Arquitetura e Organização de Computadores

Capítulo 5

Memória interna

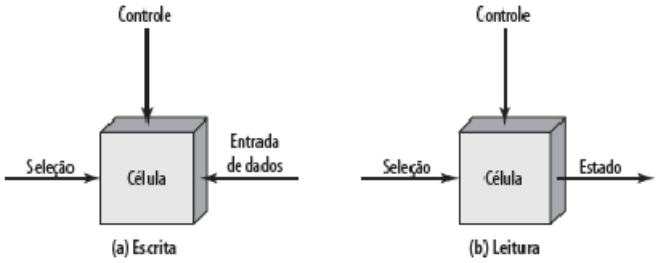
Tipos de memória de semicondutor

Tipo de memória	Categoria	Apagamento	Mecanismo de escrita	Volatilidade
Memória de acesso aleatório (RAM)	Memória de leitura-escrita	Eletricamente, em nível de byte	Eletricamente	Volátil
Memória somente de leitura (ROM)	Memória somente de leitura	Não é possível	Máscaras	Não volátil
ROM programável (PROM, do inglês *programmable ROM*)			Eletricamente	
PROM apagável (EPROM, do inglês *erasable PROM*)		Luz UV, nível de chip		
PROM eletricamente apagável (EEPROM, do inglês *electrically erasable PROM*)	Memória principalmente de leitura	Eletricamente, nível de byte		
Memória flash		Eletricamente, nível de bloco		

Memória de semicondutor

- RAM :
  - —Nome incorreto, pois toda memória de semicondutor tem acesso aleatório.
  - —Leitura/escrita.
  - —Volátil.
  - —Armazenamento temporário.
  - —Estática ou dinâmica.

Operação da célula de memória



RAM dinâmica

- Bits armazenados como carga em capacitores.
- As cargas vazam.
- Precisa de renovação mesmo se alimentada.
- Construção mais simples.
- Menor por bit.
- Mais barata.
- Precisa de circuitos de *refresh.*
- Mais lenta.
- Memória principal.
- Dispositivo basicamente analógico.
  —Nível de carga determina o valor.

Estrutura da RAM dinâmica

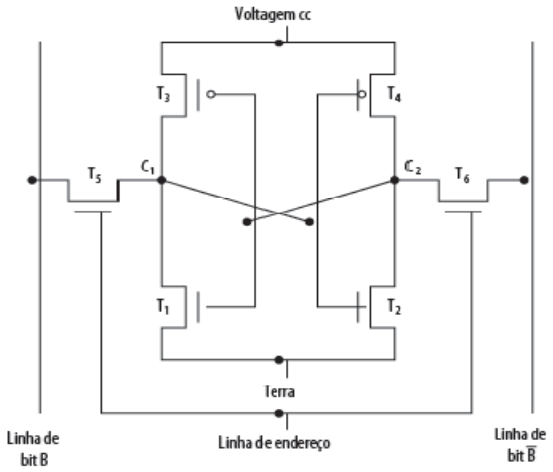
Operação da DRAM

- Linha de endereço ativa quando bit é **lido ou escrito**.
  - Chave de transistor fechada (corrente flui).
- **Escrita**:
  - Voltagem na linha de bit.
    - Alta para 1, baixa para 0.
  - Depois sinaliza linha de endereço.
    - Transfere carga ao capacitor.
- **Leitura**:
  - Linha de endereço selecionada.
    - Transistor liga.
  - Carga do capacitor alimentada por linha de bit para amplificador comparar.
    - Compara com valor de referência para determinar 0 ou 1.
  - Carga do capacitor deve ser restaurada.

RAM estática

- Bits armazenados como chaves ligado/desligado.
- Sem carga para vazar.
- Não precisa de *refresh* quando alimentada.
- Construção mais complexa.
- Maior por bit.
- Mais cara.
- Não precisa de circuitos de *refresh.*
- Mais rápida.
- Exemplo: Cache.
- Digital.
  - Usa flip-flops.

Estrutura da RAM estática



Operação da RAM estática

- Arranjo de transistores gera estado lógico estável.
- Estado 1:
  - —C_1 alta, C_2 baixa.
  - —T_1 T_4 desligados, T_2 T_3 ligados.
- Estado 0:
  - —C_2 alto, C_1 baixo.
  - —T_2 T_3 desligados, T_1 T_4 ligados.
- Linha de endereço controla dois transistores, T_5 T_6.
- Escrita – aplica valor a B e complemento de B.
- Leitura – valor está na linha B.

SRAM *versus* DRAM

- Ambas voláteis.
  - —É preciso energia para preservar os dados.
- Célula dinâmica:
  - —Mais simples de construir, menor.
  - —Mais densa.
  - —Mais barata.
  - —Precisa de *refresh.*
  - —Maiores unidades de memória.
- Estática:
  - —Mais rápida.
  - —Exemplo: Cache.

Read Only Memory (ROM)

- Armazenamento permanente.
  - —Não volátil.
- Microprogramação (ver mais adiante).
- Sub-rotinas de biblioteca.
- Programas do sistema (BIOS).
- Tabelas de função.

Tipos de ROM

- Gravada durante a fabricação:
  - —Muito cara para pequenas quantidades.
- Programável (uma vez):
  - —PROM.
  - —Precisa de equipamento especial para programar.
- Lida “na maioria das vezes”:
  - —Erasable Programmable (EPROM).
    - – Apagada por UV.
  - —Electrically Erasable (EEPROM):
    - – Leva muito mais tempo para escrever que para ler.
  - —Memória flash:
    - – Apaga memória inteira eletricamente.

