Universidade Federal de Pelotas
Instituto de Física e Matemática
Departamento de Informática
Bacharelado em Ciência da Computação

Arquitetura e Organização de Computadores II

Aula 12

3. Hierarquia de Memória: introdução, princípio da localidade. Memória cache: conceitos básicos, organização, acesso.

Prof. José Luís Güntzel

guntzel@ufpel.edu.br

www.ufpel.edu.br/~guntzel/AOC2/AOC2.html

3. Hierarquia de Memória: memória cache

A Hierarquia de Memória

- **Os programas gastam a maior parte do tempo acessando a memória**
- **Programadores gostariam de ter ao ser dispor quantidade ilimitada de memória com acesso instantâneo**
- **O projeto do sistema de memória segue alguns princípios os quais tentam dar a ilusão ao programador de que ele dispõe de uma grande quantidade de memória com tempo de acesso pequeno**

3. Hierarquia de Memória: memória cache

A Hierarquia de Memória

- **Ao estudar uma determinada matéria, tu não precisas acessar todos os teus livros/cadernos.**
- **Portanto, basta deixar sobre tua mesa os livros que estão sendo usados para o estudo. Os demais livros podem ficar nos seus lugares, nas prateleiras…**
- **Talvez tua mesa não pudesse acomodar todos os teus livros/cadernos**
- **E caso pudesse, o tempo para encontrar a matéria em um livro seria demasiado grande, dificultando o estudo**

3. Hierarquia de Memória: memória cache

A Hierarquia de Memória

- **Localidade Temporal: "se um item é referenciado, ele tende a ser referenciado novamente dentro de um espaço curto de tempo."**
  - **A maioria dos programas contém laços (instruções e dados do laço tendem a ser acessados de maneira repetitiva).**
- **Localidade Espacial: "se um item é referenciado, itens cujos endereços sejam próximos dele tendem a ser logo referenciados."**
  - **Nos programas, as instruções estão armazenadas na memória de maneira seqüencial; os itens de matrizes e de registros também se encontram armazenados de maneira seqüencial.**

3. Hierarquia de Memória: memória cache

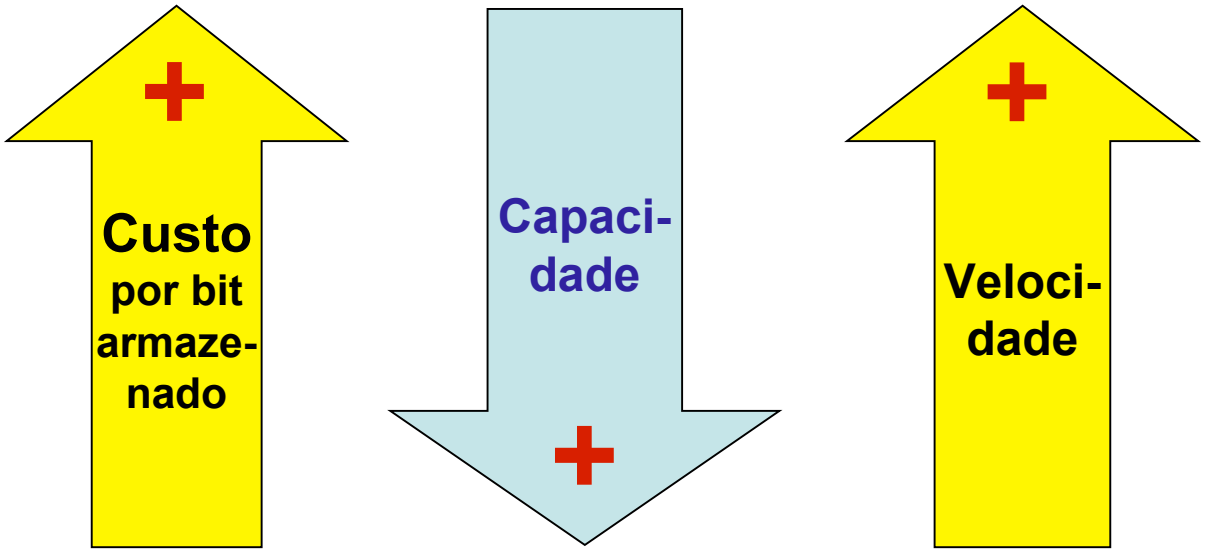
A Hierarquia de Memória

Tecnologias de Fabricação de Memórias

SRAM

DRAM

Magnética



3. Hierarquia de Memória: memória cache

A Hierarquia de Memória

Tecnologia de implementação	Tempo de acesso típico	Custo por Mbyte (em 1997)
SRAM	5-25 ns	$100 a $250
DRAM	60-120ns	$5 a $10
Disco magnético	10-20 milhões de ns	$0,10 a $0,20

