Sistemas de Memória

Conceitos básicos

João Canas Ferreira

Outubro de 2017

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Tópicos

1. Memórias
   - Aspetos gerais
   - Memórias Estáticas
   - Memórias Dinâmicas

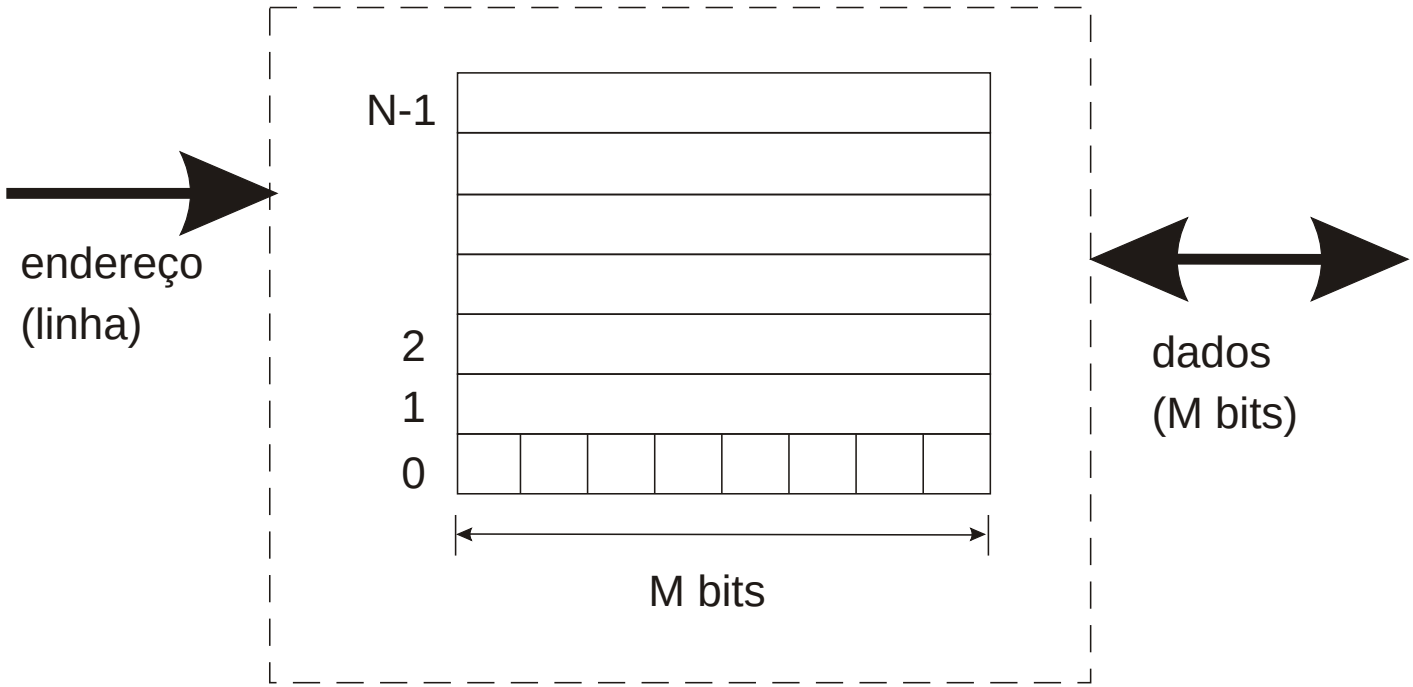
2. Descodificação de endereços
   - Organização geral
   - Descodificação total
   - Descodificação parcial

Contém figuras de "Computer Organization and Design", D. Patterson & J. Hennessey, 3ª. ed., MKP

Taxonomia

- Registos e bancos de registos permitem guardar pequenas quantidades de dados. Para maiores quantidades, usam-se **memórias de acesso direto**.
- RAM = *random access memory* (memória de acesso direto): permitem leitura e escrita em qualquer posição.
- ROM = *read-only memory*: permitem apenas leitura.
- A maior parte das memórias RAM perde os dados quando é desligada a alimentação (memória volátil). Exceções:
  - (E)EPROM: (Electrically) erasable programmable ROM
  - memórias FLASH
- Dois tipos de memórias RAM voláteis:
  - SRAM: memória estática (cada célula de memória é um anel realimentado);
  - DRAM: memória dinâmica (cada célula deve ser atualizada periodicamente).

