Hierarquia de memória

Melhorando o desempenho

Desempenho de uma CPU

- $CPU_{time\_sem\_mem}$ = #instruções x CPI x Clk_período
- $CPU_{time} = CPU_{time\_sem\_mem} + Memória_{time}$
- $CPU_{time} = (CPU_{ciclos\_sem\_mem} + Memória_{ciclos})\ Clk$
- $Memória_{ciclos} = Leitura_{ciclos} + Escrita_{ciclos}$
- $Leitura_{ciclos} = Leituras\ x\ Miss\_rate_{leitura} x\ Penalty_{leitura}$

Desempenho de uma CPU

- *Exemplo:*
  - *$Miss\_rate_{instr.}$=2%, $Miss\_rate_{dado.}$=4%, CPI=2, Penalty=40 ciclos*
  - *Taxa(Load,Store)= 36%*
  - *Qual a degradação de desempenho devido aos acessos à memória? CPI_{mem}?*

Desempenho de uma CPU

- $Miss\_instr_{ciclos} = I \times 2\% \times 40 = 0.80I$
- $Miss\_dados_{ciclos} = I \times 36\% \times 4\% \times 40 = 0.56I$
- $CPU_{time} = (CPU_{ciclos\_sem\_mem} + Memória_{ciclos})\ Clk$
- $Memória_{ciclos} = 0.80\ I + 0.56I = 1.36I$
- $CPI_{mem} = 2.0 + 1.36 = 3.36$
- $Desempenho = CPU_{time\_mem} / CPU_{time} =$
  $I \times CPI_{mem} \times clk / I \times CPI \times clk = 3.36/2.0 =$
  1.68 (mais rápida)

Processador mais rápido...

- *Diminuindo CPI*
  - $CPI_{novo} = 1.0$
  - $CPI_{mem} = 1.0 + 1.36 = 2.36$
  - *Desempenho = 2.36 / 1.0 = 2.36*
    - *memória causa maior impacto (58% tempo para acessar mem.)*
- *Duplicando o clock*
  - *Penalty = 80 ciclos*
  - $Miss_{ciclos}$*(por instr.)= 2% x 80 + 36% (4% x 80) = 2.75*
  - $CPI_{mem} = 2.0 + 2.75 = 4.75$
  - *Desempenho =* CPU_{time_slow} */* CPU_{time_fast} *=*
    $I \times CPI_{slow} \times clk / I \times CPI_{fast} \times clk/2 =$
    3.36 /(4.75 x 1/2) = 1.41 (em vez de 2)

Desempenho de uma cache

- *CPI(#clocks por instrução)*
  - *Cache Perfeita => 2.0*
  - *Cache (2% $miss_{instr}$, 4% $miss_{dado}$) => 3.36*
  - *Sem cache => 68.5*
- *Melhorando o processador*
  - *Diminuindo CPI*
    - *$desempenho_{sem\_mem}$ => 2.36 (1.68)*
  - *Duplicando clock*
    - *$desempenho_{fast}$ => 1.41 (em vez de 2.0)*

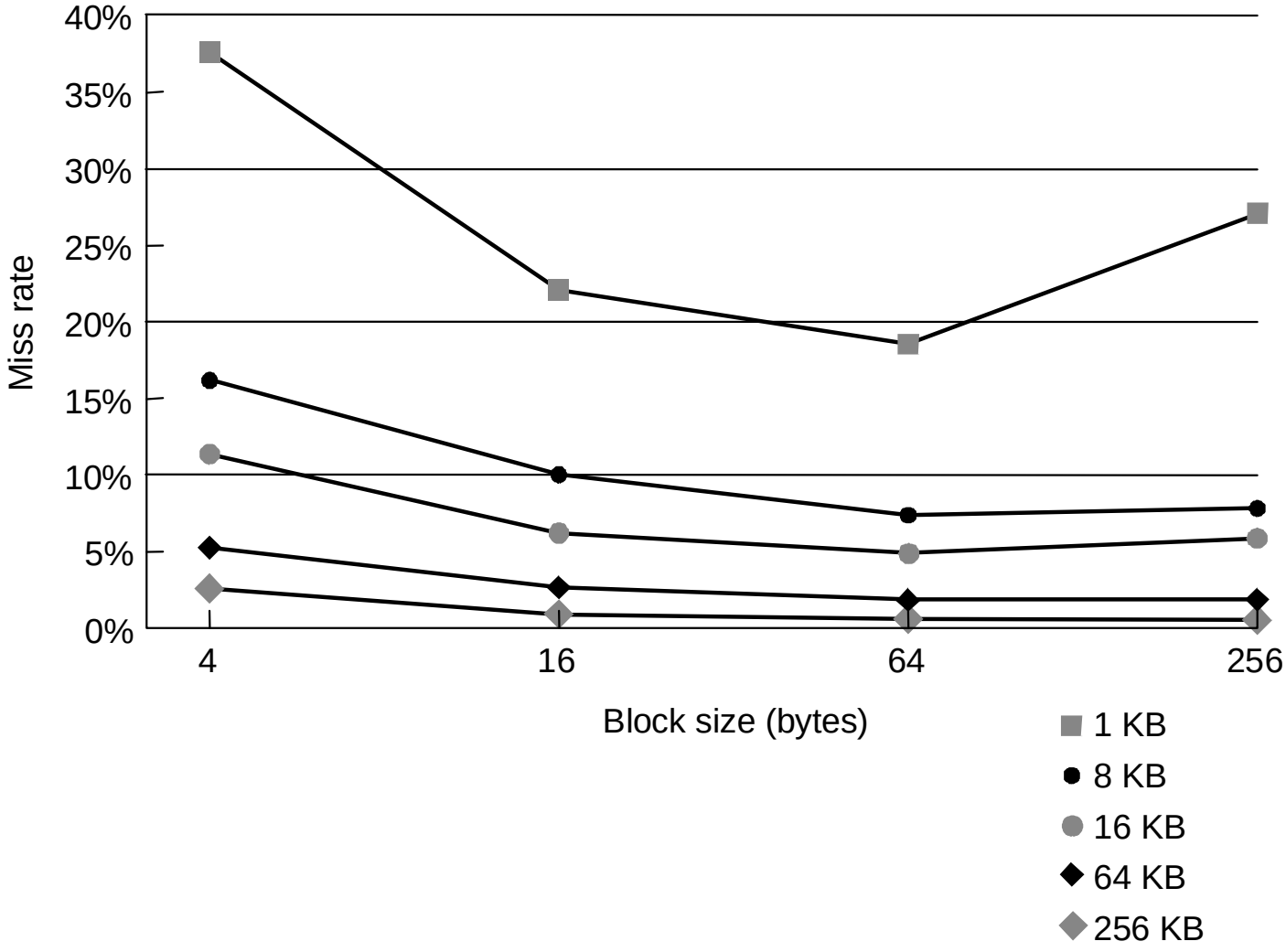
Melhorando desempenho da cache

$$Tempo\_acesso_{médio} = Hit\ time + Miss\ rate \times Miss\ penalty$$

- *Estratégias:*
  - *Redução de faltas*
  - *Redução da penalidade*
  - *Redução do tempo de acesso*

Reduzindo falta de cache

- *Aumentar tamanho do bloco*

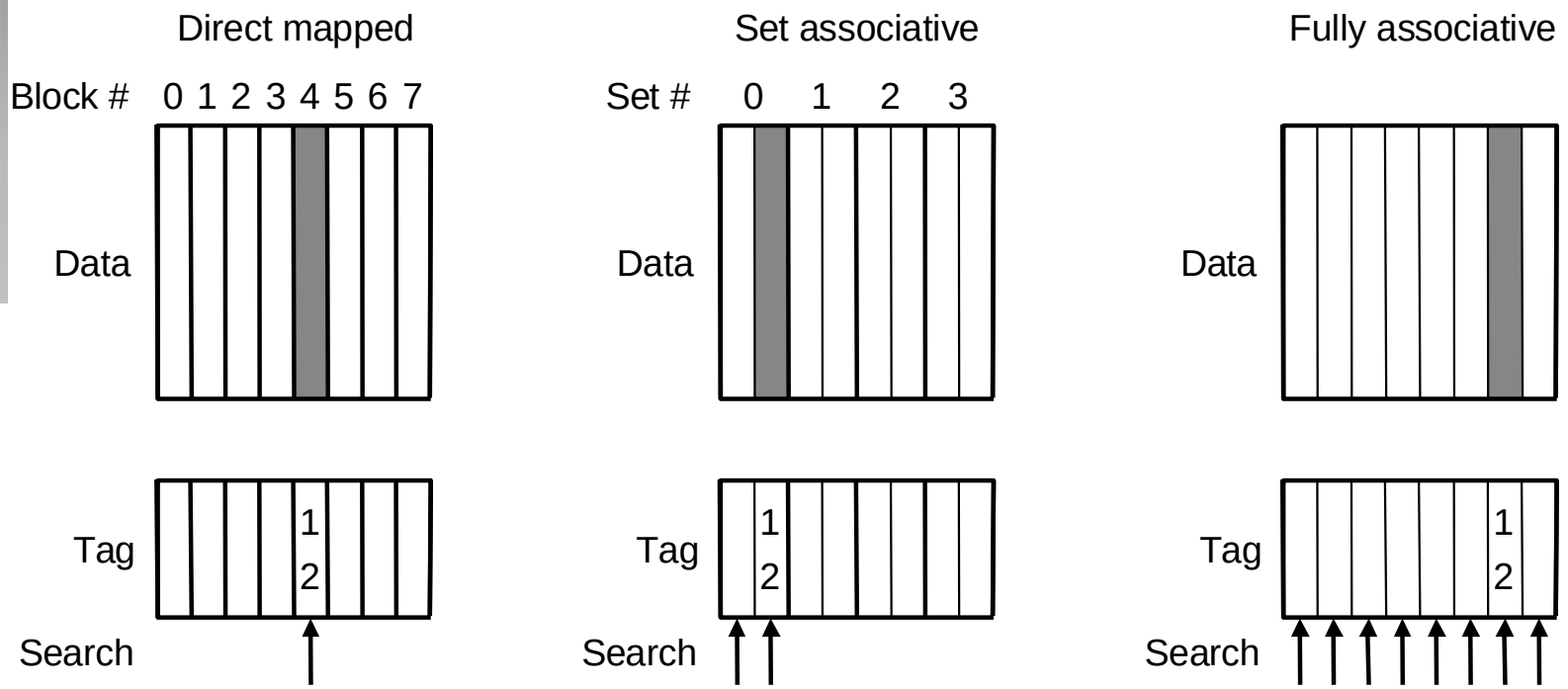


Reduzindo falta de cache

- *Aumento da associatividade*
  - *Regras práticas:*
    - *grau de associatividade 8 reduz "ao máximo" as faltas*
    - *regra 2:1: mapeamento direto de tamanho N tem a mesma taxa de falta que set associative 2 way de tamanho N/2*