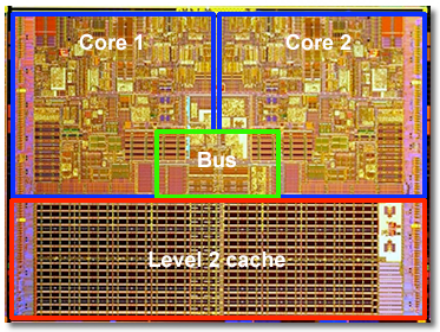
3. Hierarquia de Memória: memória cache

A Hierarquia de Memória

O Sistema de memória dos computadores é organizado de maneira hierárquica

Microprocessador:

- Caches integradas (L1, L2, L3…)
- Banco de registradores (32 a 128, tipicamente)



(SRAM)

cache externa

SRAM

memória principal

DRAM

Disco (HD)

magnética

3. Hierarquia de Memória: memória cache

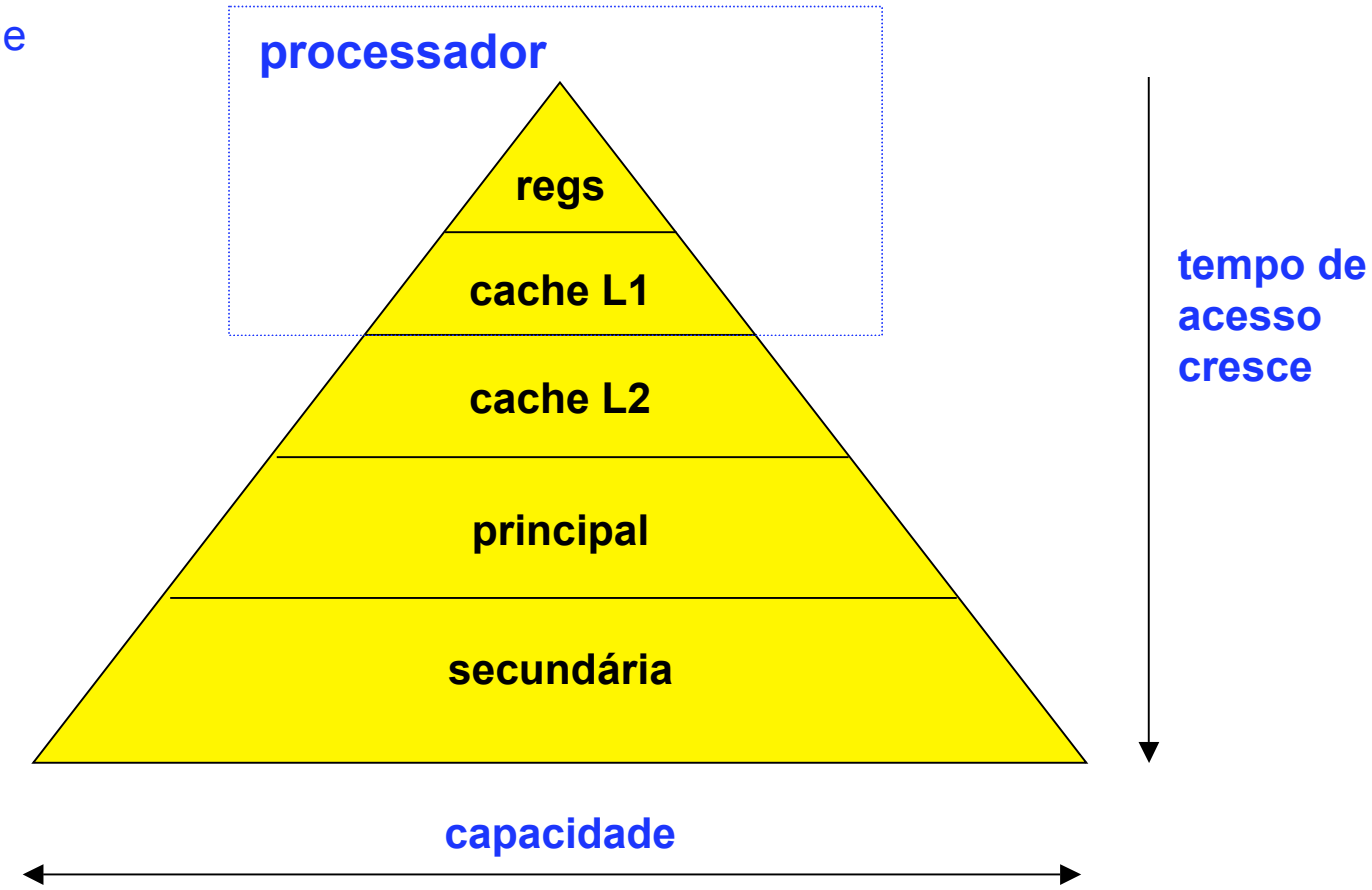
A Hierarquia de Memória

Objetivo do sistema hierárquico de memória é apresentar ao usuário uma capacidade de memória próxima à disponibilizada pela tecnologia mais barata, e um tempo de acesso próximo ao permitido pela tecnologia mais cara.

3. Hierarquia de Memória: memória cache

A Hierarquia de Memória

