Arquitetura de computadores e sistemas operacionais

Hierarquia de memória

Gabriel V C Candido
gabriel.candido@ifpr.edu.br

Instituto Federal do Paraná - Pinhais

INSTITUTO FEDERAL
Paraná
Campus Pinhais

Sumário

Introdução

Hierarquia de memória

Caches

Sumário

Introdução

Hierarquia de memória

Caches

Motivação

- o que queremos?

Motivação

- o que queremos? memória infinita

Motivação

- o que queremos? memória infinita
- o que queremos?

Motivação

- o que queremos? memória infinita
- o que queremos? memória rápida

Motivação

- o que queremos? memória infinita
- o que queremos? memória rápida
- quando queremos?

Motivação

- o que queremos? memória infinita
- o que queremos? memória rápida
- quando queremos? desde sempre

Analogia com biblioteca

Você precisa escrever sobre a história da computação, com foco nos desenvolvimentos de hardware

Analogia com biblioteca

Você precisa escrever sobre a história da computação, com foco nos desenvolvimentos de hardware

Você pega vários livros e senta à mesa para analisar. Percebe que está faltando um modelo específico e importante de hardware.

Analogia com biblioteca

Você precisa escrever sobre a história da computação, com foco nos desenvolvimentos de hardware

Você pega vários livros e senta à mesa para analisar. Percebe que está faltando um modelo específico e importante de hardware.

Vai até as estantes e encontra um livro sobre o início da computação que trata do tópico desejado.

Analogia com biblioteca

Você precisa escrever sobre a história da computação, com foco nos desenvolvimentos de hardware

Você pega vários livros e senta à mesa para analisar. Percebe que está faltando um modelo específico e importante de hardware.

Vai até as estantes e encontra um livro sobre o início da computação que trata do tópico desejado.

Volta à mesa com o livro e consulta a seleção de livros enquanto escreve o trabalho

Princípio da localidade

Localidade temporal: se um item é referenciado, a tendência é que seja referenciado novamente em breve.

Você provavelmente vai olhar para os livros que trouxe à mesa em um futuro próximo.

Princípio da localidade

Localidade espacial: se um item é referenciado, itens próximos provavelmente serão referenciados em breve.

Ao buscar o livro sobre o início da computação, notou que o livro ao lado falava sobre computadores mecânicos e também trouxe esse livro.
Posteriormente, achou coisas interessantes para escrever no trabalho.

Princípio da localidade

Programas de computador também exibem localidade:

- laços acessam os mesmos dados e instruções próximas repetidamente (tempo)
- instruções são sequenciais (espaço)