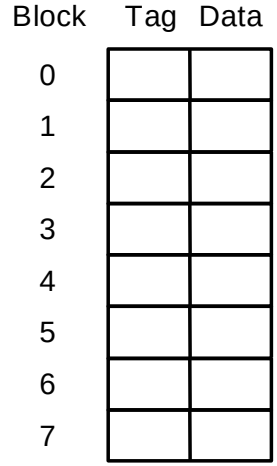
Tam. cache	one-way	two-way	four-way	eight-way
1	7.65	6.60	6.22	5.44
2	5.90	4.90	4.62	4.09
4	4.60	3.95	3.57	3.19
8	3.30	3.00	2.87	2.59
16	2.45	2.20	2.12	2.04
32	2.00	1.80	1.77	1.79
64	1.70	1.60	1.57	1.59
128	1.50	1.45	1.42	1.44

Aumento da Associatividade

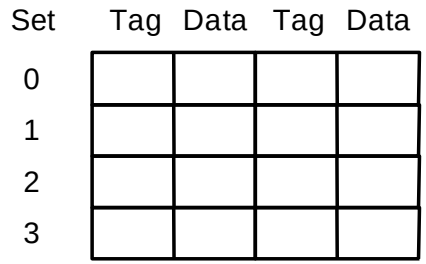


Aumento da Associatividade

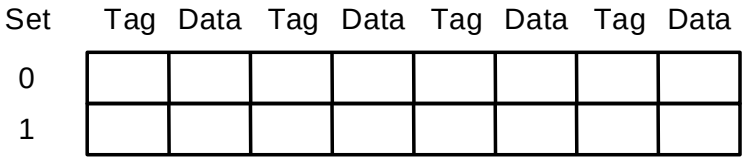
One-way set associative
(direct mapped)



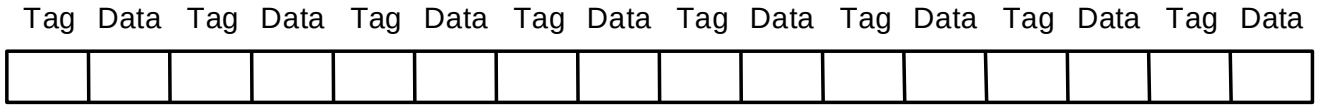
Two-way set associative



Four-way set associative



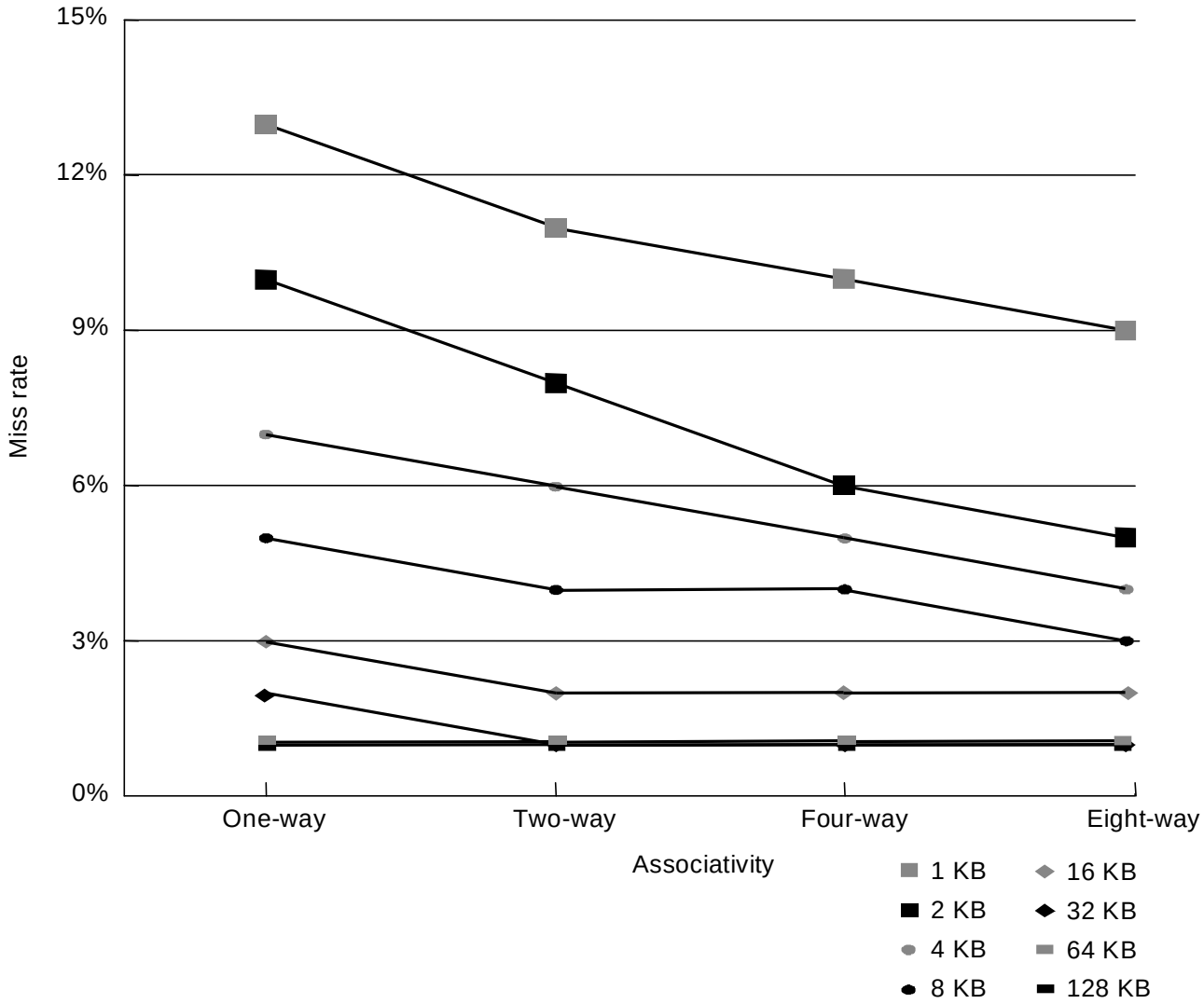
Eight-way set associative (fully associative)



Aumento da Associatividade

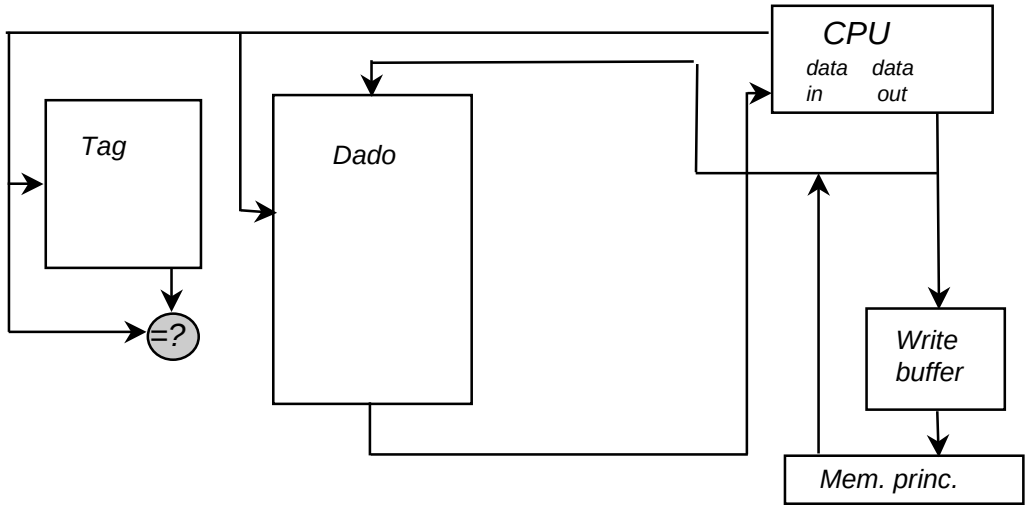
Programa	Associatidade	Miss instr.	Miss dado	Miss Total
Gcc	1	2.0%	1.7%	1.9%
Gcc	2	1.6%	1.4%	1.5%
Gcc	4	1.6%	1.4%	1.5%
Spice	1	0.3%	0.6%	0.4%
Spice	2	0.3%	0.6%	0.4%
Spice	4	0.3%	0.6%	0.4%