Obs: à medida que a tecnologia avança, outros níveis intermediários podem existir



3. Hierarquia de Memória: memória cache

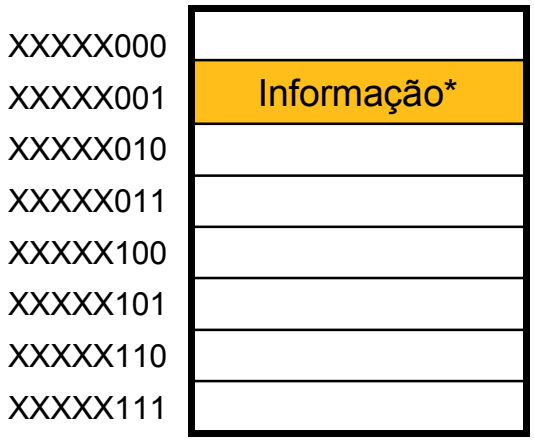
Definições

- **A princípio, uma hierarquia de memória pode ter qualquer número de níveis**
- **Entretanto, os dados sempre serão copiados entre dois níveis adjacentes (i e i+1, onde i está mais próximo do processador)**
- **Podemos concentrar nossa atenção em dois níveis quaisquer i e i+1: i, que chamaremos de superior (mais próximo do processador) e i+1, que chamaremos de inferior**

3. Hierarquia de Memória: memória cache

Definições

- **Bloco: unidade mínima de informação, contendo um certo número de palavras de memória.**
- **Exemplo, com 8 palavras (de memória)**