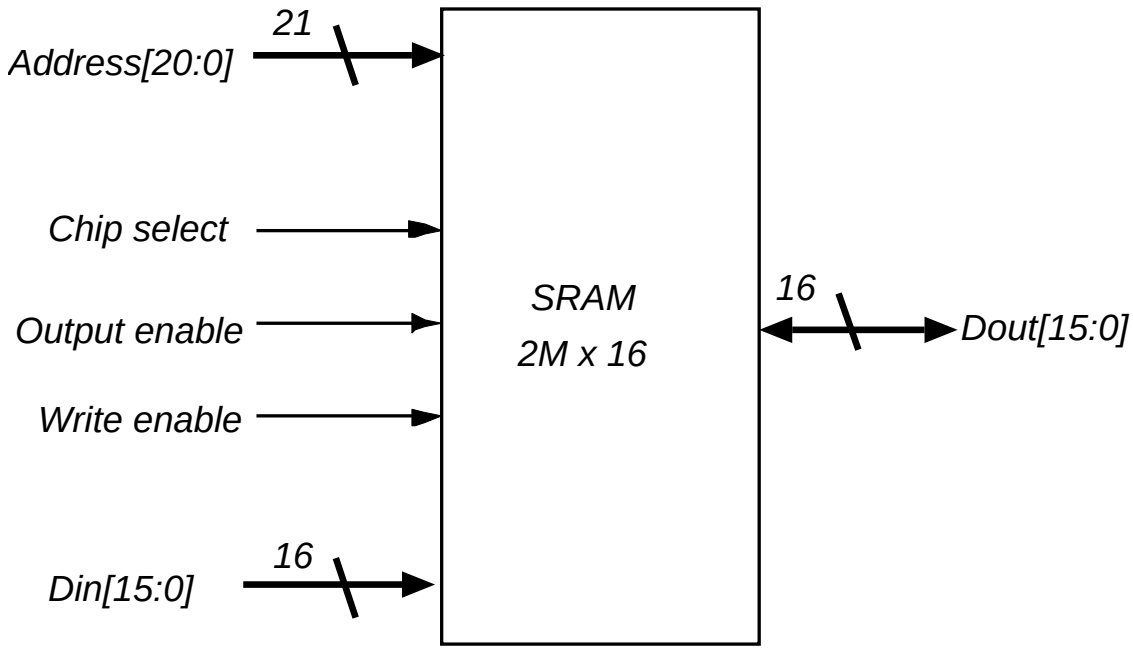
Circuitos de memória: organização conceptual



- Para P linhas de endereço: $N = 2^P$
- 2^{10} bytes $=$ 1024 bytes $=$ 1 KiB 2^{20} bytes $=$ 1048576 bytes $=$ 1 MiB
- O porto de dados é bidirecional:
  é preciso especificar o tipo de acesso (leitura ou escrita).
- M é a **largura** da memória (em número de bits).

Memórias estáticas

As memórias estáticas aproximam-se do modelo conceptual de funcionamento.



- Para aceder à memória:
  - ativar o circuito: *chip select* (CS) ativo
  - especificar o tipo de acesso:

    ativar *output enable* (leitura) **OU** *write enable* (escrita).

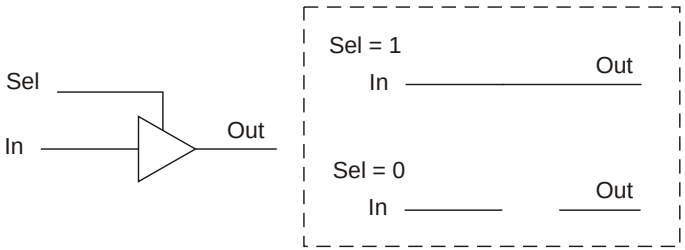
Memórias estáticas: acessos

- Tempo de acesso para leitura: intervalo entre o instante em que *output enable* e endereço estão corretos e o aparecimento de dados na saída.
- Valores típicos para memórias estáticas:
  - rápidas: 2–4 ns
  - típicas: 8–20 ns (cerca de 32 milhões de bits)
  - de baixo consumo: 5–10 vezes mais lentas
- Durante esse tempo, um processador que execute uma instrução por ciclo e use um relógio de 2 GHz, executa:
  - 4–8 instruções
  - 16–40 instruções
- Tempo de acesso para escrita: endereços e dados devem estar estáveis antes e depois do flanco. O sinal de *write enable* é sensível ao nível (não ao flanco) e deve ter uma duração mínima para que a escrita se realize.
- O tempo de escrita é superior ao tempo de leitura.

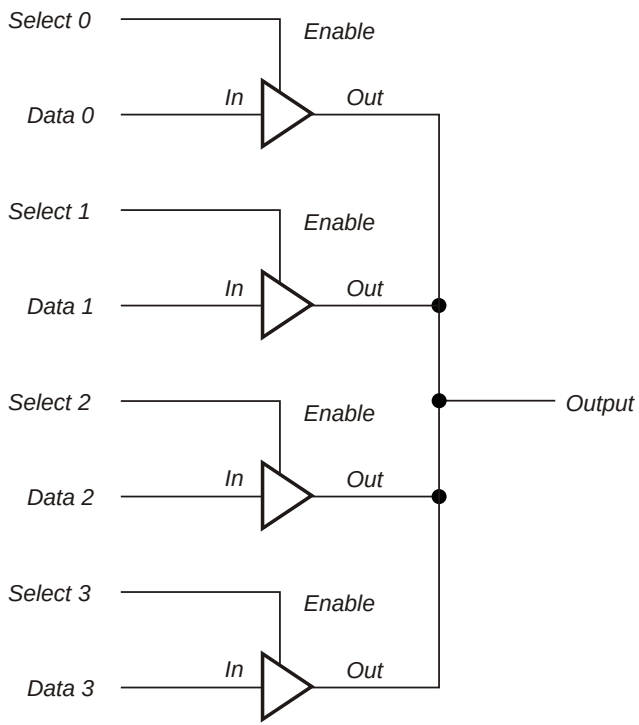
Memórias estáticas: circuito de saída

Buffer tristate

3 estados: 0, 1, desligado



Circuito de saída:



- Ao contrário de um banco de registos, o circuito de saída não pode ser baseado num multiplexador: uma SRAM 64K x 1 precisaria de ter um multiplexador 65536-para-1.