Aumento da Associatividade



Reduzindo penalidade de cache

- *Write Buffers*

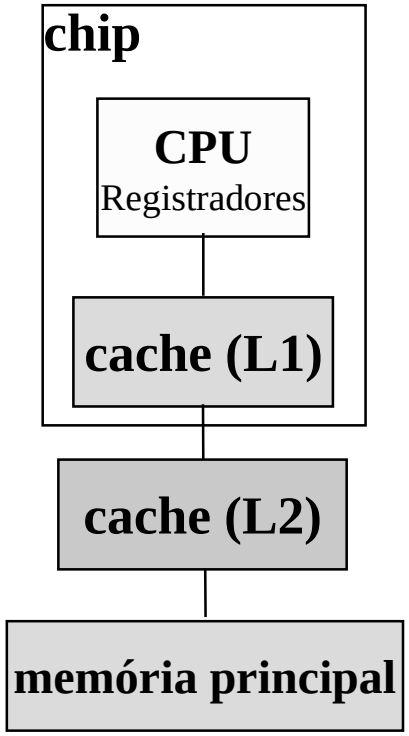


Reduzindo penalidade

- *Early Restart and Critical Word First:*
  - *Assim que palavra procurada foi carregada na cache esta é enviada para a CPU.*
  - *Requisita palavra procurada primeiro e a envia para a CPU assim que a mesma foi carregada.*
  - *Aplicável para grandes blocos*

Reduzindo a penalidade

- *Dois níveis de cache:*
  - *primeiro nível:*
    - *menor tempo de acesso*
    - *menor capacidade*
    - *maior custo*
  - *segundo nível:*
    - *maior capacidade*
    - *menor custo*
    - *maior tempo de acesso*



Reduzindo penalidade

- *Segundo nível de cache:*
  - *Desempenho:*
    - *Avrg.mem.acc.time=hit_{L1}+$miss_{L1}$x pen_{L1}*
    - *Pen_{L1}=hit_{L2}+$miss_{L2}$ x Pen_{L2}*
  - *De quanto melhora o desempenho da máquina pela inclusão do 2. nível?*

Reduzindo a penalidade

- *Exemplo: CPI_{base}=1.0, Clk=500MHz, $Time_{mem}$=200ns, $Miss\text{-}rate_{mem}$=5%.*
- *Segundo nível: $Time_{L2}$=20ns, $Miss\text{-}rate_{mem}$=2%*
- *Qual o desempenho da máquina com 2. nível de cache?*

Reduzindo a penalidade

- *Qual o desempenho da máquina com 2. nível de cache?*
  - $Penalty_{mem}$=200ns/2ns/clk = 100 ciclos
  - $CPI_{total}= CPI_{base}+Mem_{ciclos}/I$ = 1.0 + 5%x100=6.0
  - $Penalty_{L2}$=20/2=10 ciclos
  - CPI_{total}=1+L1-stalls+L2stalls = 1+((5%-2%)x10)+(2%x(10+100))= 1+0.3+2.2=3.5
  - Desempenho= 6.0/3.5 = 1.7

Reduzindo penalidade

- *Primeiro nível de cache:*
  - *Redução da penalidade*
    - *Redução do tempo de acesso*
    - *Uso de técnicas como early-restart e critical-word-first*

Reduzindo penalidade

- *Segundo nível de cache:*
  - *Redução da taxa de falta*
    - *cache do segundo nível maior que a do primeiro nível*
  - *E quanto a duplicação de dados nos dois níveis?*
    - *Os dados devem ser duplicados (consistência)*

Memória principal

- *Duplo papel:*
  - *satisfazer a demanda da cache*
  - *servir como interface para E/S*
- *Medidas de performance:*
  - *latência -> cache*
  - *Largura de banda -> E/S*

Tecnologias de memória

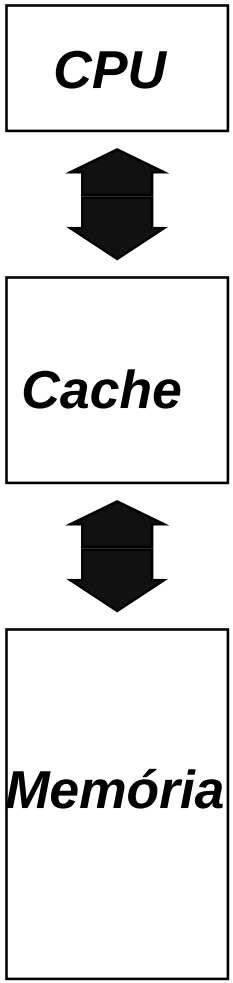
- *DRAM (Dynamic Random Acess Memory)*
  - *grande capacidade de integração*
    - *– multiplexação das linhas de endereço*
  - *perda de informação*
    - *– necessidade de refreshing*
- *SRAM (Static Random Acess Memory)*
  - *pequeno tempo de acesso*
  - *não existe necessidade de refreshing*

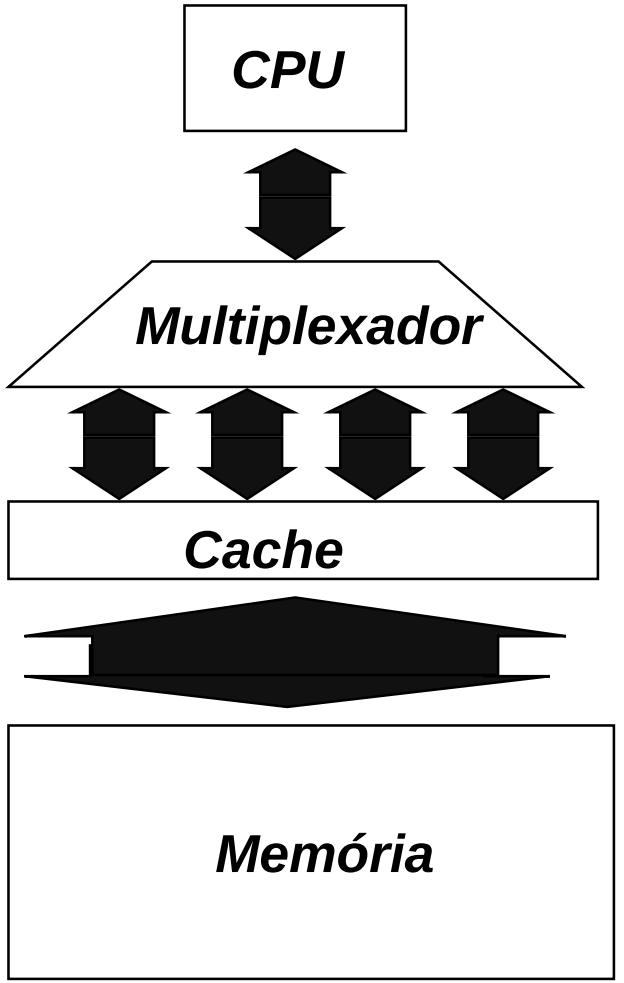
DRAMs vs. SRAMs

- *Capacidade de integração:*
  - *DRAMs -> 4 - 8 vezes SRAMs*
- *Tempo de acesso:*
  - *SRAMs -> 8 - 16 vezes mais rápidas que DRAMs*
- *Custo:*
  - *SRAMs -> 8 - 16 vezes mais caras que DRAMs*

