

MÓDULOS DE MEMÓRIA - VII

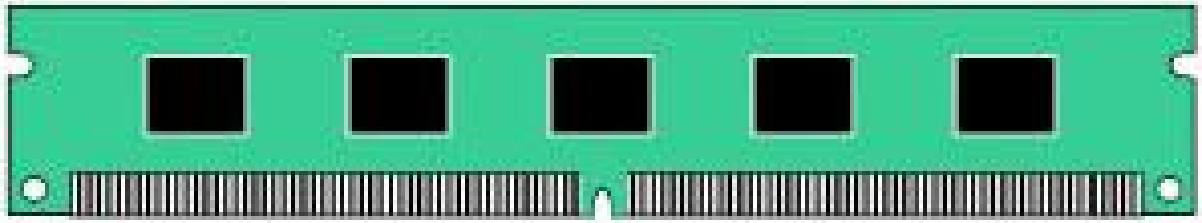
DIMM DDR -184



A evolução da tecnologia DDR permitiu construir memórias ainda mais rápidas. A memória SDRAM DIMM DDR2 é duas vezes mais rápida que a DDR, e quatro vezes (2^2) mais rápida que uma DIMM comum. Possui 120 contatos de cada lado do módulo, totalizando 240 pinos. A DDR3 é duas vezes (2^1) mais rápida que a DDR-2, quatro vezes (2^2) mais rápida que a DDR e oito vezes (2^3) mais rápida que uma DIMM comum. Também possui 240 pinos, mas os módulos DDR2 e DDR3, embora semelhantes, são eletricamente incompatíveis, ou seja, o slot (cavidade da placa mãe) que recebe um módulo DDR2 não funciona com um módulo DDR3 (e vice-versa).

MÓDULOS DE MEMÓRIA - VIII

DIMM DDR2 ou
DDR3 - 240 pinos



A memória RAM mais rápida atualmente é a DDR3. Entretanto existem memórias tão rápidas quanto a DDR3 (a DDR4 e a DDR5) que são exclusivas de placas gráficas de alto desempenho, funcionando em conjunto com as GPU (*Graphics Processing Unit* - Unidade de Processamento Gráfico), que são processadores construídos exclusivamente para trabalhar com instruções gráficas.

MÓDULOS DE MEMÓRIA - IX

Esse tipo de memória (DDR4 e DDR5) é diferente daqueles utilizados em módulos de memória RAM. A sigla correta deveria ser GDDR4 (Graphics DDR version 4) e GDDR5 (Graphics DDR version 5).

A GDDR4 é baseada na tecnologia da DDR2, sendo porém mais rápida, conseguindo taxas de transferência de até 16 GB/s com a GPU. A GDDR5 é baseada na tecnologia da DDR3, sendo também mais rápida, conseguindo taxas de até 28,2 GB/s com a GPU.

MÓDULOS DE MEMÓRIA - X

Existem diversas empresas que fabricam chips para memória (Samsung, Siemens, Fujitsu, OKI, etc), mas o fato de fabricarem os chips não significa que sejam fabricantes de módulos de memória completos.

É importante utilizar memórias de qualidade. Muitas empresas como a HP, Dell, SONY, etc, costumam encomendar lotes especiais para fabricantes e colocam seus selos.

Os fabricantes de módulos de memória de qualidade que eu conheço são a Kingston (www.kingston.com), a OCZ (www.ocz.com) e a Corsair (www.corsair.com).

MÓDULOS DE MEMÓRIA - XI

Usuários caseiros e e de pequenos escritórios (segmento SOHO - *Small Office Home Office*) não exigem computadores com mais de 1 GB de RAM. Em geral utilizam o Windows XP Home ou o Windows 7 Starter. Em caso de baixa performance do computador, a adição de mais 1 GB costuma resolver o problema, tornando sua performance adequada. Pequenos servidores e usuários que façam uso de aplicações gráficas (jogos, por exemplo) devem ter pelo menos 4 GB de RAM. Conforme a aplicação, pode ser necessário chegar até a 128 GB de RAM ou mais.

HIERARQUIA DE MEMÓRIA

Quanto mais em cima na lista a seguir, a memória é mais cara, mais rápida e com menor capacidade para guardar informações. Quanto mais em baixo, a memória é mais barata, mais lenta e com maior capacidade para gravar informações:

+ barata
+ lenta
+ capacidade

1. Registradores
2. Cache
3. Memória RAM
4. Disco rígido
5. Fitas DAT
6. Dispositivos ópticos?

+ cara
+ rápida
- capacidade

MEMÓRIA SOMENTE DE LEITURA

Já sabemos que a memória RAM é volátil (perde as informações quando o computador é desligado). Existe um tipo de memória que mantém a informação gravada mesmo quando o computador está desligado, e não estou falando dos discos rígidos (memória secundária). A memória a que me refiro é a ROM . Nos primeiros computadores, tanto o sistema operacional como o firmware vinham gravados juntos em uma memória chamada ROM (Read Only Memory). Neste tipo de memória, os bytes são gravados durante o processo de fabricação desta memória. Eles ficam gravados permanentemente e não podem ser modificados.

Com o advento dos sistemas operacionais em disco, apenas o firmware (BIOS) continuou a ser gravado em memórias apenas de leitura. Estas memórias também evoluíram: PROM, EPROM, E2PROM e Flash.

Outros tipos de ROM's

PROM: Neste tipo de memória (Programmable ROM), as informações não são "queimadas" (armazenadas) durante o processo de fabricação. Ela é fabricada como uma série de posições vazias que são armazenadas de forma permanente depois. Os fabricantes que compram este tipo de memória utilizam um gravador, que vai gravar as informações nesta memória, apenas uma vez. Uma vez gravada, o seu conteúdo não pode ser modificado.

EPROM: Neste tipo de memória (Erasable Programmable ROM), as informações são gravadas de forma permanente, mas podem ser apagadas e reprogramadas a partir da utilização de um gravador de EPROM. O processo para gravar é o mesmo da PROM. Para apagar, existe uma "janela" neste tipo de memória, que permite apagar as informações quando exposta a raios ultravioleta. Uma vez apagada, pode ser gravada novamente.

Outros tipos de ROM's

E2PROM: Neste tipo de memória (Electrically Erasable Programmable ROM – EEPROM ou E2PROM), as informações são gravadas de forma permanente, mas podem ser apagadas e reprogramadas sem ter que retirar o módulo da placa mãe. Para isto, basta aplicar uma tensão em um pino específico. As informações gravadas neste tipo de memória podem ser lidas infinitas vezes, mas ela não pode ser apagada e gravada indefinidamente. Dependendo da tecnologia utilizada em sua fabricação, existe um número máximo de ciclos de regravação, geralmente de 10 mil vezes (quanto utiliza a tecnologia SLC) e 1 milhão de vezes (quando utiliza a tecnologia MLC).

Flash: É uma evolução da memória E2PROM, que permite leituras e gravações mais rápidas.