Hierarquias de memórias aproveitam o princípio da localidade

Sumário

Introdução

Hierarquia de memória

Caches

Hierarquia de memória

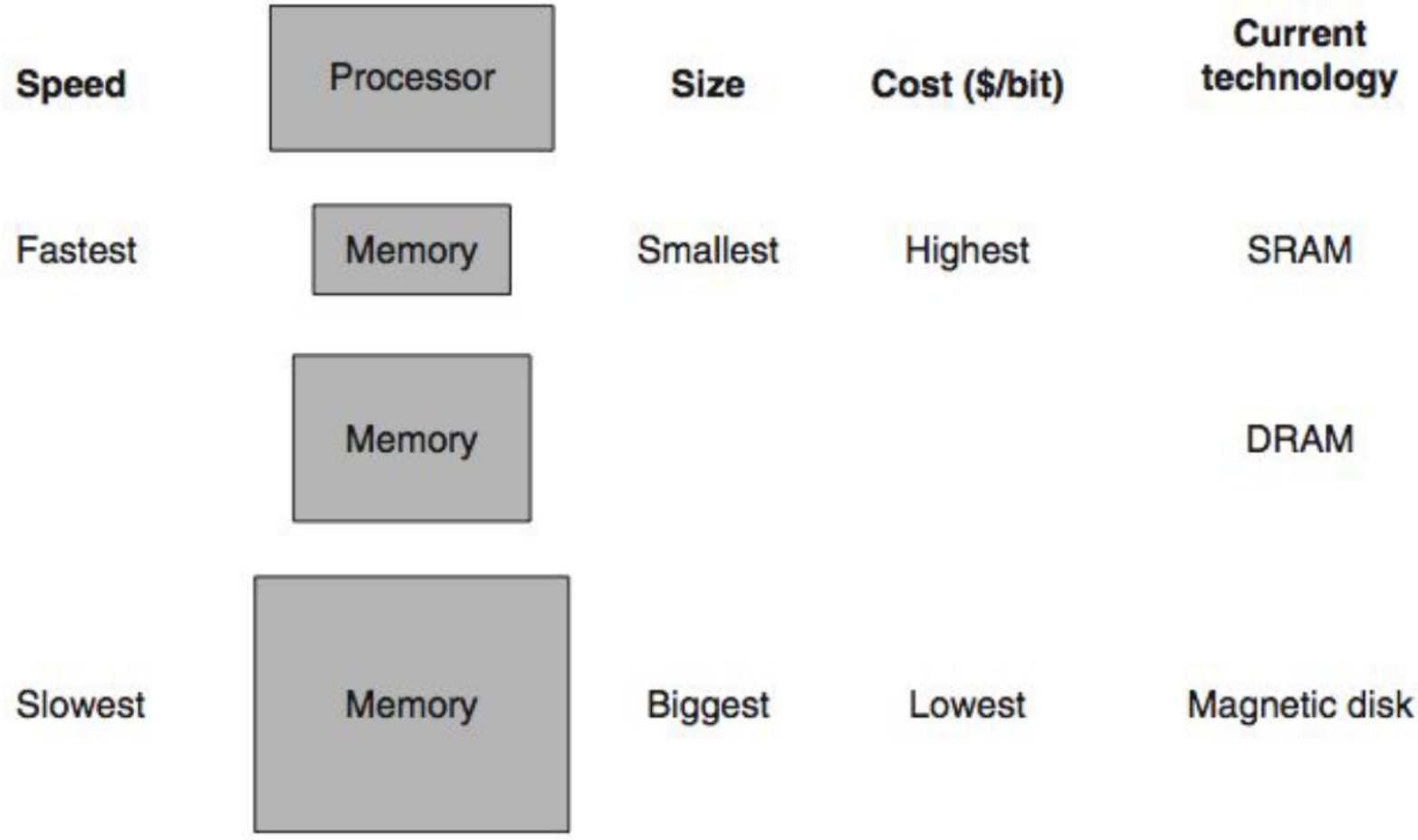

Figura: Esquema básico de memória hierárquica. Fonte: PH 5.1

Hierarquia de memória

Memory technology	Typical access time	$ per GB in 2008
SRAM	0.5–2.5 ns	$2000–$5000
DRAM	50–70 ns	$20–$75
Magnetic disk	5,000,000–20,000,000 ns	$0.20–$2

Figura: Tipos de memória, tempos de acesso e custo. Fonte: PH 5.1

Sumário

Introdução

Hierarquia de memória

Caches

Memórias cache

Memória entre o processador e a memória principal (RAM), "escondida": o processador não acessa a cache diretamente

Memórias cache

Memória entre o processador e a memória principal (RAM), "escondida": o processador não acessa a cache diretamente

Como saber se um dado está na cache?

Memórias cache

Memória entre o processador e a memória principal (RAM), "escondida": o processador não acessa a cache diretamente

Como saber se um dado está na cache?

Como achar o dado?

Memórias cache

Memória entre o processador e a memória principal (RAM), "escondida": o processador não acessa a cache diretamente

Como saber se um dado está na cache?

Como achar o dado?

Primeiro grande problema: como mapear a memória principal (grande) na memória cache (pequena)?