⇨ *Necessidade de melhora no desempenho*

Memórias mais largas

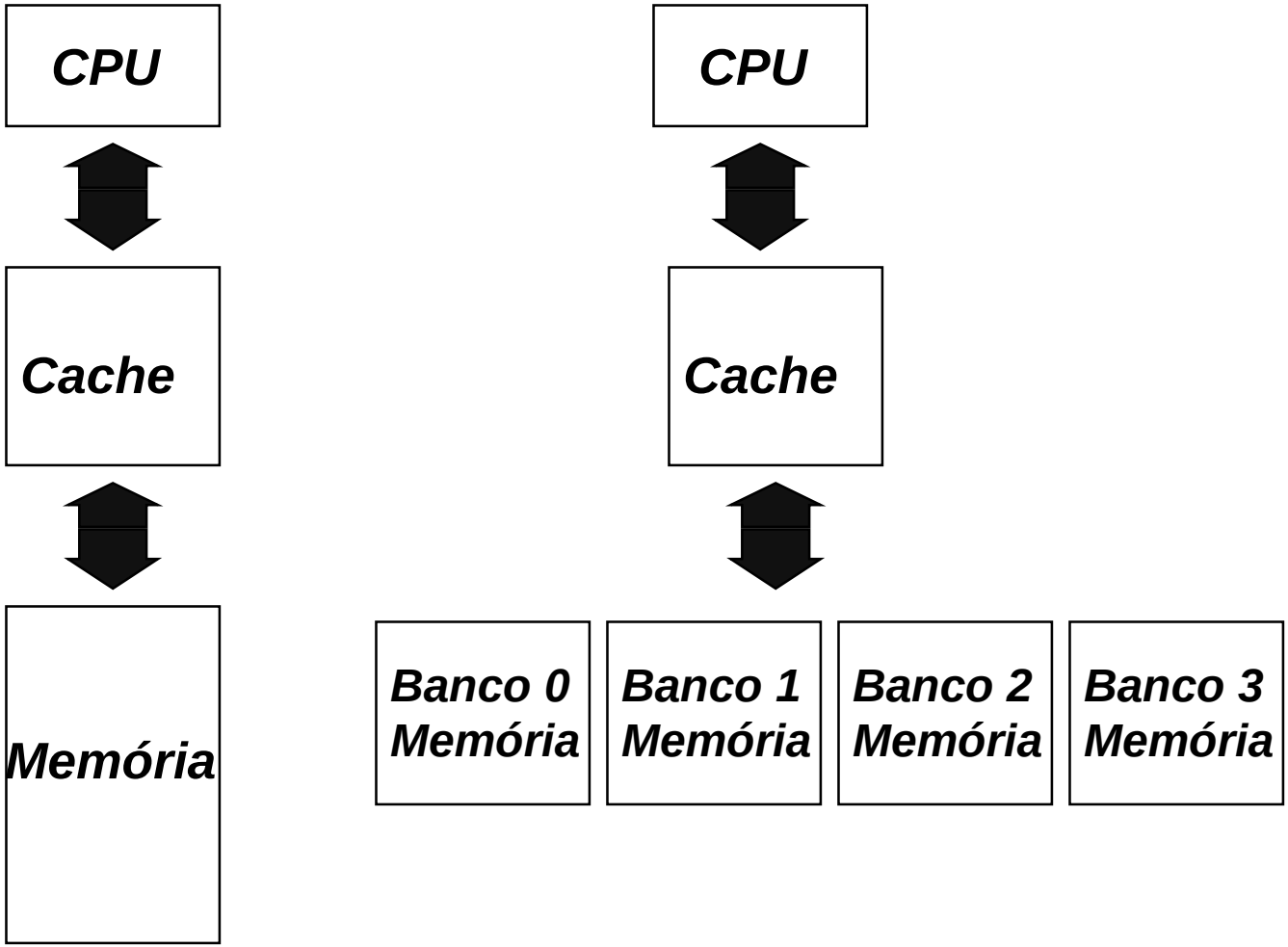




Memórias mais largas

- *Redução da penalidade de cache*
- *Necessidade de barramento e multiplexadores*
- *Expansão condicionada a largura*
- *Dificuldade em corrigir erros*
- *Ex: Alpha :*
  - *cache e mem. principal => 256 bits*

Memória "Interleaved"



Memória Interleaved

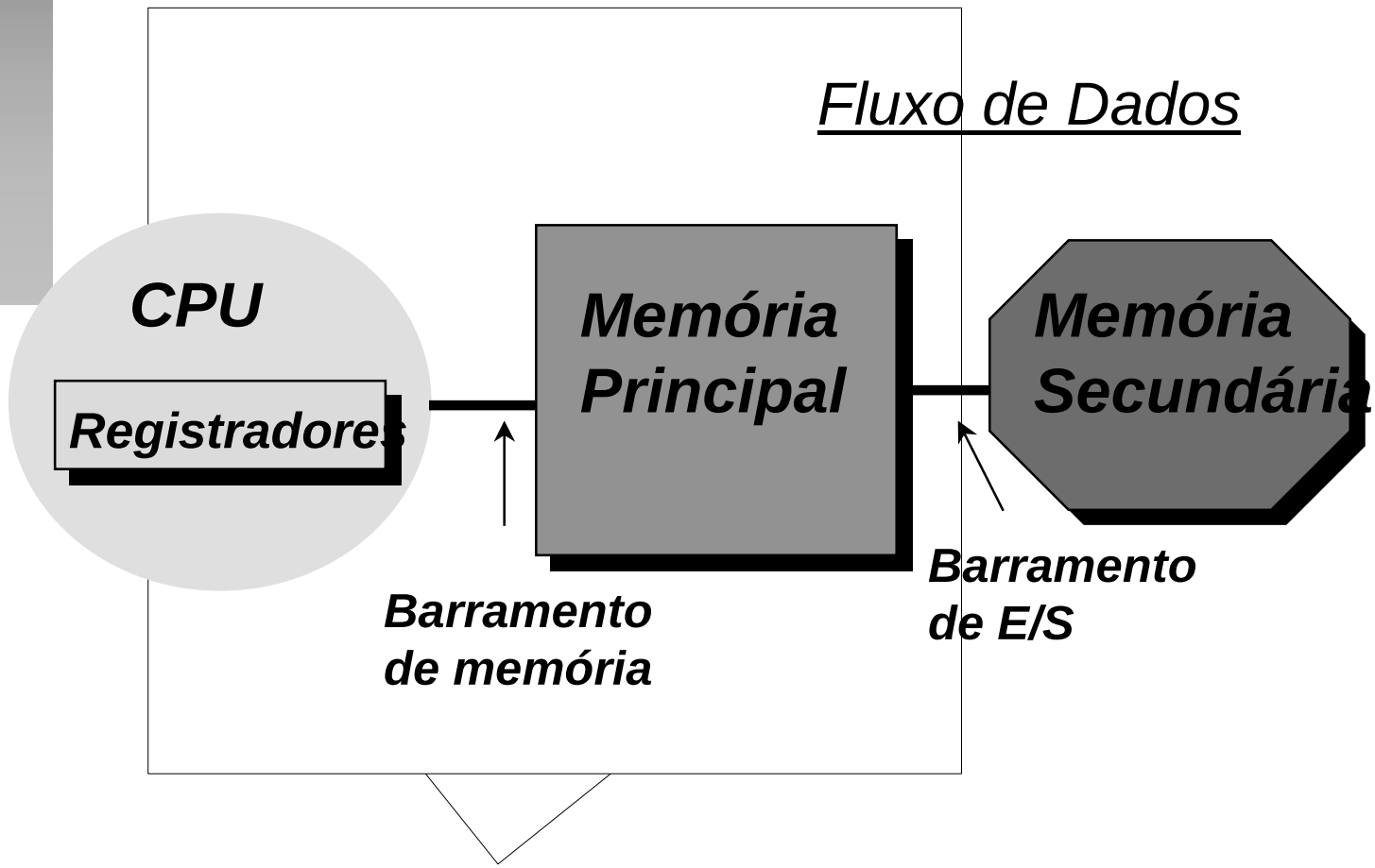
- *Bancos de memória para escrita/leitura de múltiplas palavras*
- *Reduz penalidade*
- *Necessita pouco hardware adicional*

Memória "larga" vs. Interleaved

- *CPI (# clocks por instrução)*
  - *32 bits, sem interleaving=3.54*
  - *32 bits, interleaving=2.86*

Tecnologias de Memória

Sistema Básico de Memória

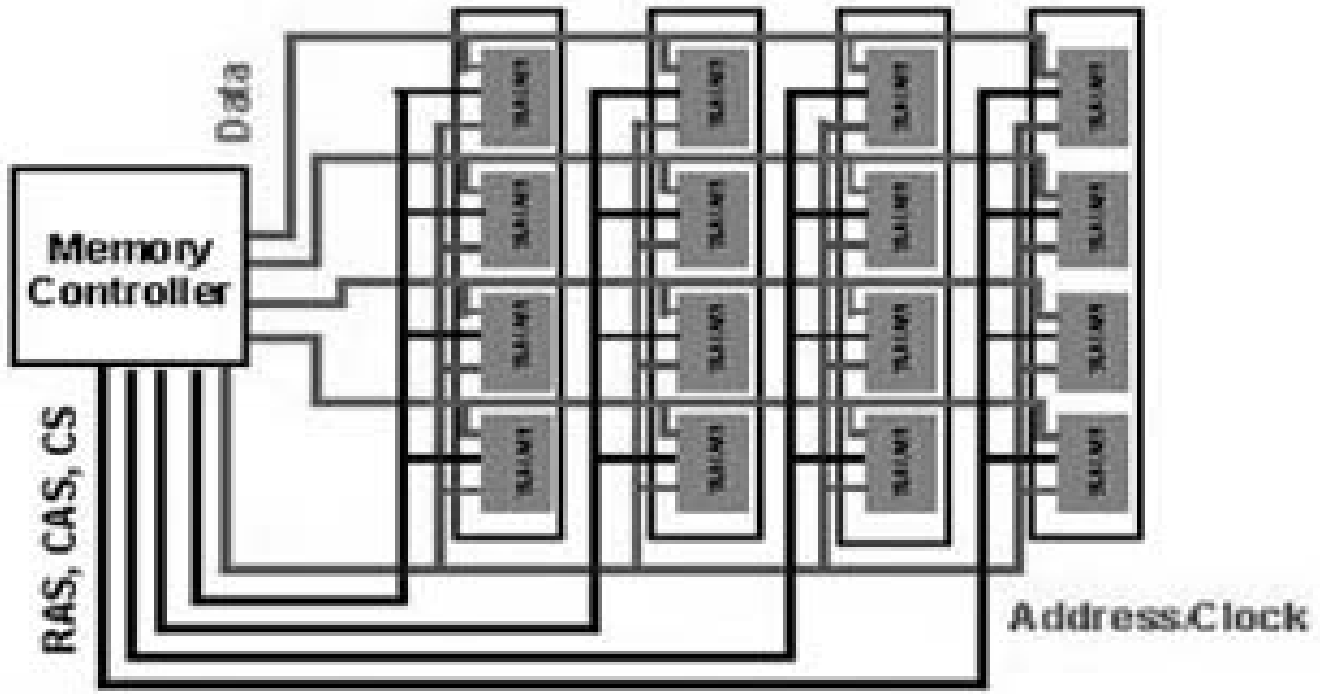


Características de Memória

- *Localização*
  - *CPU*
  - *Placa mãe (primária)*
  - *Externa (secundária)*
- *Capacidade*
  - *Tamanho e Núm. de palavras*
- *Unidade de Transf.*
  - *Palavra ou bloco*
- *Método de Acesso*
  - *Sequencial (ex. fita)*
  - *Acesso direto (ex. disco)*
  - *Acesso randômico*
  - *Acesso associativo*
- *Performance*
  - *Tempo de acesso*
  - *Ciclo*
  - *Taxa de transferência*
- *Implementação*
  - *Semicondutor*
  - *Superfície magnética*
  - *Superfície ótica*
- *Manutenção de dados*
  - *Volátil ou não*
  - *Apagável ou não*