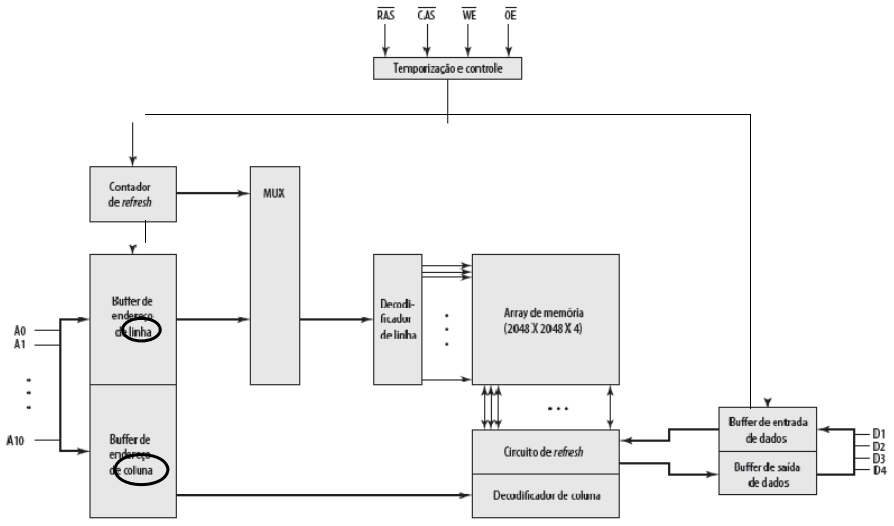
Organização em detalhes

- Um chip de 16 Mbits pode ser organizado como 1M de palavras de 16 bits.
- Um sistema de **um bit** por chip tem 16 lotes de chip de 1 Mbit com bit 1 de cada chip no chip 1, e assim por diante.
- Um chip de 16 Mbits pode ser organizado como um array de 2048 x 2048 x 4 bits.
  - —Reduz número de pinos de endereço.
    - – **Multiplexa endereço de linha e endereço de coluna**.
    - – 11 pinos para endereçar (2^{11}=2048).
    - – Aumentar um pino dobra o intervalo de valores, de modo que a capacidade multiplica por 4.

Refreshing

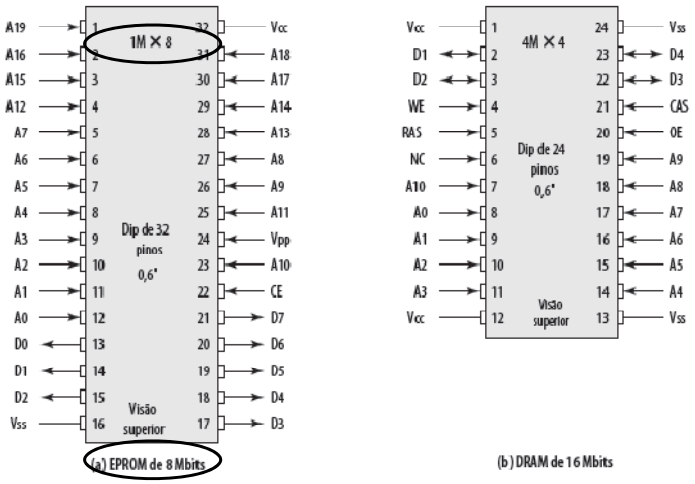
- Circuito de *refresh* incluído no chip.
- Desabilita chip.
- Conta por linhas.
- Lê e escreve de volta.
- Leva tempo.
- Atrasa o desempenho aparente.

DRAM típica de 16 Mb (4M x 4)