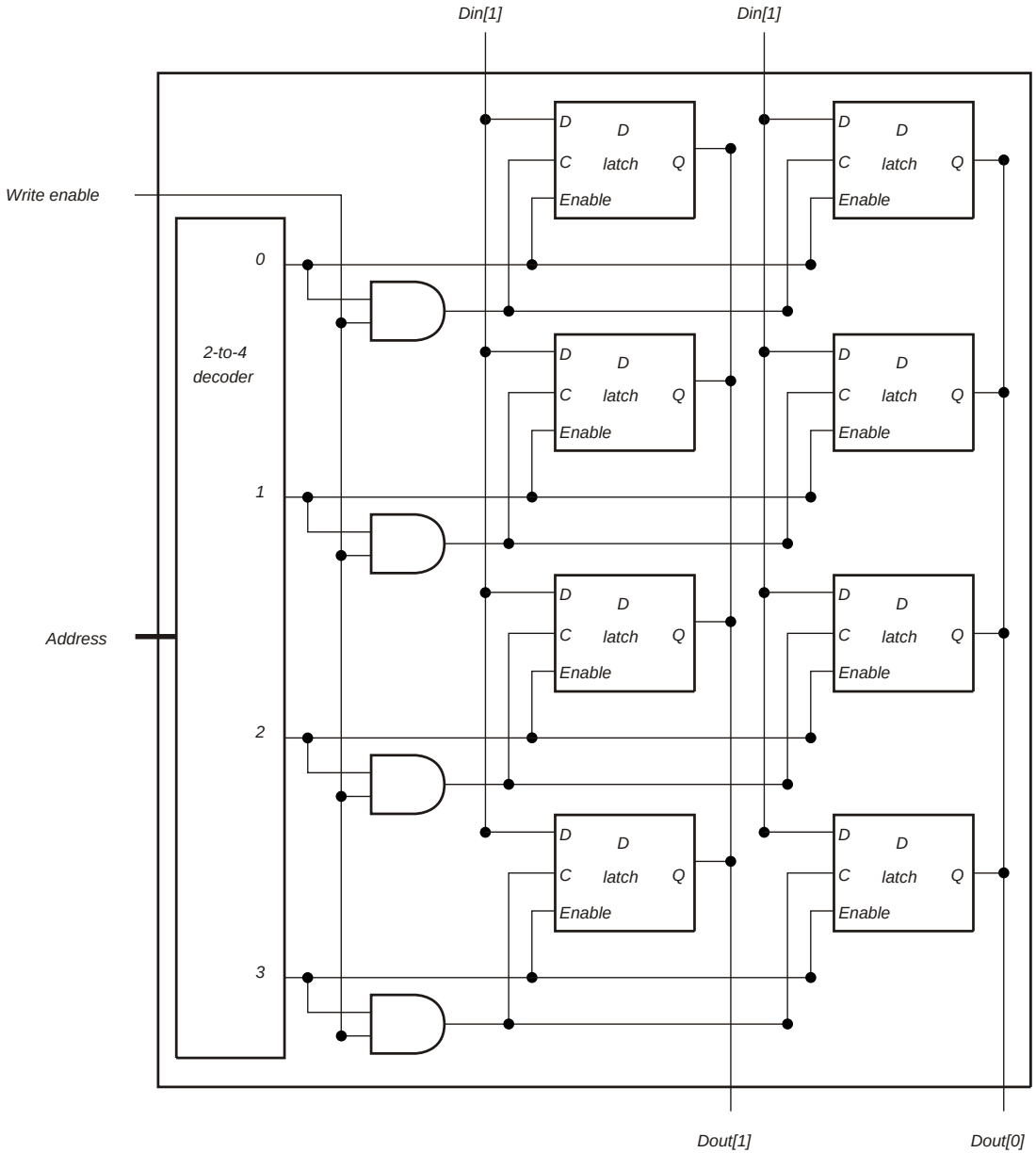
- Solução: utilizar *buffer tristate*, cuja saída pode ter 3 estados (0, 1 ou alta-impedância).

- No estado de alta-impedância, a saída do circuito está *desligada*.

- O estado da saída é determinado por uma entrada de controlo: `Sel`.

- Todas as saídas são ligadas em paralelo. **Não pode haver mais que uma saída ativa** (i.e., não em alta-impedância) a cada instante.