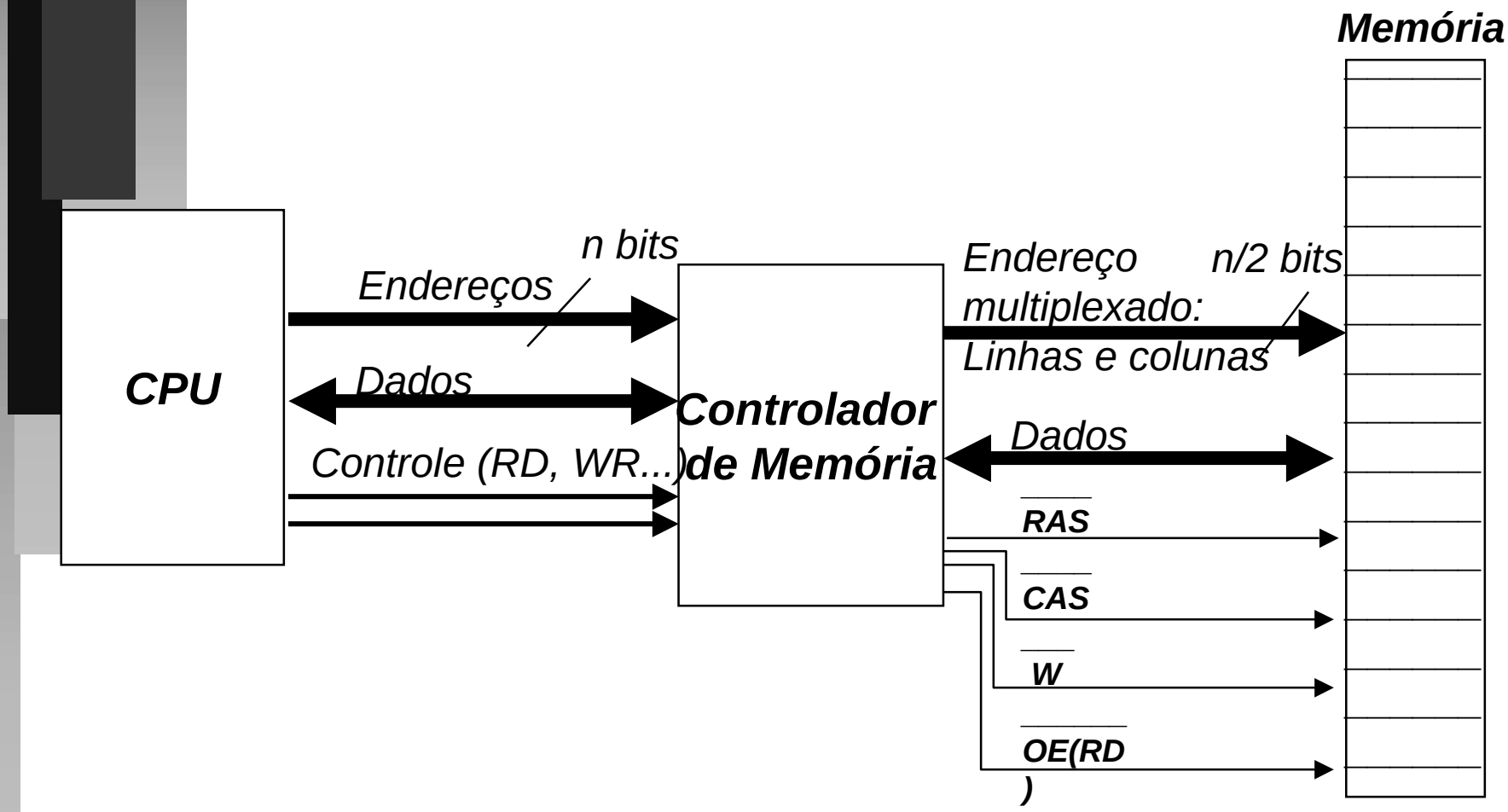
DIMM

SDRAM Architecture



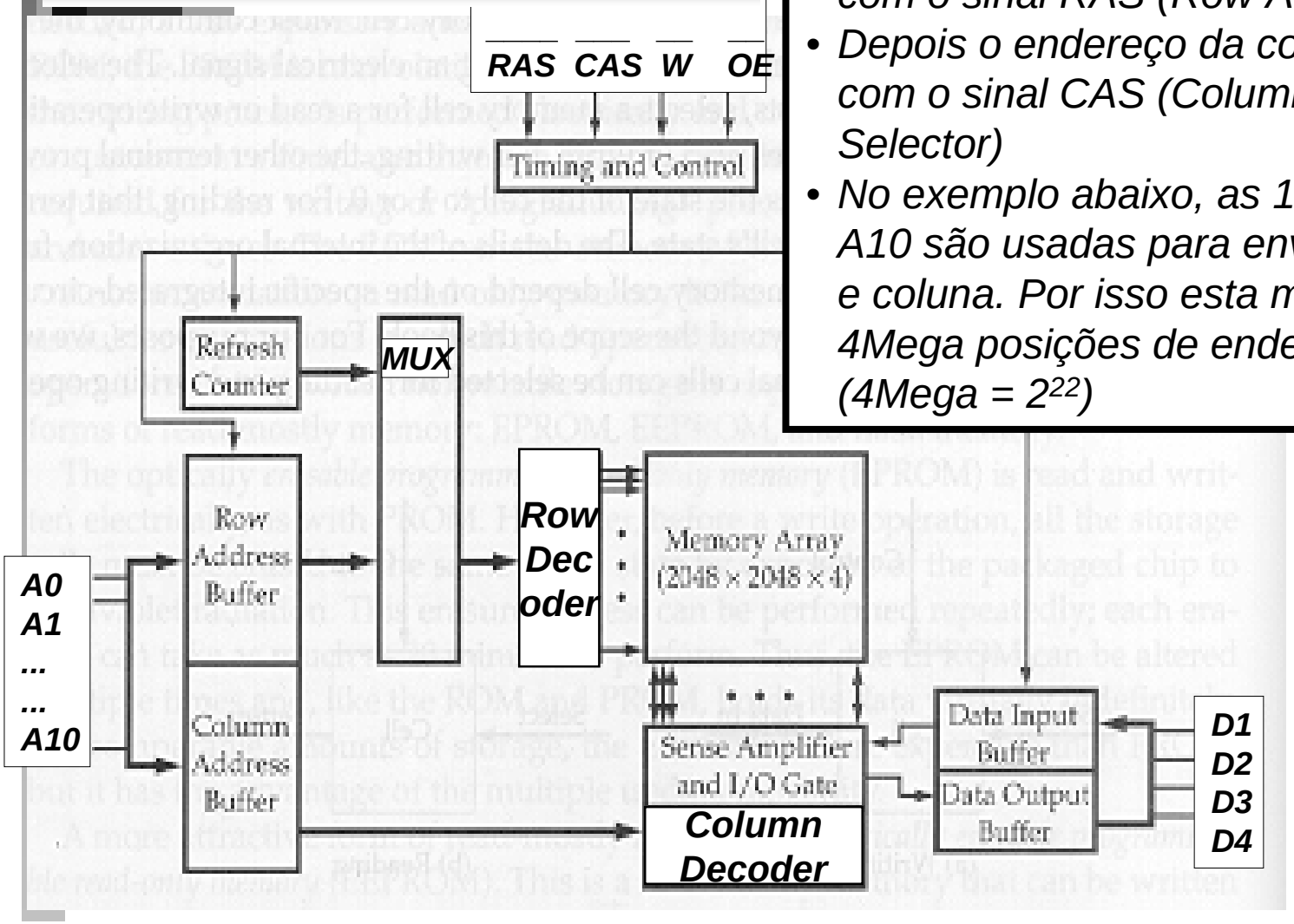
- Variable length wires, different routes
- 66-133Mhz

Acesso a memória principal (DRAM)



DRAM Organização Interna e Método de Acesso

- *O endereço das linhas e colunas são enviados separadamente (prim. linha, depois coluna).*
- *Para acessar uma posição (leitura ou escrita), o endereço da linha é posto junto com o sinal RAS (Row Addres Selector).*
- *Depois o endereço da coluna vai junto com o sinal CAS (Column Address Selector)*
- *No exemplo abaixo, as 11 linhas de A0 a A10 são usadas para enviar end. de linha e coluna. Por isso esta memória tem 4Mega posições de endereçamento (4Mega = 2^{22})*