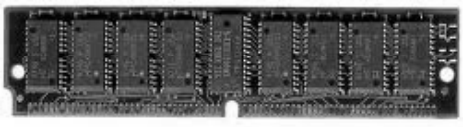


Empacotamento ou Encapsulamento

Chips de EPROM e DRAM típicos



Empacotamento ou Encapsulamento

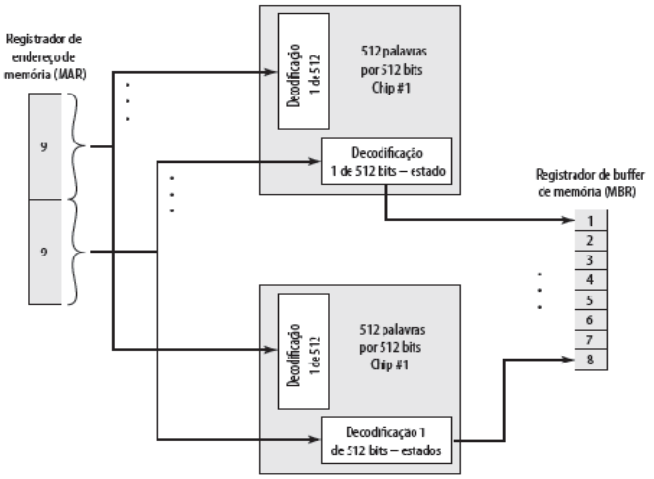


16 MB de memória DRAM

512 MB de SDRAM

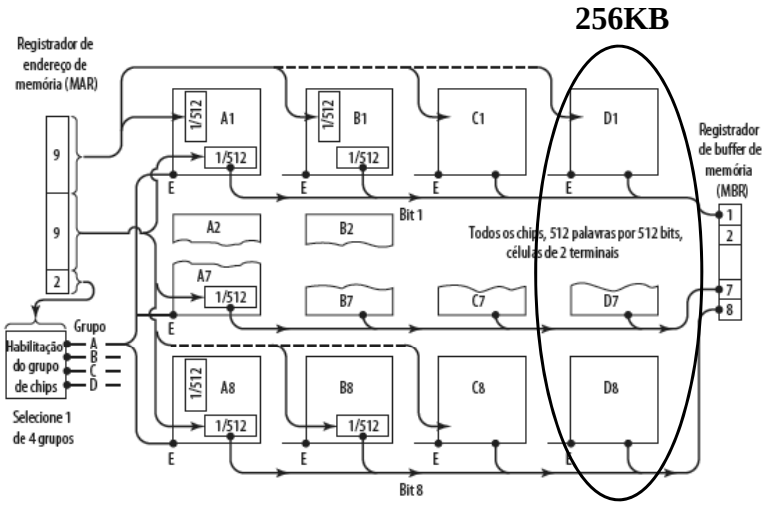
techfuels.com

Organização do módulo de 256 KBytes



8 Chips de 256K x 1 bit

Organização do módulo de 1 MBytes



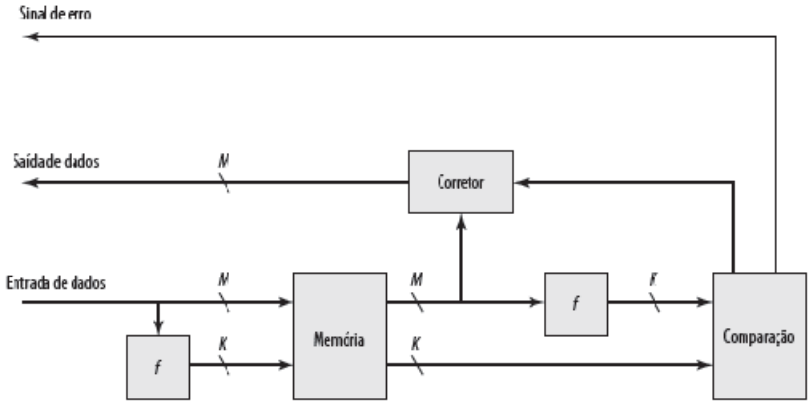
Memória intercalada

- Coleção de chips de DRAM.
- Agrupada em **banco de memória**.
- Bancos atendem a solicitações de leitura ou escrita independentemente.
- K bancos podem atender a k solicitações simultaneamente.

Correção de erro

- Falha permanente.
  - —Defeito permanente.
- Erro não permanente:
  - —Aleatório, não destrutivo.
  - —Sem dano permanente à memória.
- Detectado usando código de correção de erro de Hamming.

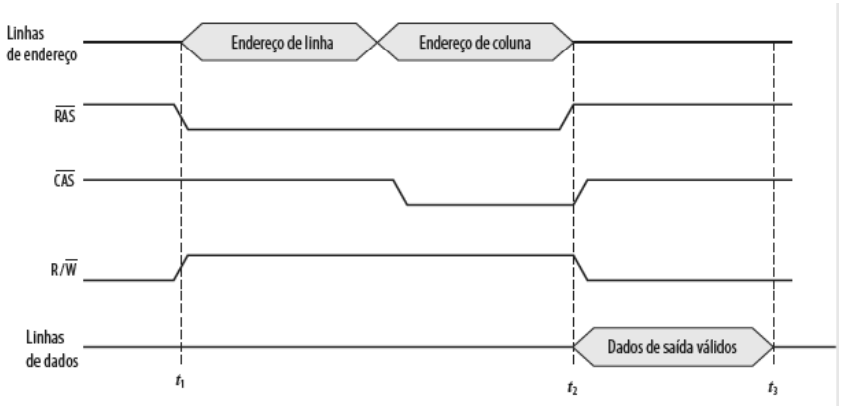
Função do código de correção de erro



Organização avançada da DRAM

- DRAM básica igual desde primeiros chips de RAM.
- DRAM avançada.
  - —Também contém pequena SRAM.
  - —SRAM mantém última linha lida.
- Cache DRAM:
  - —Maior componente em tamanho é a SRAM.
  - —Usa-se como cache ou buffer serial.

Temporização de leitura de DRAM simplificada



Desempenho de algumas alternativas a DRAM

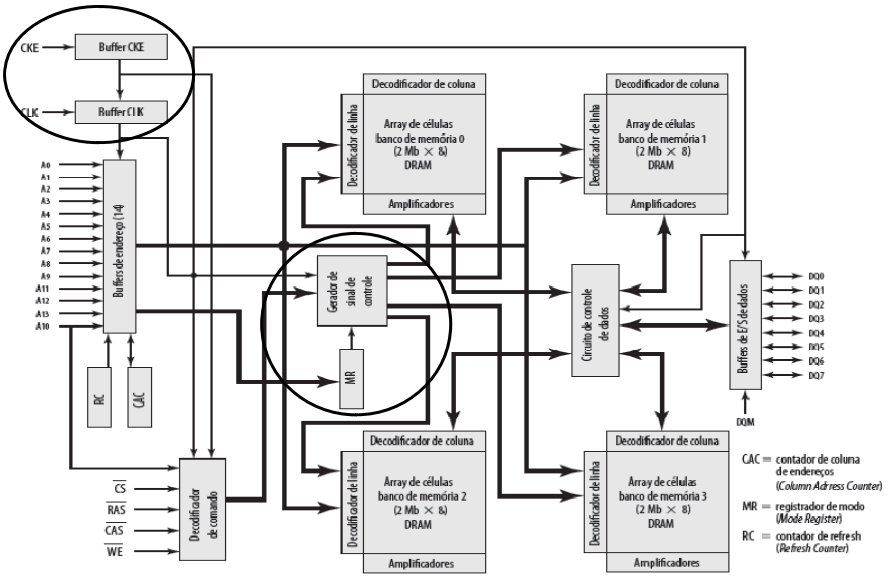
Tabela 5.3 Comparação do desempenho de algumas alternativas à DRAM

	Frequência de clock (MHz)	Taxa de transferência (GB/s)	Tempo de acesso (ns)	Contagem de pinos
SDRAM	166	1,3	18	168
DDR	200	3,2	12,5	184
RDRAM	600	4,8	12	162

DRAM síncrona (SDRAM)

- Acesso sincronizado com clock externo.
- Endereço é apresentado à RAM.
- RAM encontra dados (CPU espera na DRAM convencional).
- Como a SDRAM move dados em tempo com o clock do sistema, CPU sabe quando os dados estarão prontos.
- CPU não precisa esperar, e pode fazer alguma outra coisa.
- Modo de rajada permite que SDRAM defina fluxo de dados e o dispare em bloco.