Estrutura básica de uma memória estática

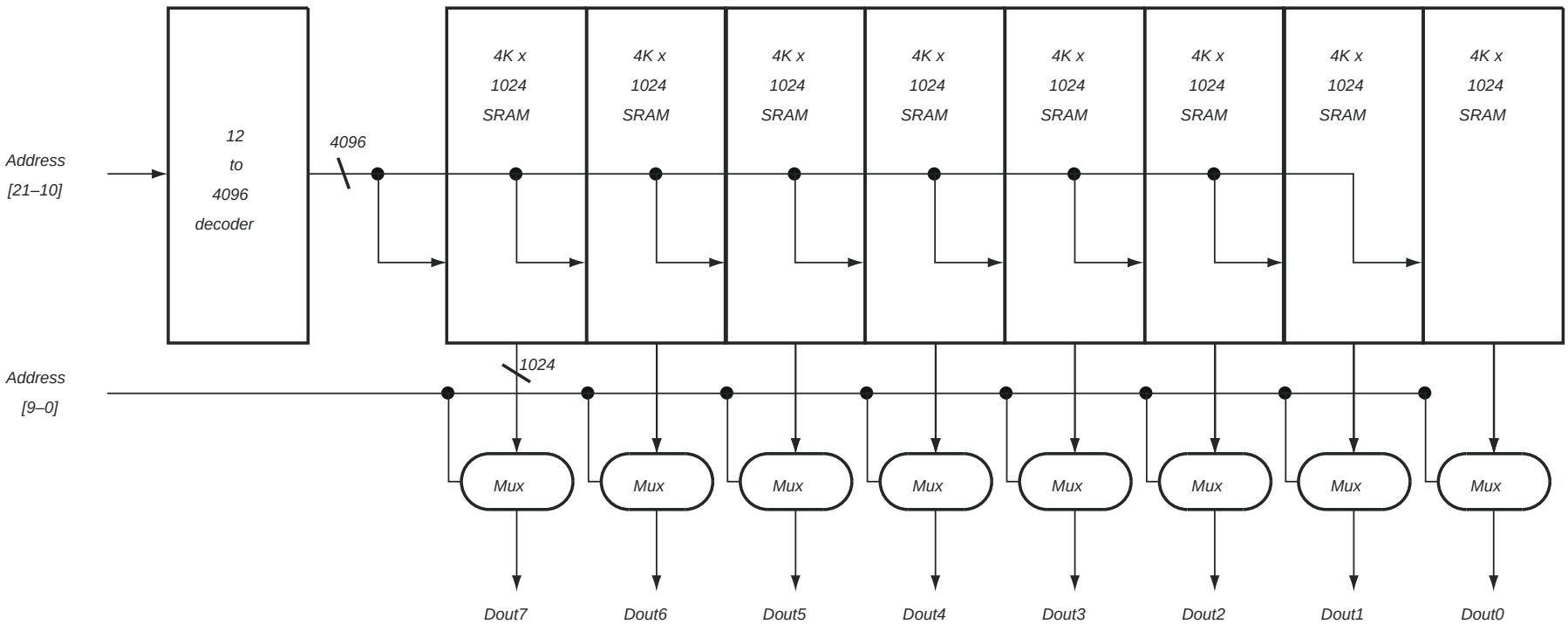


SRAM 4×2

Fonte: [COD3]

Memória estática organizada por bancos

Para limitar o tamanho do descodificador de endereços:



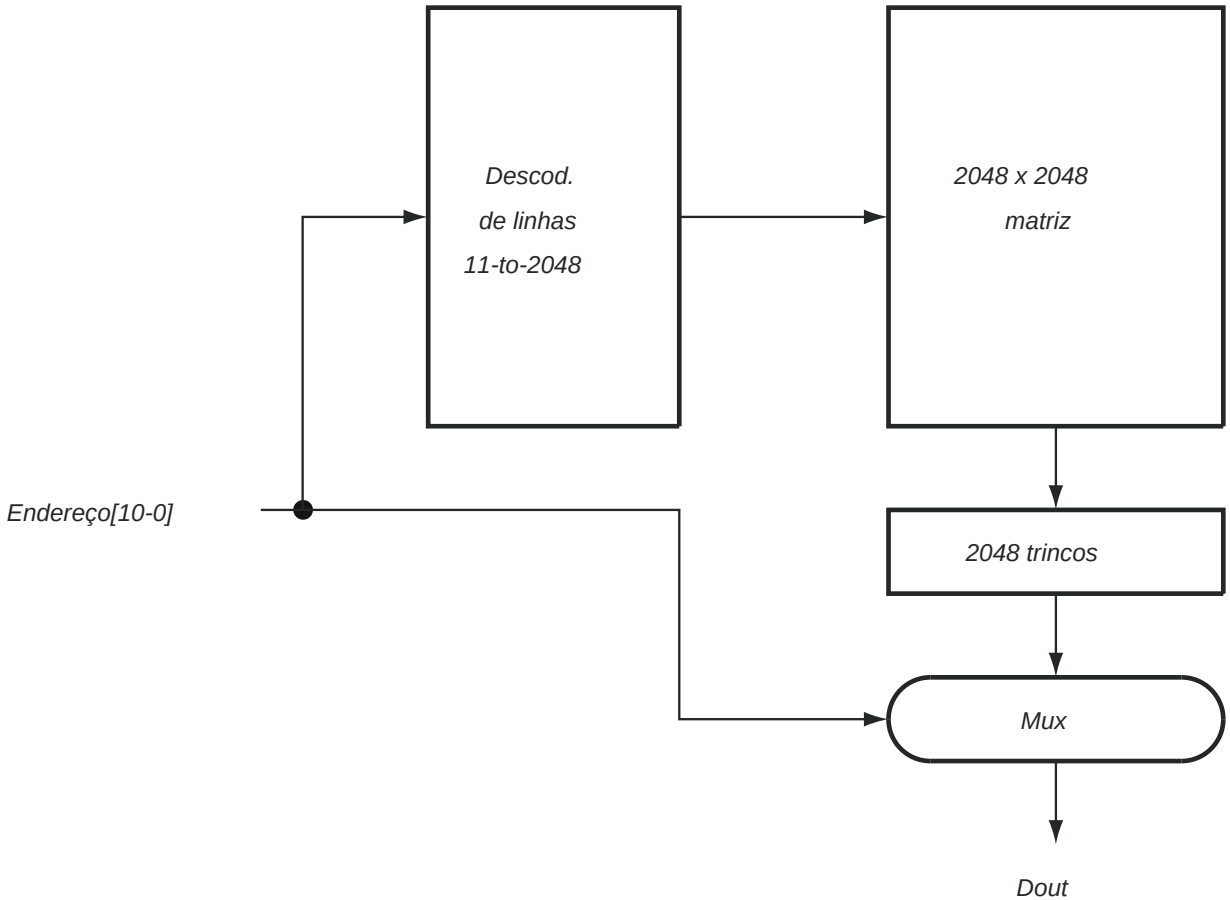
Fonte: [COD3]