* informação = instrução ou dado

3. Hierarquia de Memória: memória cache

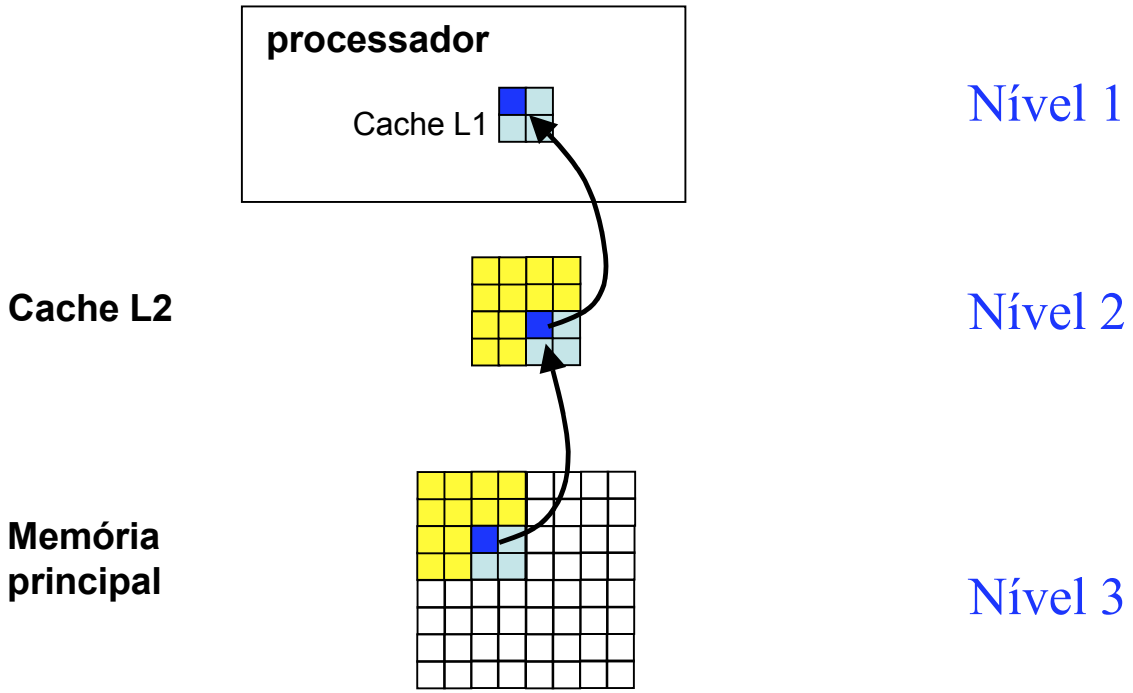
Definições

- **Se a informação solicitada pelo processador estiver presente no nível superior da hierarquia, ocorre um acerto (*hit*)**
- **Se a informação solicitada pelo processador não puder ser encontrada no nível superior, a tentativa de encotrá-la gera uma falta (*fault*)**
- **Quando ocorre uma falta, o nível imediatamente inferior é acessado, na tentativa de se recuperar o bloco com a informação solicitada pelo processador**

3. Hierarquia de Memória: memória cache

Definições

- **Se um bloco está presente no nível i, então ele também está presente no nível i+1**



3. Hierarquia de Memória: memória cache

Definições

- **A taxa de acertos ou razão de acertos (*hit ratio*) corresponde à fração dos acessos à memória encontrados no nível superior (com freqüência, é usada como medida de desempenho do sistema de memória)**
- **A taxa de faltas (= 1- taxa de acertos) é a fração de acessos à memória não encontrados no nível superior**

3. Hierarquia de Memória: memória cache

Definições

- **Tempo de acerto (*hit time*) é o tempo necessário para acessar o nível superior da hierarquia, que inclui o tempo necessário para determinar se a tentativa de acesso à informação vai gerar um acerto ou uma falta**
- **A penalidade por falta (*fault penalty*) é o tempo necessário para substituir um dos blocos do nível superior pelo bloco do nível inferior que contém a informação desejada, mais o tempo para enviar a informação ao processador**

Tempo de acerto << Tempo de acesso ao nível imediatamente inferior

3. Hierarquia de Memória: memória cache

Memória Cache

- ***Cache*, em inglês: lugar seguro para esconder ou guardar algo**
- ***Cacher*, em francês: esconder, guardar**
- **Nome usado para designar o nível de memória entre o processador e a memória principal**
- **Este nome foi usado pela máquina que introduziu pioneiramente (no início dos anos 1960) este nível de memória (entre a memória principal e o processador)**
- **Cache explora o princípio da localidade**
- **Hoje em dia, usa-se este nome para designar qualque memória que explore o princípio da localidade**

3. Hierarquia de Memória: memória cache

Memória Cache

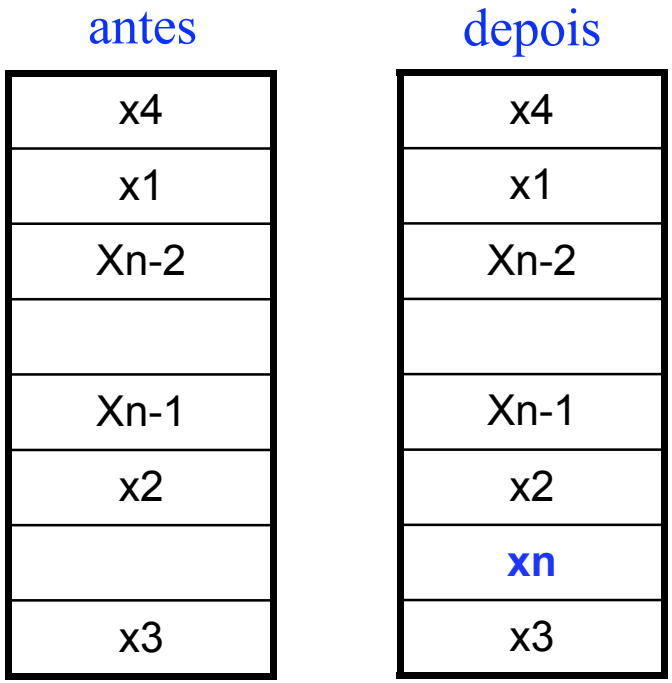
Assumamos as seguintes características de um sistema de memória extremamente simples:

- **O processador sempre requisita uma única palavra**
- **Existe apenas um nívvel de memória cache (L1)**
- **Os blocos de L1 são constituídos por somente uma palavra**

3. Hierarquia de Memória: memória cache

Memória Cache

Fazendo referência ao dado xn



- **Como saber se uma informação está na cache?**
- **Caso ela esteja, como encontrá-la?**

3. Hierarquia de Memória: memória cache

Memória Cache

Mapeamento Direto:

- **Para cada palavra na cache, atribuir um endereço com base no endereço da palavra na memória principal**
- **A maioria das caches que usa mapeamento direto o faz usando o seguinte processo:**

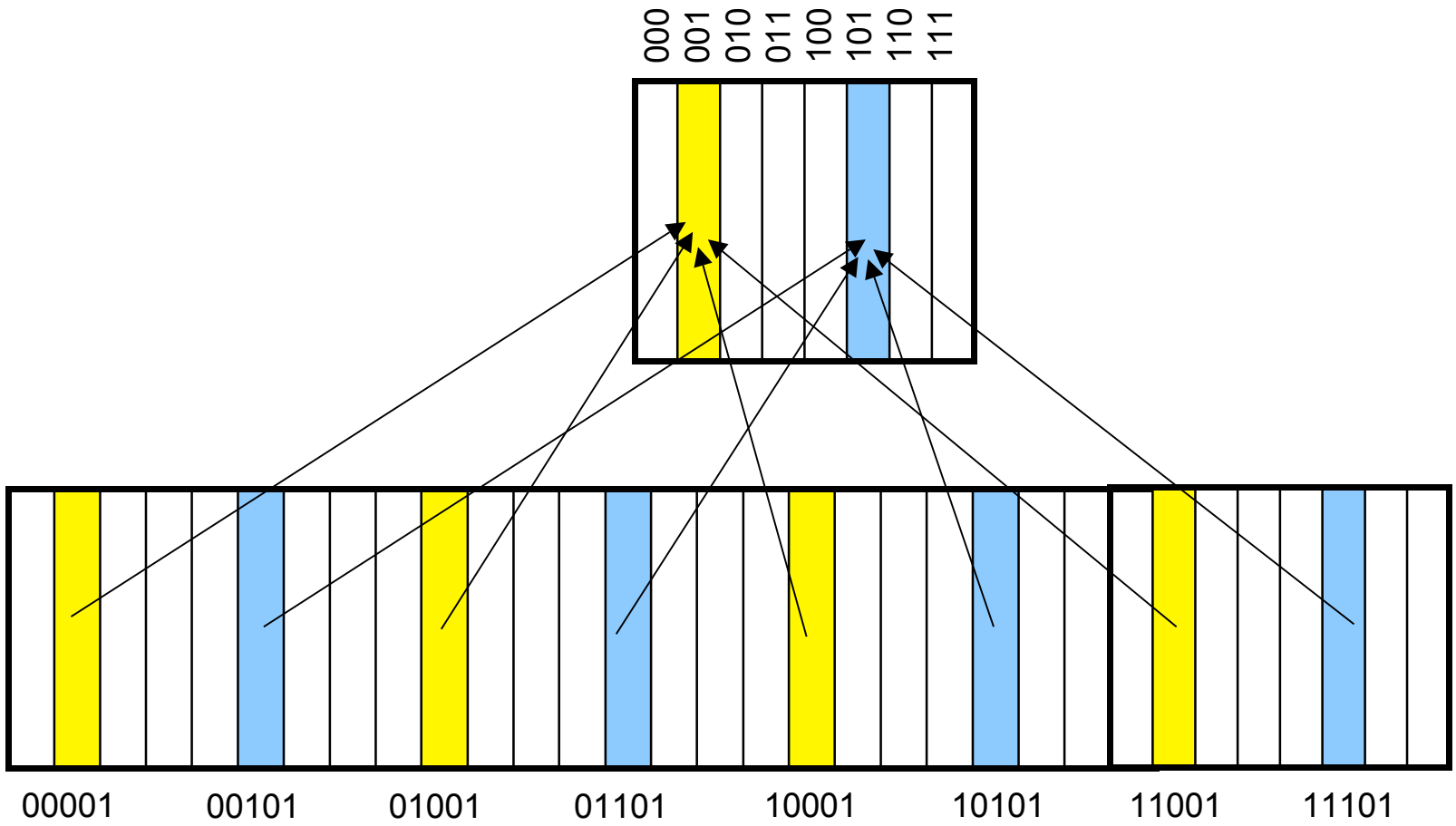
(Endereço do bloco) módulo (Número de blocos da cache)