SDRAM de 64 MB da IBM

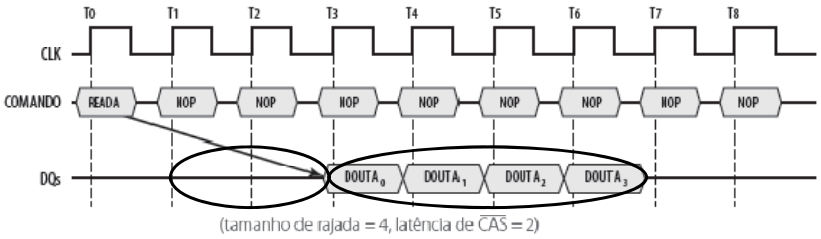


SDRAM

Tabela 5.4 Atribuições de pino da SDRAM

A0 a A13	Entradas de endereço
CLK	Entrada de clock
CKE	Habilitação de clock
$\overline{CS}$	Seleção se chip
$\overline{RAS}$	*Strobe* de endereço de linha
$\overline{CAS}$	*Strobe* de endereço de coluna
$\overline{WE}$	Habilitação de escrita
DQ0 a DQ7	Entrada/saída de dados
DQM	Máscara de dados

Temporização de leitura da SDRAM

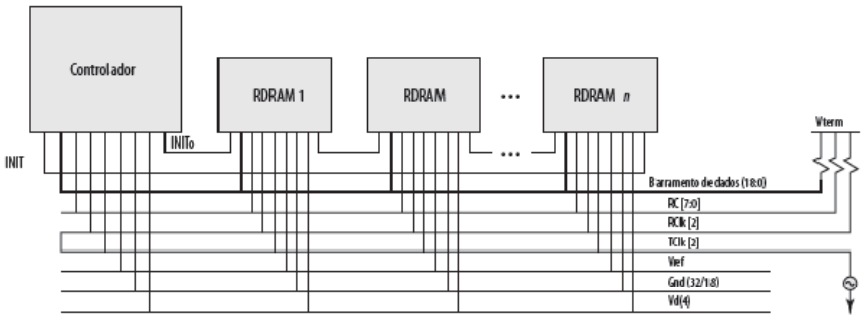


(tamanho de rajada = 4, latência de $\overline{CAS}$ = 2)

DRAM RamBus (ou RDRAM)

- Adotada pela Intel para Pentium & Itanium.
- Concorrente principal da SDRAM.
- Encapsulamento vertical – todos os pinos em um lado.
- Troca de dados com processador por **28 fios < 12 cm**.
- Barramento endereça até 320 chips RDRAM a 1,6GBps.
- Barramento especial: informações de endereço e controle usando protocolo
- Protocolo assíncrono orientado a bloco :
  - —Tempo de acesso de 480ns.
  - —Então, 1,6 GBps.

Estrutura da RAMBUS

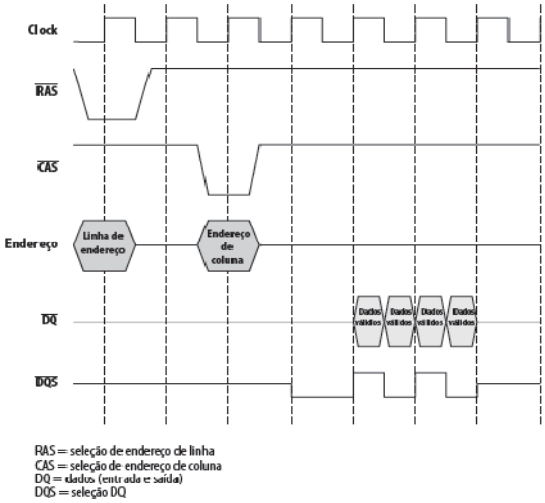


RAMBUS

DDR – SDRAM

- Desenvolvida pela *JEDEC Solid State Technology Association*, agência de padronização da EIA (*Electronics Industries Alliance*)
- SDRAM só pode enviar dados uma vez por ciclo de clock.
- **DDR-SDRAM** (*Double Data Rate*) envia dados duas vezes por ciclo de clock
  —Transição de subida e transição de descida.

Temporização de leitura da SDRAM DDR



RAS = seleção de endereço de linha
CAS = seleção de endereço de coluna
DQ = dados (entrada e saída)
DQS = seleção DQ

Melhorias em DDR – SDRAM

- **DDR2** aumenta a frequência operacional do chip
- DDR2 aumenta o buffer de pré-busca de 2 bits para 4 bits por chip
- Buffer de pré-busca é uma **cache no chip de RAM**
- Portanto, aumenta a taxa de transferência de dados
- **DDR3** (introduzida em 2007) aumenta tamanho do buffer de pré-busca para 8 bits

Módulo	Faixa de Operação (MHz)
DDR	200 a 600
DDR2	400 a 1300
DDR3	800 a 1600

Memórias DDR – SDRAM

Nome padrão	Clock dos chips	Ciclo de tempo	Clock real	Dados por segundo	Nome do módulo	Taxa de Transferência
DDR-200	100 MHz	10 ns	100 MHz	200 Milhões	PC-1600	1600 MB/s
DDR-266	133 MHz	7.5 ns	133 MHz	266 Milhões	PC-2100	2100 MB/s
DDR-300	150 MHz	6.67 ns	150 MHz	300 Milhões	PC-2400	2400 MB/s
DDR-333	166 MHz	6 ns	166 MHz	333 Milhões	PC-2700	2700 MB/s
DDR-400	200 MHz	5 ns	200 MHz	400 Milhões	PC-3200	3200 MB/s

Memórias DDR2 – SDRAM

Nome padrão	Clock dos chips	Ciclo de tempo	Clock real	Dados por segundos	Nome do módulo	Taxa de transferência
DDR2-400	100 MHz	10 ns	200 MHz	400 Milhões	PC2-3200	3200 MB/s
DDR2-533	133 MHz	7.5 ns	266 MHz	533 Milhões	PC2-4200 PC2-4300	4266 MB/s
DDR2-667	166 MHz	6 ns	333 MHz	667 Milhões	PC2-5300 PC2-5400	5333 MB/s
DDR2-800	200 MHz	5 ns	400 MHz	800 Milhões	PC2-6400	6400 MB/s
DDR2-1066	266 MHz	3.75 ns	533 MHz	1066 Milhões	PC2-8500 PC2-8600	8533 MB/s
DDR2-1300	325 MHz	3.1 ns	650 MHz	1300 Milhões	PC2-10400	10400 MB/s