- Organização típica de uma memória 4 MiB como um agrupamento de 8 blocos de memória de capacidade $(4 \times 2^{10}) \times 1024$ bits.
- Os blocos MUX são realizados por *buffers* de três estados.

Memória dinâmica (DRAM)

- Valor guardado como *carga num condensador*.
- O acesso é feito através de um transístor a operar como interruptor.
- Consequência: maior densidade (bit/mm^2), logo circuitos de maior capacidade e menor custo.
- Comparação: SRAM requer 4 a 6 transístores por bit armazenado.
- Acesso a DRAM é feito em duas etapas:
  1. seleção de coluna (usando uma parte do endereço);
  2. seleção de linha (usando os restantes bits do endereço).
- DRAM é mais lenta que SRAM. Exemplo:
  Capacidade: 2 Gibit ($(512 \times 2^{20}) \times 4$ bits); tempo de acesso 55 ns.
- Condensador vai perdendo a carga e deve ser periodicamente "refrescado", fazendo uma leitura seguida de escrita (circuito dinâmico). Acessos para refrescamento constituem 1 % to 2 % dos acessos.

Acesso a uma memória dinâmica (exemplo)

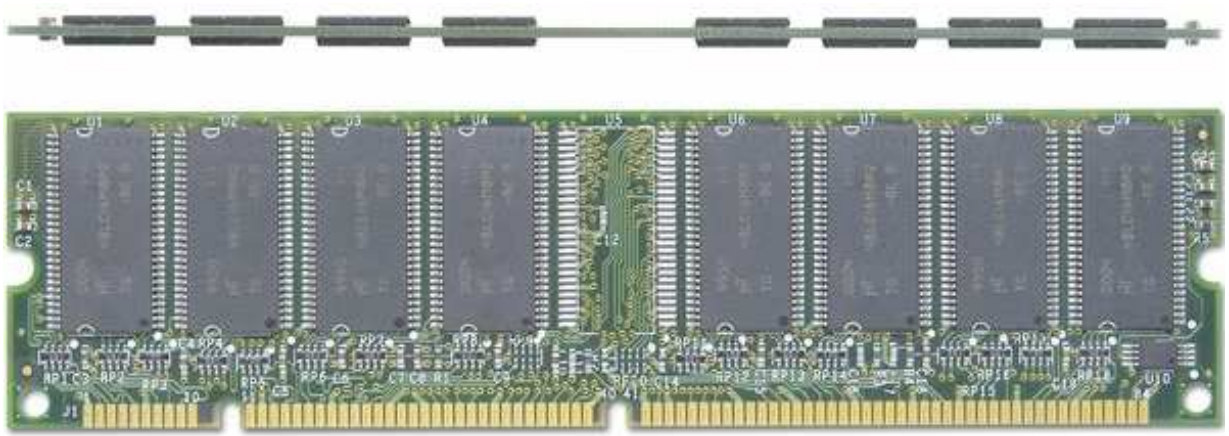


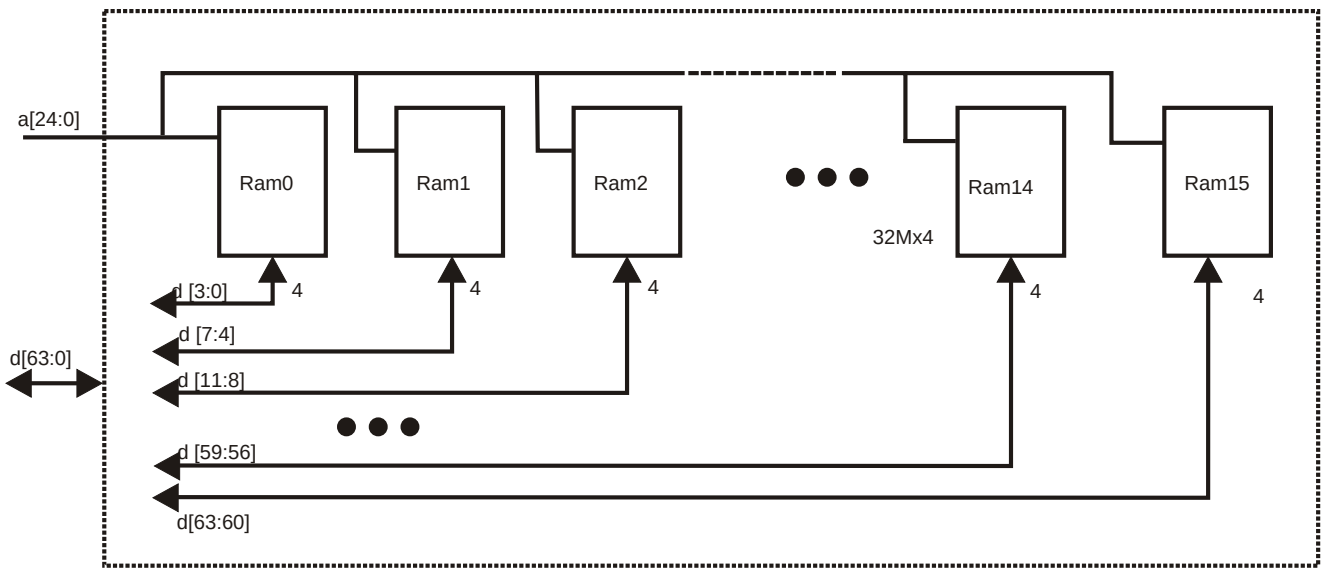
- Endereço: 11+11 bits.
- DRAM $(2 \times 2^{10}) \times 4$ bit: 11 bits selecionam uma linha, que é temporariamente armazenada em 2048 trincos.
- Multiplexador seleciona uma de 2048 entradas.

Módulos de memória: DIMM

CIs individuais podem ser agrupados em módulos.

Ex: módulo $(32 \times 2^{20}) \times 64$, i.e, de capacidade 256 MiB, pode usar 16 componentes $(32 \times 2^{20}) \times 4$.