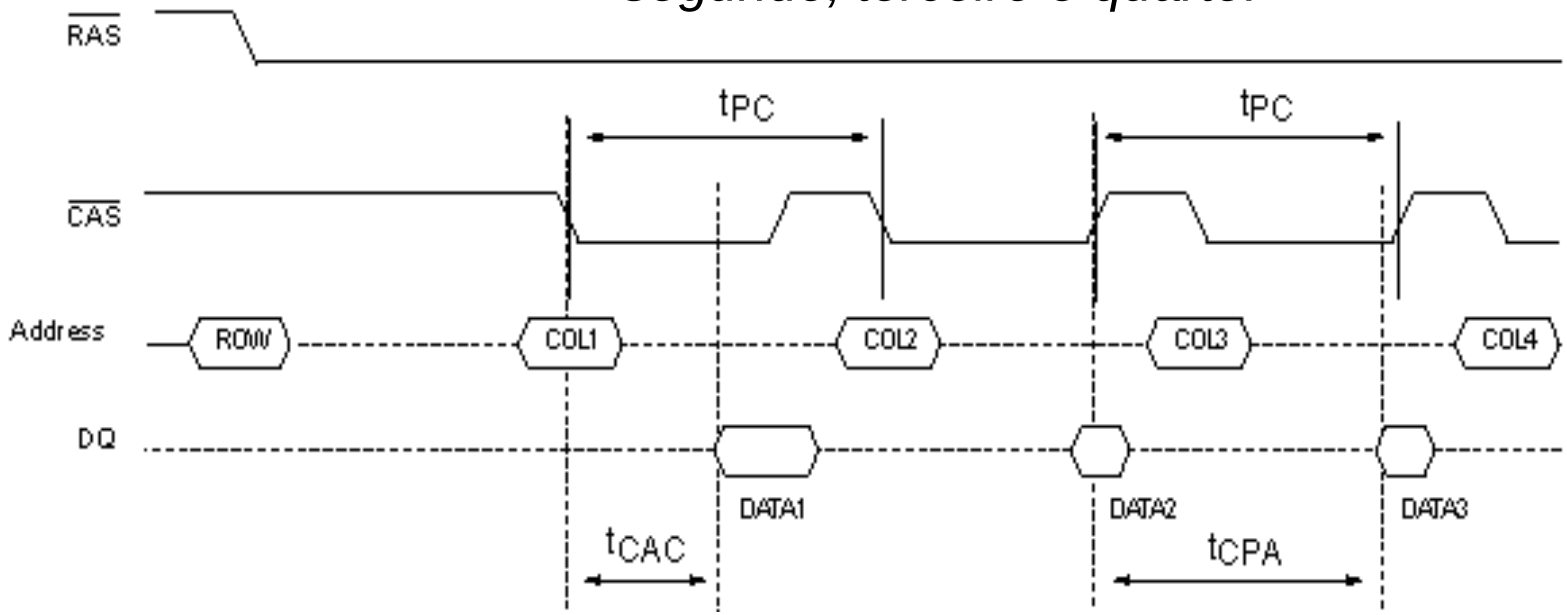
Modo de Acesso

Ex: FPM RAM

- *Ativa uma fila (RAS)*
- *Acessos sequenciais a colunas (vários pulsos de CAS)*

- *No Page Mode, o controlador de memória faz até 4 acessos em sequência à DRAM. É comum designar o núm. de pulsos de clock de cada acesso para cada tipo de memória.*
- *Ex. a FPM RAM tem acesso 5/3/3/3 (a 66MHz) em page mode, o que significa 5 pulsos de clock para obter o primeiro dado, e 3 para o segundo, terceiro e quarto.*

Fast Page Mode:



FPM DRAM - Fast Page Mode DRAM
EDO DRAM - Extended Data-Output DRAM

FPM DRAM

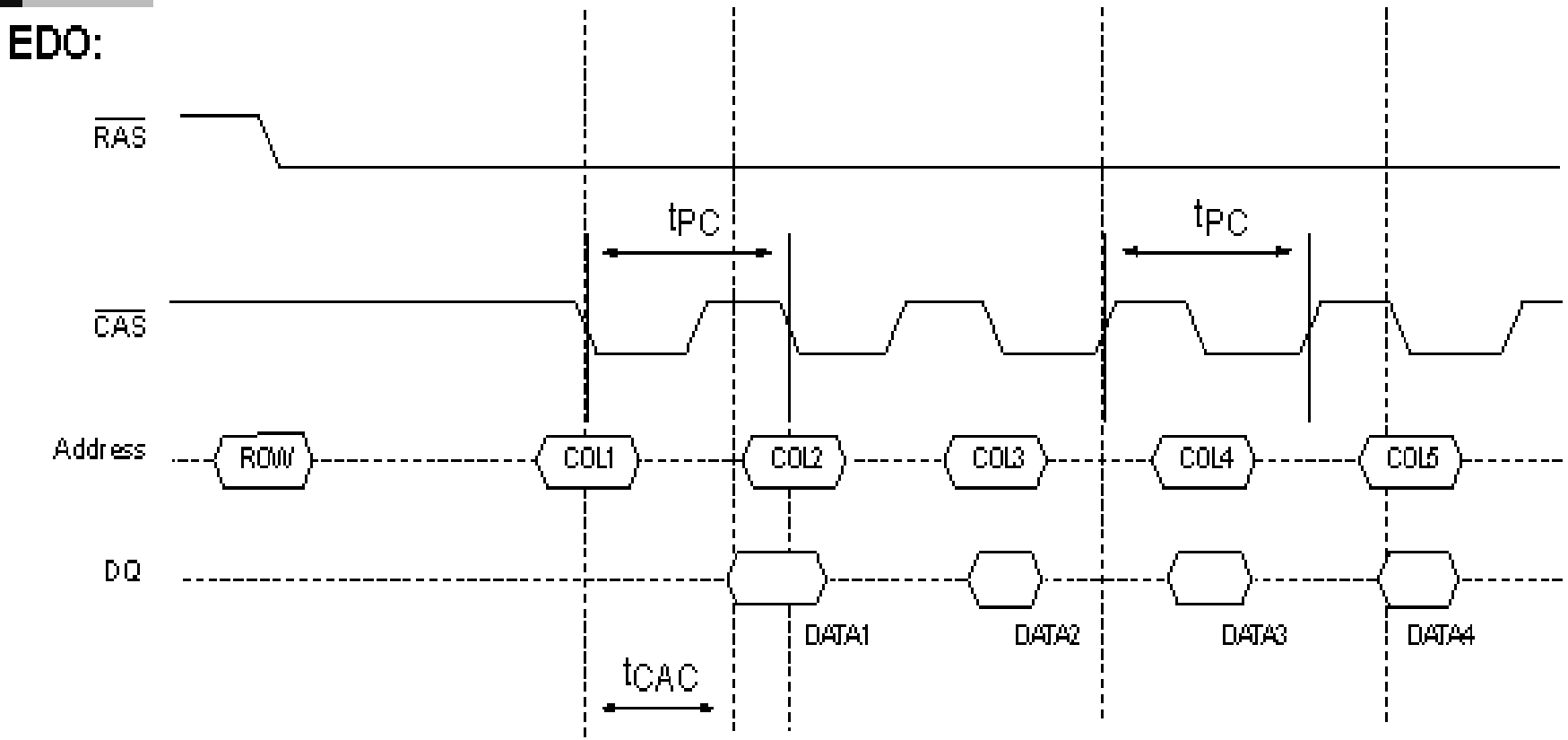
- *DRAM mais simples*
- *Tempo de acesso 70ns and 60ns.*
- *Acesso page mode = 5/3/3/3 a 66MHz*

EDO DRAM

- *Tempo de acesso 70ns, 60ns and 50ns*
  Para barramento de 66 MHz use 60 ns ou melhor.
- *Acesso page mode = 5/2/2/2 a 66MHz*

EDO DRAM
Método de Acesso

- ***No page mode, um latch na saída de dados permite o acesso simultâneo a novas posições de memória enquanto os dados estão sendo lidos na saída.***
- ***Isso permite a diminuição do tempo entre pulsos de CAS***

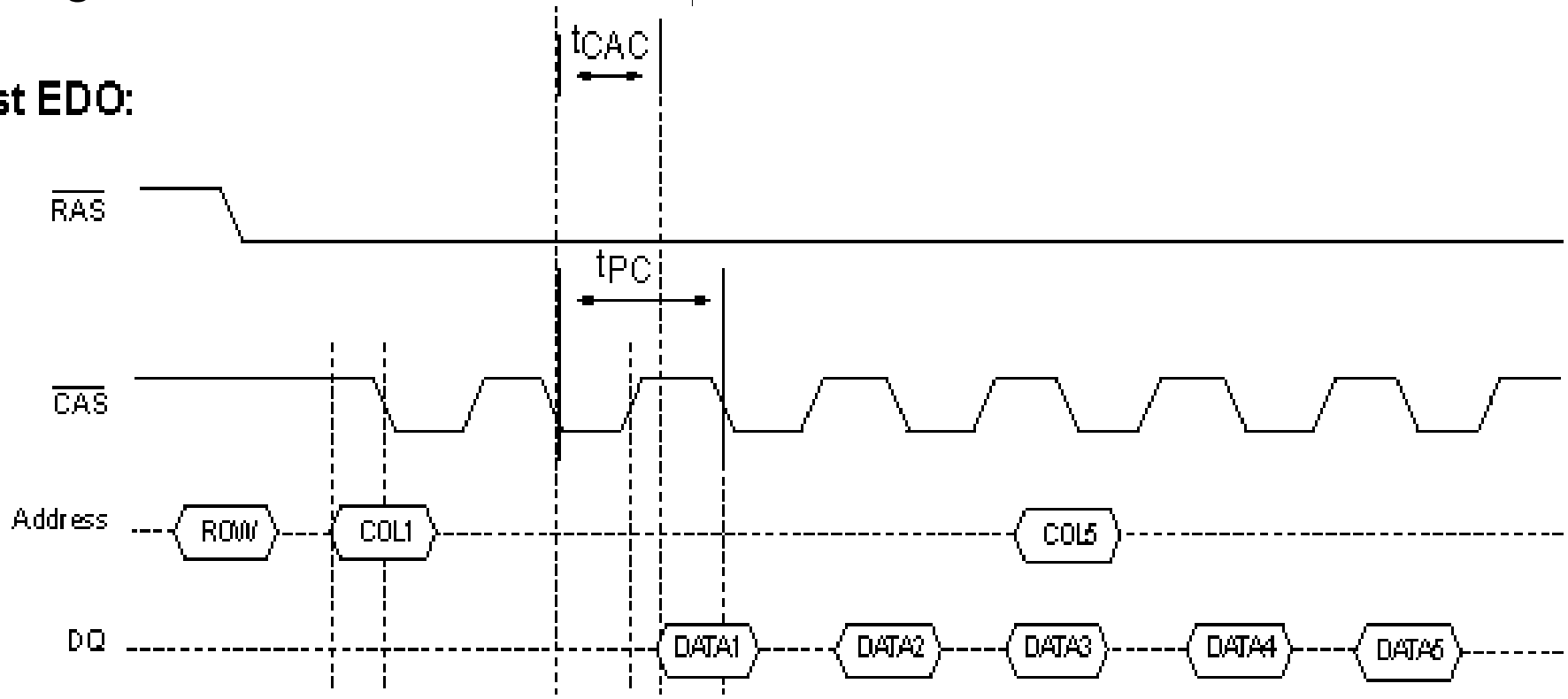


BEDO - Burst EDO DRAM

- **Page (burst) mode = 5/1/1/1 a 66MHz**
- **Tempo de acesso randômico é igual ao FPM ou EDO RAM**