Memórias DDR3 – SDRAM

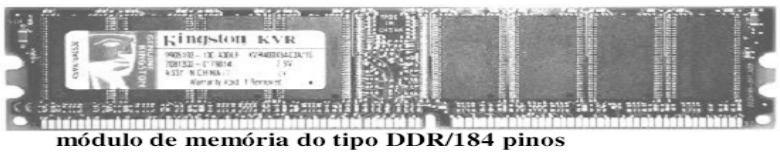
Nome padrão	Clock de memória	Tempo de ciclo	Velocidade de clock	Taxa de dados	Nome do módulo	Pico de taxa de transferência	Tempos
DDR3-1066	133 MHz	7.5 ns	533 MHz	1066 MT/s	PC3-8500	8533 MB/s	6-6-6 7-7-7 8-8-8
DDR3-1333	166 MHz	6 ns	667 MHz	1333 MT/s	PC3-10600	10667 MB/s	7-7-7 8-8-8 9-9-9 10-10-10
DDR3-1600	200 MHz	5 ns	800 MHz	1600 MT/s	PC3-12800	12800 MB/s	8-8-8 9-9-9 10-10-10 11-11-11
DDR3-2133	266⅔ MHz	$3\ ^{3}/_{4}$ ns	1066⅔ MHz	2133⅓ MT/s	PC3-17000	17066⅔ MB/s	11-11-11 12-12-12 13-13-13 14-14-14

- DDR3 consome 30% menos que DDR2
- MT/s – Mega Transferências/segundo

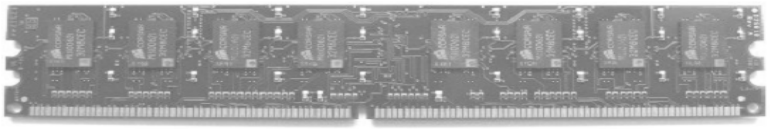
Memórias DDR3 – SDRAM

- **Tempos**: (CL-tRCD-tRP)
- CL – ciclos de clock entre o envio de um endereço de coluna para a memória e o início da recepção dos dados.
- tRCD – ciclos de clock entre a ativação da linha e leituras/escritas.
- tRP – ciclos de clock entre a carga de linha e sua ativação.
- **Latência do CAS**: latência do CAS de 9 em 1000 MHz (DDR3-2000) é de 9 ns, enquanto que a latência do CAS de 7 em 667 MHz (DDR3-1333) é de 10.5 ns. Quanto mais baixa, melhor.
  - (CAS Latency / Frequency (MHz)) × 1000 = X ns

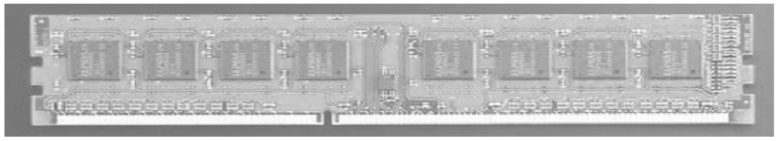
Comparação entre Módulos DDR3 – SDRAM



módulo de memória do tipo DDR/184 pinos

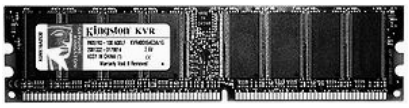


módulo de memória do tipo DDR2/240 pinos

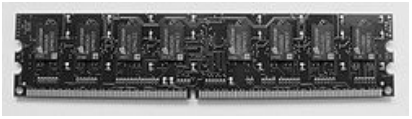


módulo de memória do tipo DDR3/240 pinos

Memórias DDR, DDR2 e DDR3



DDR 400 1G

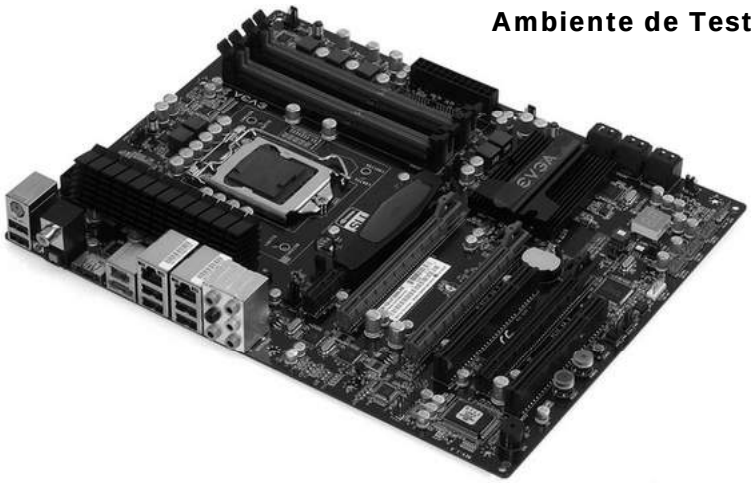


DDR2 400 512M

DDR3 1333 2G

Benchmark de Kits DDR3

Ambiente de Teste



Fonte: http://www.tomshardware.com/reviews/dual-channel-ram-ddr3-4gb,2618.html

Benchmark de Kits DDR3

Placa-mãe EVGA P55 SLI E655-1

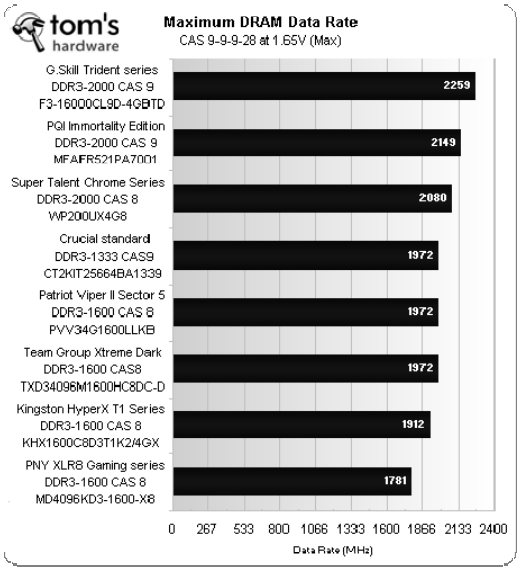
Benchmark de Kits DDR3

Ambiente de Teste

Processadores Core i7-870 and Core i7-860

Benchmark de Kits DDR3

1º Teste



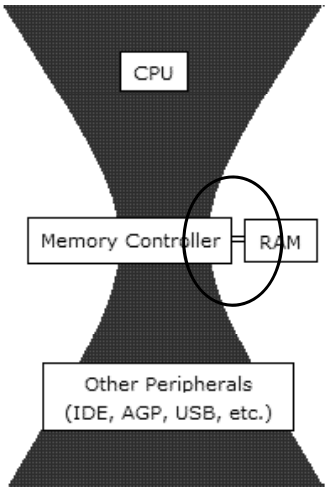
Taxas de dados em MHz

Benchmark de Kits DDR3

Conclusão – kit Trident 4GB DDR3-2000 da G.Skill

Tecnologias DDR e Dual-channel

Se o processador é rápido demais, ocorre um Gargalo de RAM !