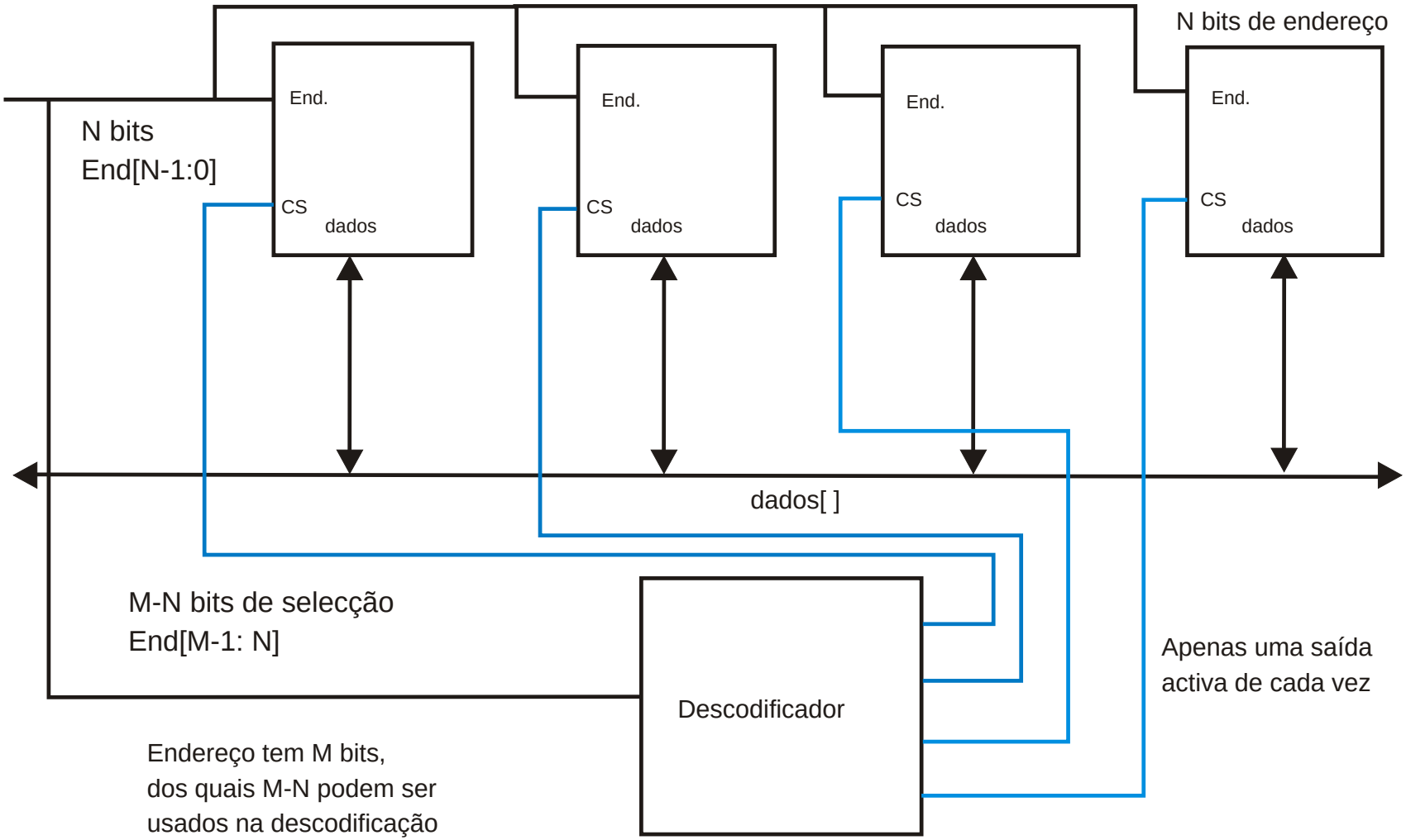




Organização da memória de um computador

- A memória física de um computador é geralmente composta por vários módulos (circuitos integrados, DIMM, etc.) por forma a ser possível obter maiores capacidades de armazenamento.
- Para além dos módulos de memória é necessário ter um circuito de descodificação de endereços que seleciona quais os módulos ativos durante um dado acesso (com base no endereço apresentado pelo CPU).
- Organização típica:
  - os bits menos significativos são ligados diretamente aos módulos individuais;
  - os bits mais significativos são usados para definir a ativação dos módulos.