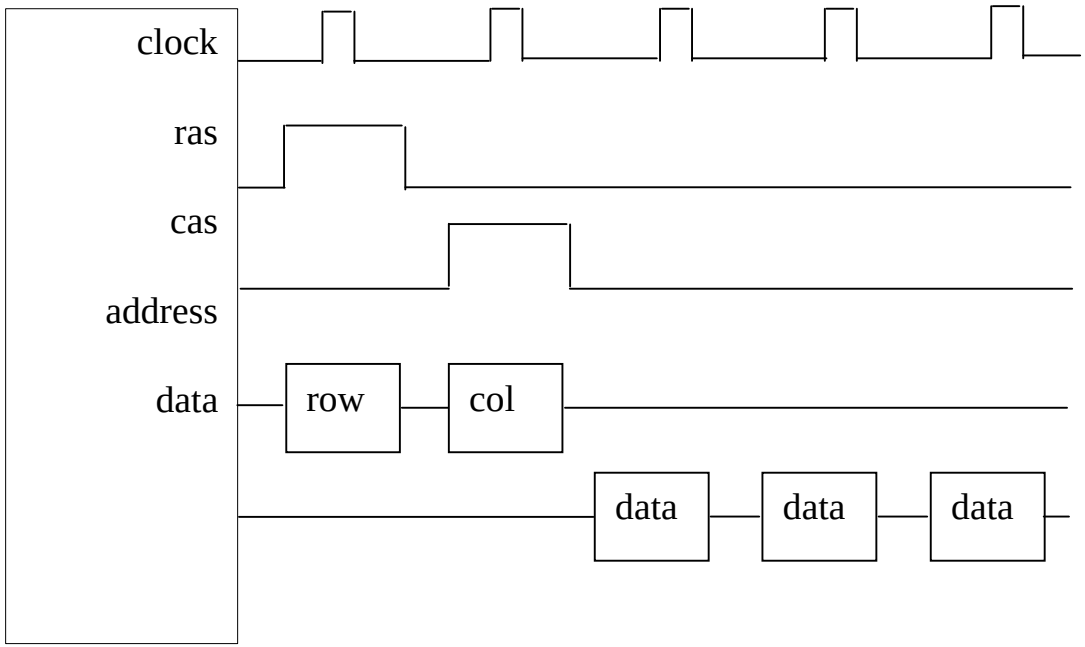
- **Possui registrador e gerador interno de ender. sequenciais**

Burst EDO:



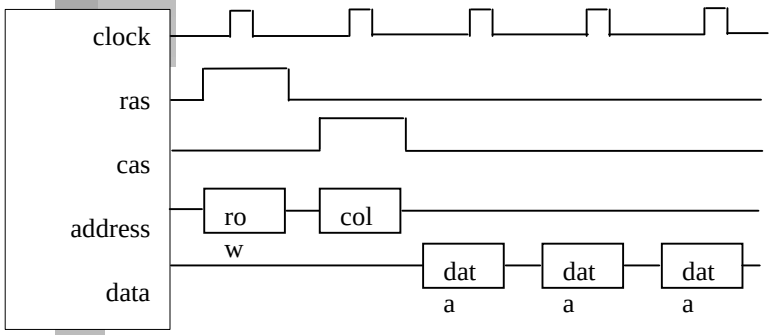
SDRAM - Synchronous DRAM

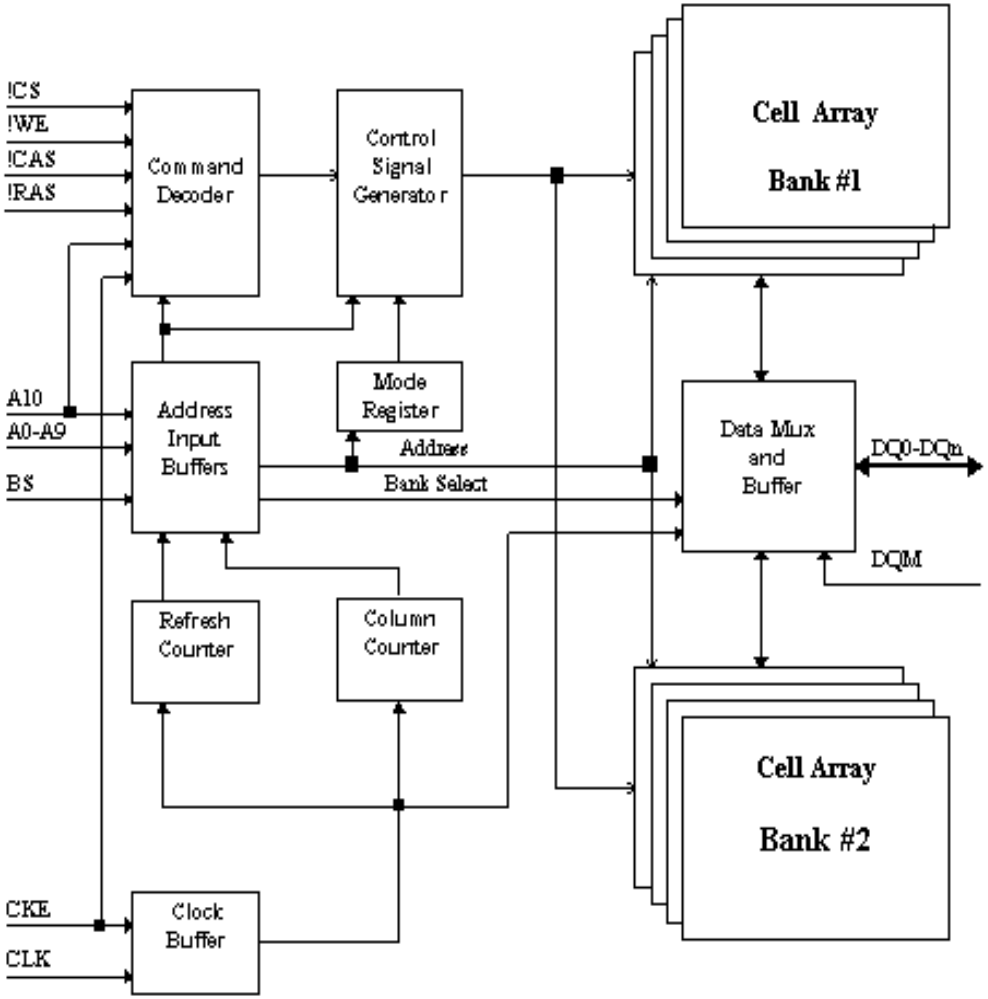
- Page (burst) mode = 5/1/1/1 a 100MHz
- Tempo de acesso randômico é igual à FPM ou EDO RAM.
- Trabalha na velocidade do bus, por isso o nome.



SDRAM - Synchronous DRAM

- Page (burst) mode = 5/1/1/1 a 100MHz
- Interleaved
- Uso de Serial Presence Detect para Plug-and-Play