Endereço absoluto (i.e., em relação à memória principal)

resto da divisão inteira

3. Hierarquia de Memória: memória cache

Memória Cache: mapeamento direto



3. Hierarquia de Memória: memória cache

Memória Cache: mapeamento direto

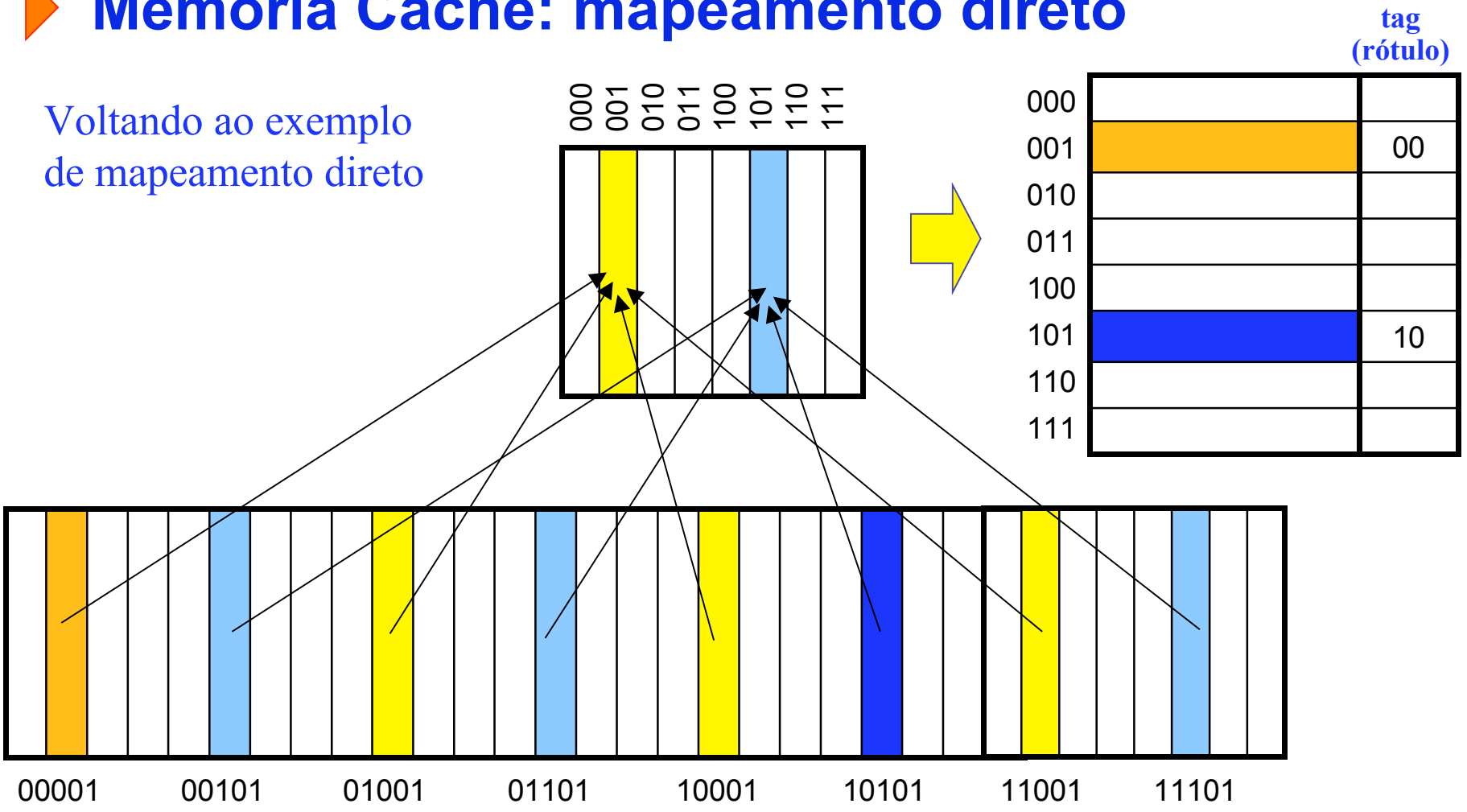
Dado que uma "entrada" da cache pode armazenar o conteúdo de diversos endereços da memória (principal), como identificar se o dado armazenado na cache corresponde ao solicitado?