- Linhas de dados podem ligadas a mais que um módulo (usando *buffers tristate*).

Organização da memória: diagrama de blocos



Nota: Para memórias DRAM, a descodificação de endereços é mais complicada; apenas abordaremos o caso das memórias SRAM e ROM (que é análogo).

Regras para descodificação de endereços

- Para que esta organização funcione bem, a descodificação de endereços deve garantir que:

  > Para o conjunto de todos os módulos que partilham uma mesma linha de dados: **apenas um** deve ser ativado durante um acesso.

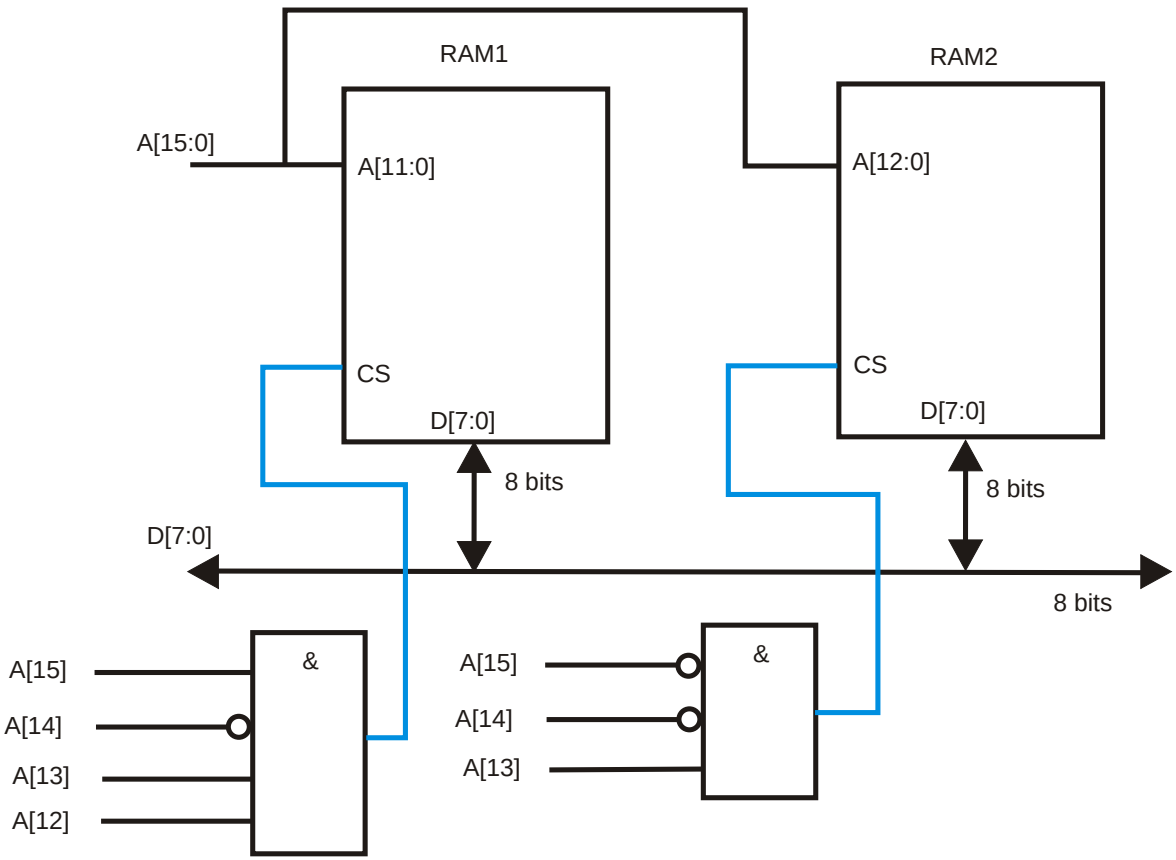
  Se esta condição não for respeitada, os componentes podem ser definitivamente danificados.

  ☞ O que acontece se nenhum módulo ser selecionado?