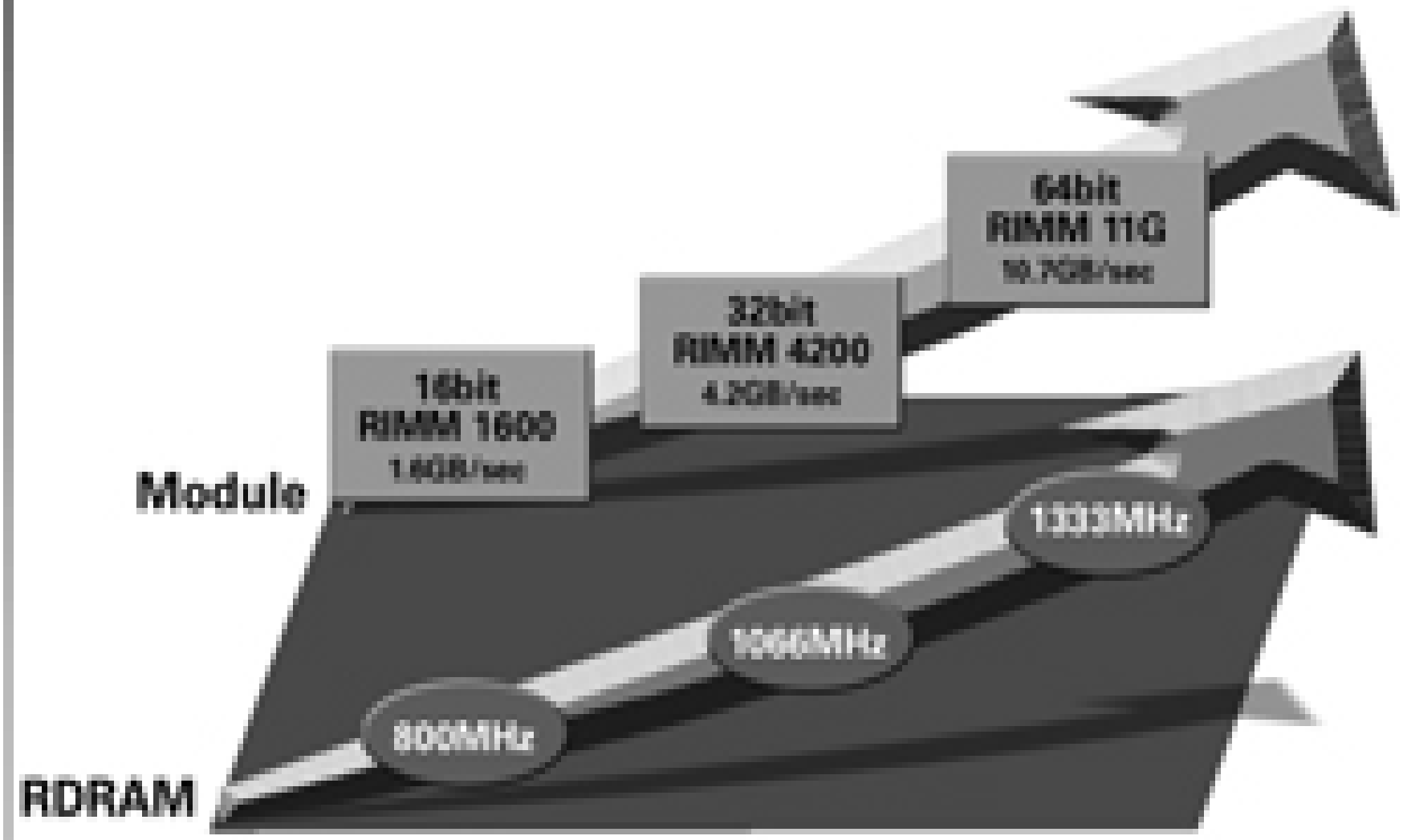
DDR SDRAM – Double Data Rate SDRAM

- Transfere dados na subida e descida do clock (compensa o uso de barramentos lentos)
- Uso de Serial Presence Detect para Plug-and-Play

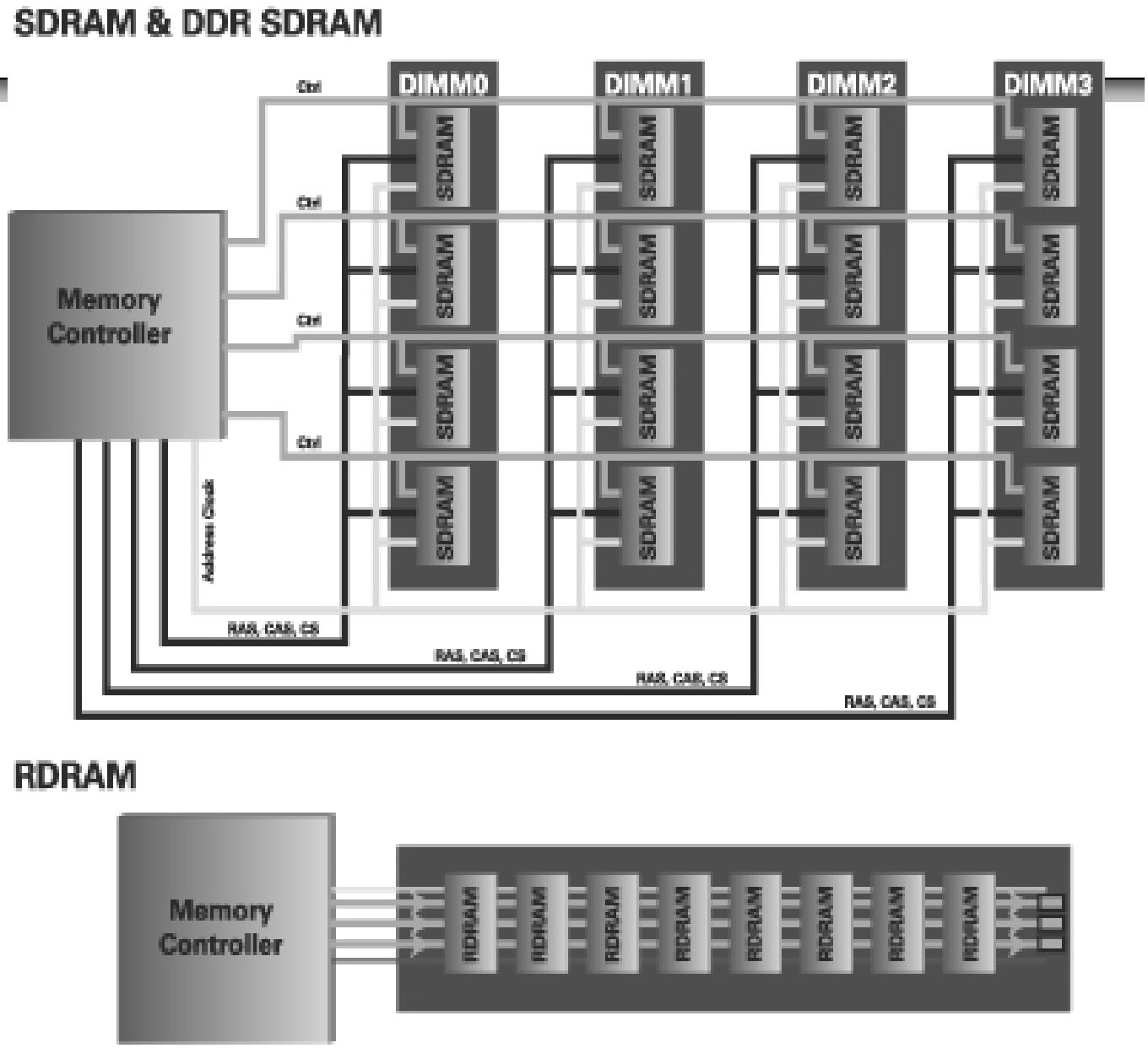
Direct Rambus DRAM ou RDRAM

- Padrão proprietário da Rambus Inc.
- Usa o Direct Rambus Channel – 400MHz, 16 bits (o barramento estreito permite estas taxas altas)
- Transferência de dados na subida e descida do clock
  ➔ 1,6 GByte/s
- Uso de Serial Presence Detect para Plug-and-Play

Direct Rambus DRAM ou RDRAM



Direct Rambus DRAM ou RDRAM



Synchronous Link DRAM - SLDRAM

- Desenvolvido pelo SLDRAM Consortium (20 empresas)
- Uso de tecnologia mais comum: bus de 64-bit a 200 MHz
- Transferência de dados na subida e descida do clock
  ➔ 3,2 GByte/s

Memórias para Controlador de Vídeo

- *VRAM - Vídeo RAM WRAM - Windows RAM*
- *Têm duas portas de acesso: uma para o controlador de vídeo e outra para a CPU*
- *WRAM é 50% mais rápida e 20% mais barata*
- *MDRAM - Multibank DRAM*
- *Vários bancos de DRAMs (cada um 32 KB) com I/O próprio ligados num bus interno.*
- *Dados podem ser acessados em vários bancos simultaneamente.*
- *Chips de qualquer tamanho podem ser fabricados. Ex: 2.5MB, usados em cont. de vídeo com resolução de 1,024x768 por 24bit*

Avanços em implementação de caches

- *Sincronizada ao clock do sistema*
  - *Os sinais são sincronizados ao clock.*
  - *Simplifica o projeto da memória*
- *Burst*
  - *Incorpora controle interno que permite acesso rápido a posições subsequentes (ex. reg. e gerador de endereços)*
- *Pipelining*
  - *Usa registradores na entrada e/ou saída o que permite fornecer dados e acessar novos endereços em paralelo.*

Async SRAM

- *A mais antiga*
- *É mais rápida que a DRAM apenas por ser estática.*
- *20, 15 or 12 ns.*
- *Não é suficientemente rápida para permitir acesso síncrono.*

SB SRAM - Syncronous Burst SRAM

- 8.5ns to 12ns
- Acesso em page mode
  - 2/1/1/1 em 66 MHz
  - 3/2/2/2 em > 66 MHz
- Para velocidadades de barramento até 66 MHz, SB SRAM apresenta a melhor performance.

PB SRAM - Pipeline Burst SRAM (cache)

- Usa registradores na entrada ou saída.
- Gasta um clock a mais para carregar o registrador, mas depois permite acesso simultâneo a novas posições de memória enquanto os dados estão sendo lidos na saída.
- Acesso page mode 3/1/1/1.
- Mais lenta que SB SRAM em bus $\leq$ 66MHz. Melhor se > 66MHz.
- 4.5ns to 8ns.

Velocidade de Acesso a Memórias Cache

A melhor memória para cada velocidade de barramento

Vel. Bus (MHz)	***33***	***50***	***60***	***66***	***75***	***83***	***100***	***125***
Async SRAM	***2/1/1/1***	***3/2/2/2***	***3/2/2/2***	***3/2/2/2***	***3/2/2/2***	***3/2/2/2***	***3/2/2/2***	***3/2/2/2***
Sync SRAM	***2/1/1/1***	***2/1/1/1***	***2/1/1/1***	***2/1/1/1***	***3/2/2/2***	***3/2/2/2***	***3/2/2/2***	***3/2/2/2***
PB SRAM	***3/1/1/1***	***3/1/1/1***	***3/1/1/1***	***3/1/1/1***	***3/1/1/1***	***3/1/1/1***	***3/1/1/1***	***3/1/1/1***

Comparando as tecnologias...

Memory Technology	Typical System Bus Speeds	Ideal Timing	Usual DRAM Speed (ns)
Conventional	4.77-40	5-5-5-5 or worse	80-150
FPM	16-66	5-3-3-3	60-80
EDO	33-75	5-2-2-2	50-60
BEDO	**60-100**	**5-1-1-1**	!?
SDRAM	**60-100+**	**5-1-1-1**	6-12