- **Solução: atribuir à cache um conjunto de rótulos (*tags*)**
- **Os rótulos são usados em conjunto com o endereço do mapeamento, de modo a compor o endereço completo, com relação à memória principal**

3. Hierarquia de Memória: memória cache

Memória Cache: mapeamento direto

Voltando ao exemplo
de mapeamento direto



3. Hierarquia de Memória: memória cache

Memória Cache: mapeamento direto

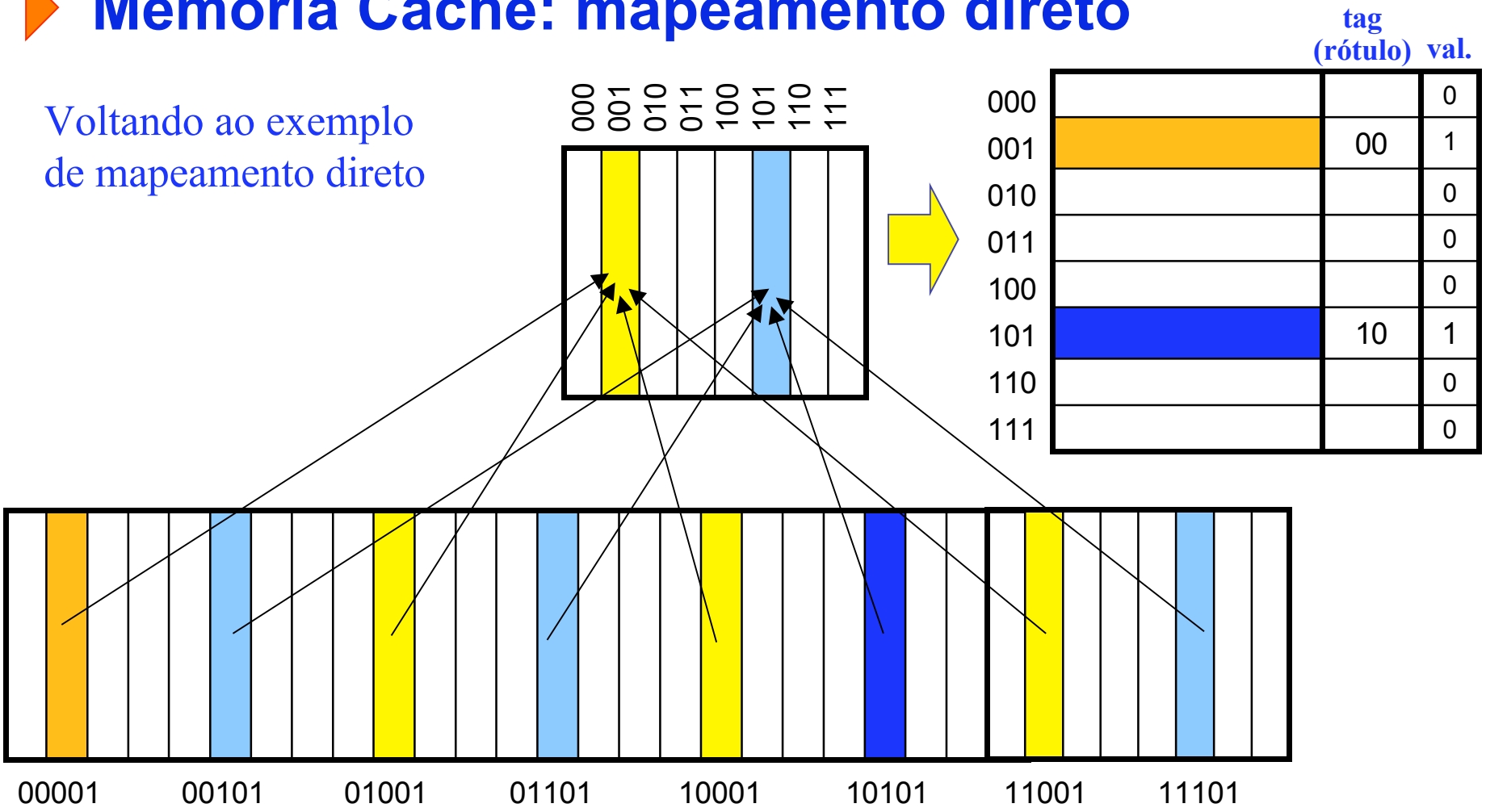
E como reconhecer se um bloco da cache possui uma informação válida? (Quando o processador é inicializado, por exemplo, algo deve sinalizar que a cache está "vazia")