- O mapeamento de endereços para componentes pode ser classificado de acordo com o número de endereços que é mapeado na mesma posição física:
  - **total:** 1 endereço $\rightarrow$ 1 posição
  - **parcial:** N endereços $\rightarrow$ 1 posição

- Na descodificação total, todos os bits do endereço são usados: ligados diretamente aos componentes ou utilizados na seleção dos componentes.

Descodificação total: exemplo



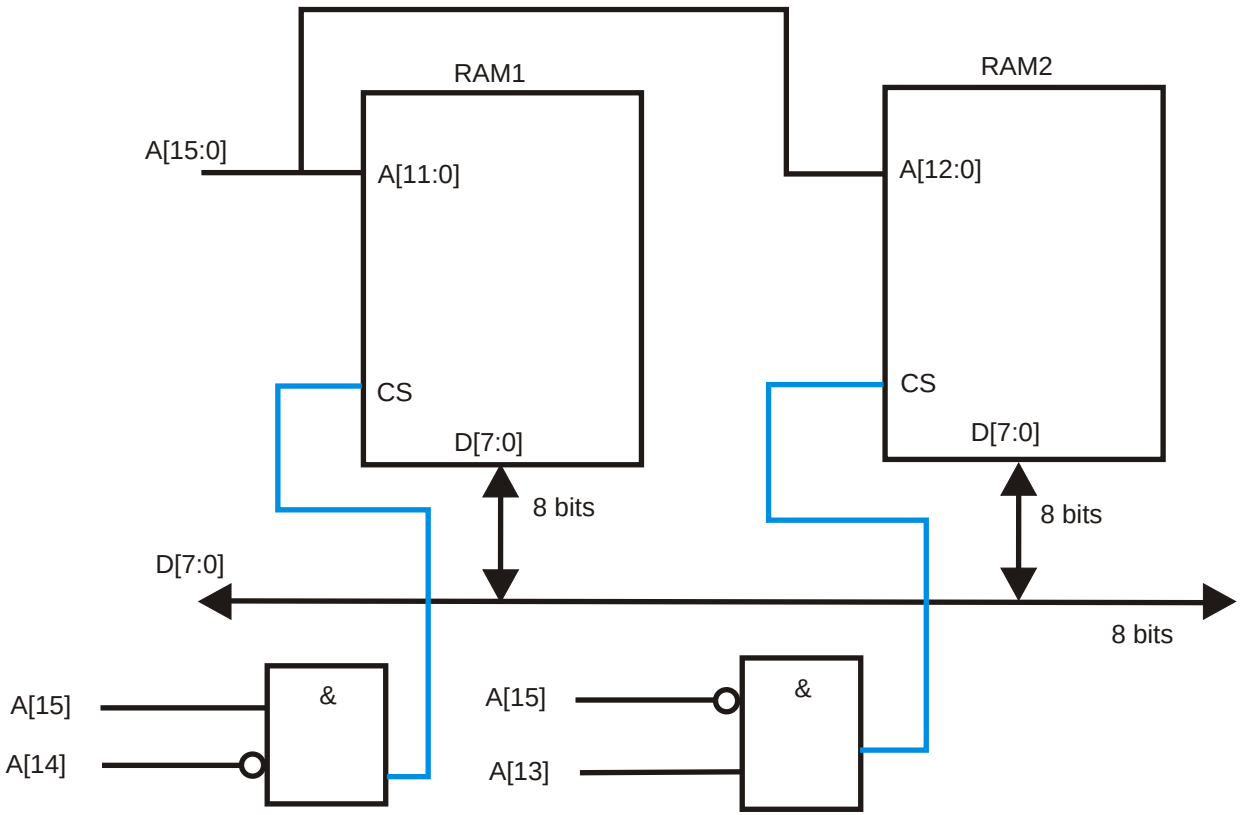
RAM1: 4Kx8 RAM2: 8Kx8
Espaço de endereçamento do CPU:
64 K, 1 byte por endereço

RAM 1:
1011 XXXX XXXX XXXX
Gama: B000H a BFFFH

RAM2:
001X XXXX XXXX XXXX
Gama: 2000H a 3FFFH

- Endereço B712H (46866) → RAM1
- Endereço C1E0H (49632) → nenhum circuito!

Descodificação parcial: exemplo



RAM1: 4Kx8 RAM2: 8Kx8
Espaço de endereçamento do CPU:
64 K, 1 byte por endereço

RAM 1:
10?? XXXX XXXX XXXX
Gamas:
8000H a 8FFFH
9000H a 9FFFH
A000H a AFFFH
B000H a BFFFH

RAM2:
0?1X XXXX XXXX XXXX
Gamas:
2000H a 3FFFH
6000H a 7FFFH

- O byte 10 de RAM1 pode ser acedido através de que endereços?
- 800AH, 900AH, A00AH e B00AH

Referências

COD4 D. A. Patterson & J. L. Hennessey, Computer Organization and Design, 4 ed.

COD3 D. A. Patterson & J. L. Hennessey, Computer Organization and Design, 3 ed.

Os tópicos tratados nesta apresentação são descritos na seguinte secção de [COD4]:

- apêndice C, secção C.9

Também são tratados na seguinte secção de [COD3]:

- apêndice B, secção B.9