DDR é diferente de Dual-channel

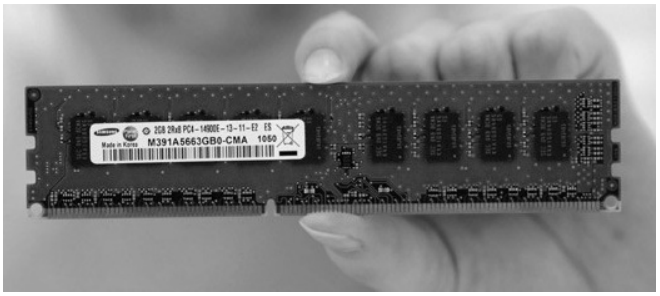
- **DDR** significa que **cada módulo de memória é acessado 2 vezes por clock**.
- **Dual-channel** é uma arquitetura que descreve uma tecnologia que teoricamente dobra a vazão de dados **da memória para o controlador de memória (dobra a largura de banda)**.
- Controladores de memória com Dual-channel habilitado usam **2 canais de dados simultâneos de 64 bits**.
- Usa a tecnologia de memória existente, mas muda controlador de memória (placa-mãe).
- As duas tecnologias são independentes uma da outra. Uma placa-mãe eventualmente pode usar ambas, ou seja, usar **memórias DDR na configuração dual-channel.**

Velocidade & Largura de Banda ?

- **Exemplo:**
- Uma estrada de mão dupla com **2 pistas**, cuja velocidade máxima permitida é 100 Km/hora
  - Velocidade = 100 Km/hora
  - Largura de Banda = 200 Km/hora
- Uma estrada de mão dupla com **4 pistas**, cuja velocidade máxima permitida é 100 Km/hora
  - Velocidade = 100 Km/hora
  - Largura de Banda = 400 Km/hora
- No caso de um canal simples significa uma única pista de ida e volta (leitura e escrita)
- No caso Dual-channel significa 2 pistas de ida e volta (leitura e escrita)

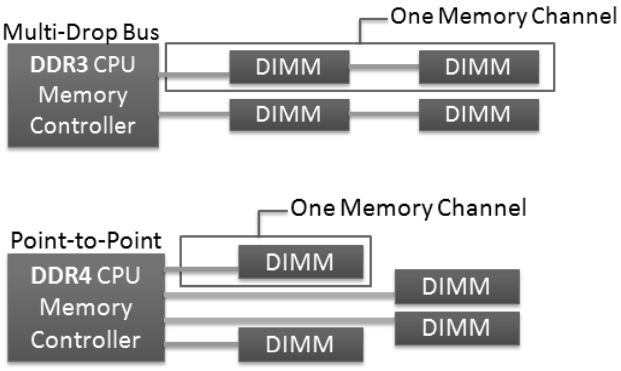
Ver: http://en.wikipedia.org/wiki/List_of_device_bandwidths

Memórias DDR4 (DDR Graphics RAM) ou GDDR4 (Graphics Double Data Rate, versão 4)



- Primeiro módulo **DDR4** lançado pela Sansung em janeiro de 2011 – 2 GB, depois Hynix em abril de 2011
- Consome 40% menos energia que **DDR3**
- Produção em massa esperada para o 2º semestre de 2012
- pré-busca de 16 bits/clock (DDR3 é 8 bits/clock)

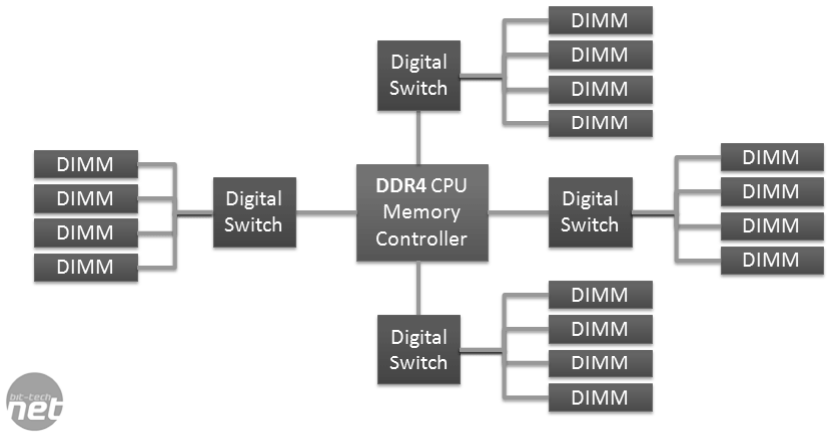
Memórias DDR4



Topologia Ponto-a-Ponto

Fonte: http://www.bit-tech.net/hardware/memory/2010/08/26/ddr4-what-we-can-expect/2

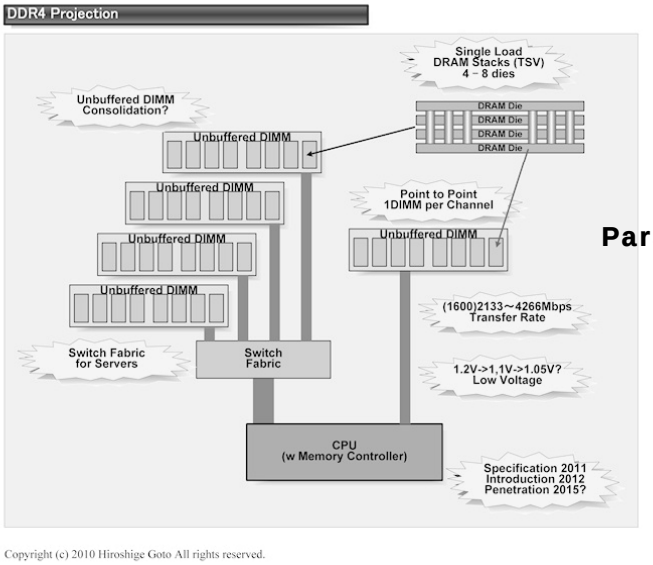
Memórias DDR4



Para Servidores

Fonte: http://www.bit-tech.net/hardware/memory/2010/08/26/ddr4-what-we-can-expect/2