- **Solução: incluir um bit de validade**
- **Se o bit de validade = 0, a informação contida naquele bloco da cache não é válida**

3. Hierarquia de Memória: memória cache

Memória Cache: mapeamento direto

Voltando ao exemplo
de mapeamento direto



3. Hierarquia de Memória: memória cache

Acesso a uma Cache (para leitura)

(1) Estado inicial da cache, após inicialização da máquina

OBS: tags e endereços em binário

índice	val.	tag	informação
000	0		
001	0		
010	0		
011	0		
100	0		
101	0		
110	0		
111	0		

OBS: todos os valores neste exemplo estão em binário

3. Hierarquia de Memória: memória cache

Acesso a uma Cache (para leitura)

(2) Referência ao endereço 10110: falta

índice	val.	tag	informação
000	0		
001	0		
010	0		
011	0		
100	0		
101	0		
110	0		
111	0		

3. Hierarquia de Memória: memória cache

Acesso a uma Cache (para leitura)

(2) Referência ao endereço 10110: falta: tratamento da falta

índice	val.	tag	informação
000	0		
001	0		
010	0		
011	0		
100	0		
101	0		
110	1	10	Memória(10110)
111	0		

Tratamento da falta = buscar no nível inferior (neste exemplo, a memória principal) o bloco com endereço 10110

3. Hierarquia de Memória: memória cache

Acesso a uma Cache (para leitura)

(3) Referência ao endereço 11010: falta

índice	val.	tag	informação
000	0		
001	0		
010	0		
011	0		
100	0		
101	0		
110	1	10	Memória(10110)
111	0		

3. Hierarquia de Memória: memória cache

Acesso a uma Cache (para leitura)

(3) Referência ao endereço 11010: falta: tratamento da falta

índice	val.	tag	informação
000	0		
001	0		
010	1	11	Memória(11010)
011	0		
100	0		
101	0		
110	1	10	Memória(10110)
111	0		

Tratamento da falta = buscar no nível inferior (neste exemplo, a memória principal) o bloco com endereço 11010

3. Hierarquia de Memória: memória cache

Acesso a uma Cache (para leitura)

(4) Referência ao endereço 10000: falta

índice	val.	tag	informação
000	0		
001	0		
010	1	11	Memória(11010)
011	0		
100	0		
101	0		
110	1	10	Memória(10110)
111	0		

3. Hierarquia de Memória: memória cache

Acesso a uma Cache (para leitura)

(4) Referência ao endereço 10000: falta: tratamento da falta

índice	val.	tag	informação
000	1	10	Memória(10000)
001	0		
010	1	11	Memória(11010)
011	0		
100	0		
101	0		
110	1	10	Memória(10110)
111	0		

Tratamento da falta = buscar no nível inferior (neste exemplo, a memória principal) o bloco com endereço 10000

3. Hierarquia de Memória: memória cache

Acesso a uma Cache (para leitura)

(5) Referência ao endereço 00011: falta

índice	val.	tag	informação
000	1	10	Memória(10000)
001	0		
010	1	11	Memória(11010)
011	0		
100	0		
101	0		
110	1	10	Memória(10110)
111	0		

3. Hierarquia de Memória: memória cache

Acesso a uma Cache (para leitura)

(5) Referência ao endereço 00011: falta: tratamento da falta

índice	val.	tag	informação
000	1	10	Memória(10000)
001	0		
010	1	11	Memória(11010)
011	1	00	Memória(00011)
100	0		
101	0		
110	1	10	Memória(10110)
111	0		

Tratamento da falta = buscar no nível inferior (neste exemplo, a memória principal) o bloco com endereço 00011

3. Hierarquia de Memória: memória cache

Acesso a uma Cache (para leitura)

(6) Referência ao endereço 10010: falta

índice	val.	tag	informação
000	1	10	Memória(10000)
001	0		
010	1	11	Memória(11010)
011	1	00	Memória(00011)
100	0		
101	0		
110	1	10	Memória(10110)
111	0		

3. Hierarquia de Memória: memória cache

Acesso a uma Cache (para leitura)

(6) Referência ao endereço 10010: falta: tratamento da falta

índice	val.	tag	informação
000	1	10	Memória(10000)
001	0		
010	1	10	Memória(10010)
011	1	00	Memória(00011)
100	0		
101	0		
110	1	10	Memória(10110)
111	0		

Tratamento da falta = buscar no nível inferior (neste exemplo, a memória principal) o bloco com endereço 00011, escrevendo-o na posição 010 da cache