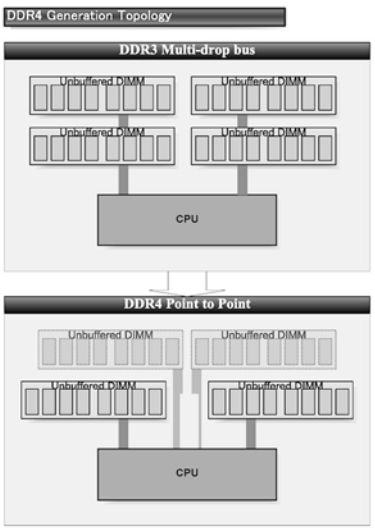
Memórias DDR4

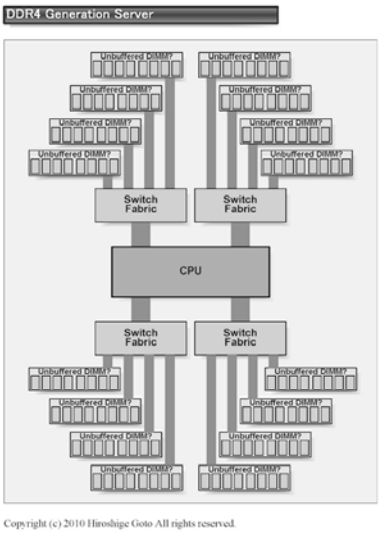


Para Desktops

Fonte: http://pc.watch.impress.co.jp/docs/column/kaigai/20100816_387444.html

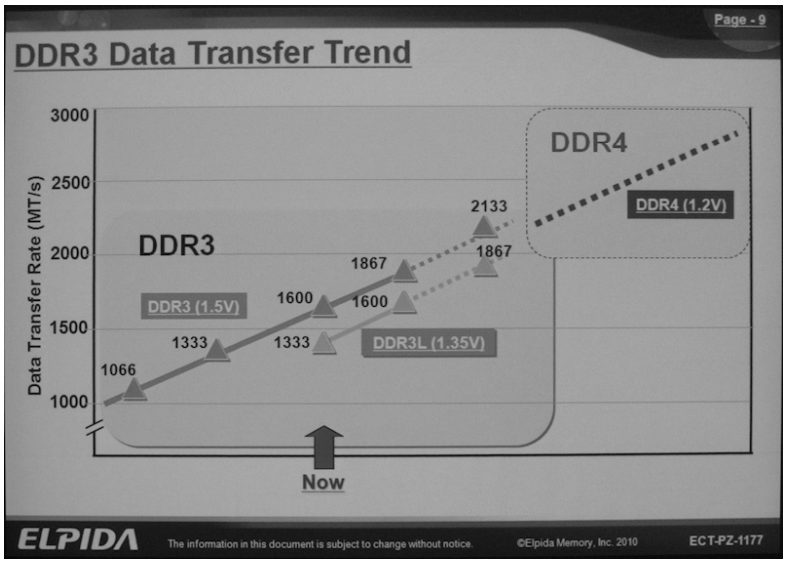
Memórias DDR4





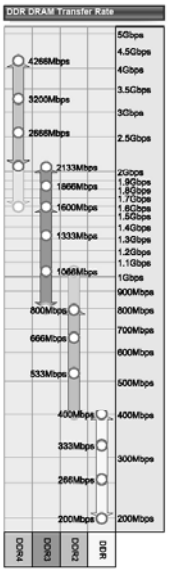
Topologia Ponto-a-Ponto

Memórias DDR4



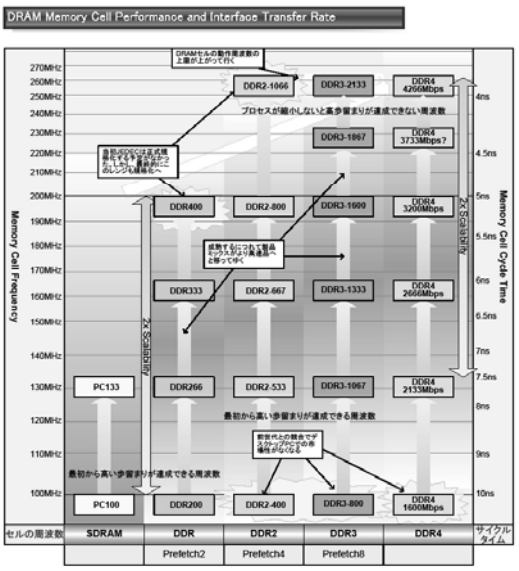
Memórias DDR4

Taxa de Transferência das memórias DDR



Memórias DDR4

Tempo de ciclo e Frequência de operação



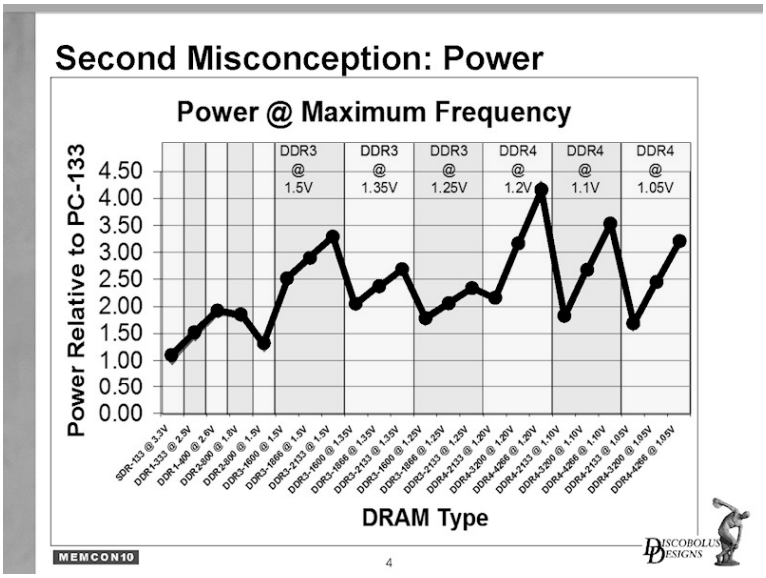
Memórias DDR4

DRAM Technology Transition

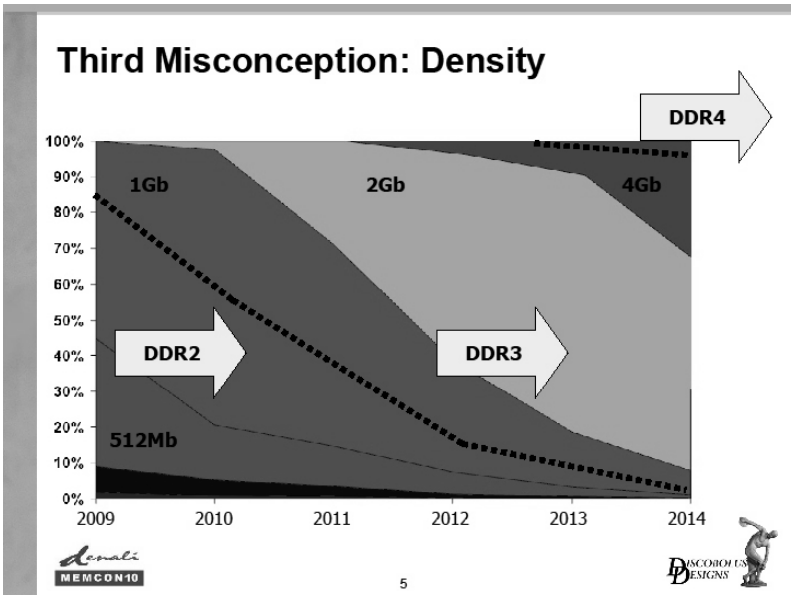
		2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
½ Pitch (ITRS)	プロセス (ITRS)	80nm	70nm	68nm	59nm	52nm	45nm	40nm	36nm	32nm	28nm	25nm
Density (ITRS)	容量世代 (ITRS)	1Gb	2Gb					4Gb			8Gb	
Cell Size	セルサイズ	8F2				6F2			4F2			
Volume Technology	プロセス 量産出荷	90nm	80nm	70nm	6xnm	5xnm	4xnm	4x-3xnm	3xnm			
Volume Density	PC向け 容量世代	256Mb	512Mb			1Gb			2Gb		4Gb	
Type	技術	DDR2						DDR3			DDR4	

Expectativa de Transição de Tecnologia

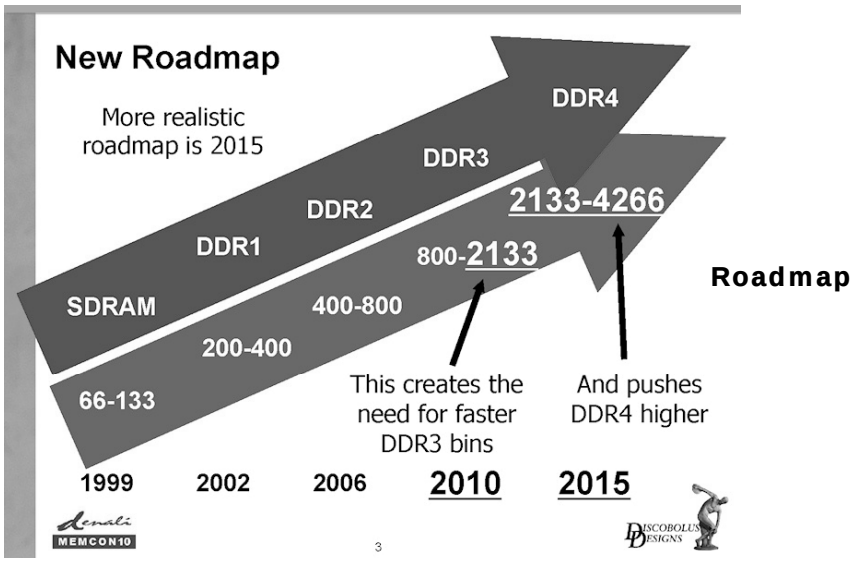
Memórias DDR4



Memórias DDR4



Memórias DDR4



Memórias DDR5 (DDR Graphics RAM) ou GDDR5 (Graphics Double Data Rate, versão 5)

- **GDDR5** é baseada na SDRAM DDR3.
- Também possui buffer de pré-busca de 8 bits, como a DDR3.
- Opera com 2 diferentes tipos de clock (CK e WCK).
- CK é usado como referência para endereços e entradas de comando. WCK é usado como referência para leitura e escrita de dados. WCK é o dobro da frequência de CK.
- Exemplo: 1 CK de 1.25 GHz, WCK de 2.5 GHz => taxa de dados de 5Gb/s, já que 2 dados serão em 1 pulso de WCK.
- A Hynix introduziu o primeiro chip de memória de 1Gb GDDR5, largura de banda de 20GB/s sobre um barramento de 32 bits.
- Chips de memória de 2 Gbit GDDR5 permitirá placas de vídeo com 2 GB ou mais de memória onboard com uma largura de banda de pico de 224 GB/s ou maior.

Cache DRAM (CDRAM)

- Desenvolvida pela Mitsubishi.
- Integra pequena cache SRAM (16 kb) no chip de DRAM genérico.
- Usada como cache verdadeira.
  - —linhas de 64 bits.
  - —Efetiva para acesso aleatório comum.
- Para admitir acesso serial de bloco de dados.
  - —Por exemplo, ao renovar tela de mapa de bits.
    - – CDRAM pode previamente buscar os dados da DRAM no buffer de SRAM.
    - – Acessos subsequentes unicamente à SRAM.