Acesso à cache

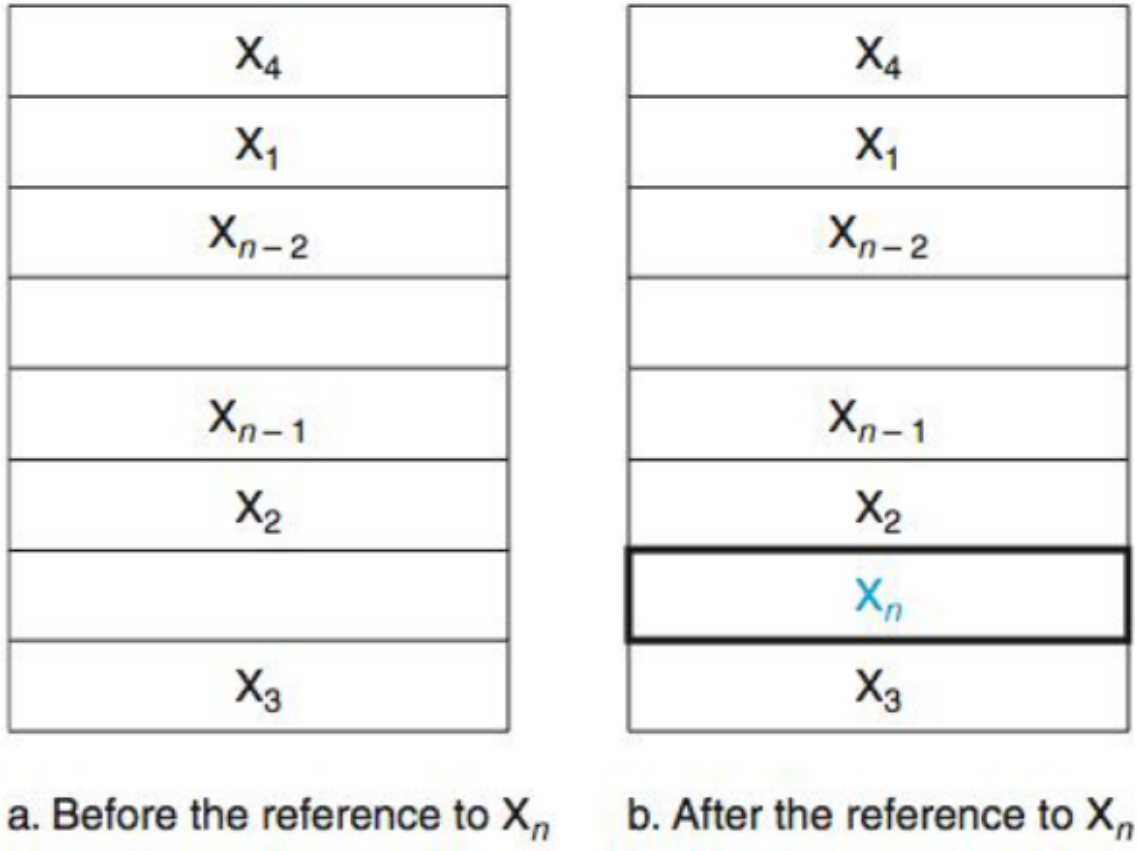

Figura: Acesso à cache. Fonte: PH 5.2

Caches em processadores comerciais

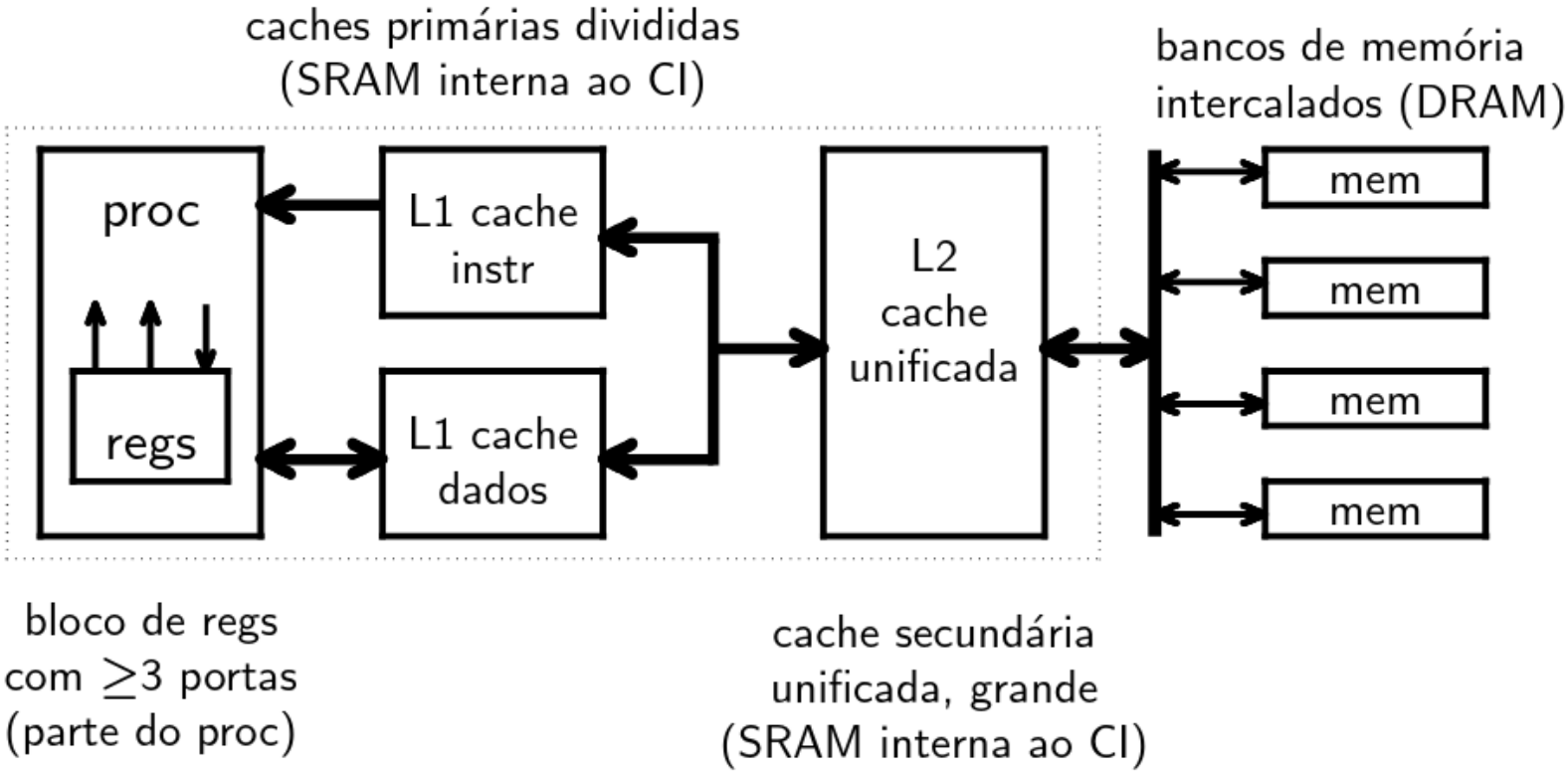

Figura: Cache de processadores comerciais. Fonte: RH

Cache diretamente mapeada

Mapear para a cache usando os bits menos significativos do endereço de memória. Isso resulta em posições fixas para cada endereço na memória principal. *Potências de 2!*

Cache diretamente mapeada

Mapear para a cache usando os bits menos significativos do endereço de memória. Isso resulta em posições fixas para cada endereço na memória principal. *Potências de 2!*

Tags indicam qual a entrada da memória principal está mapeada naquele momento. Um bit de validade indica se o conteúdo da cache pode ser usado ou não (por exemplo, ao ligar o computador, o conteúdo é lixo)

A ideia é dividir o trabalho em uma linha de produção para aumentar a eficiência

Cache diretamente mapeada

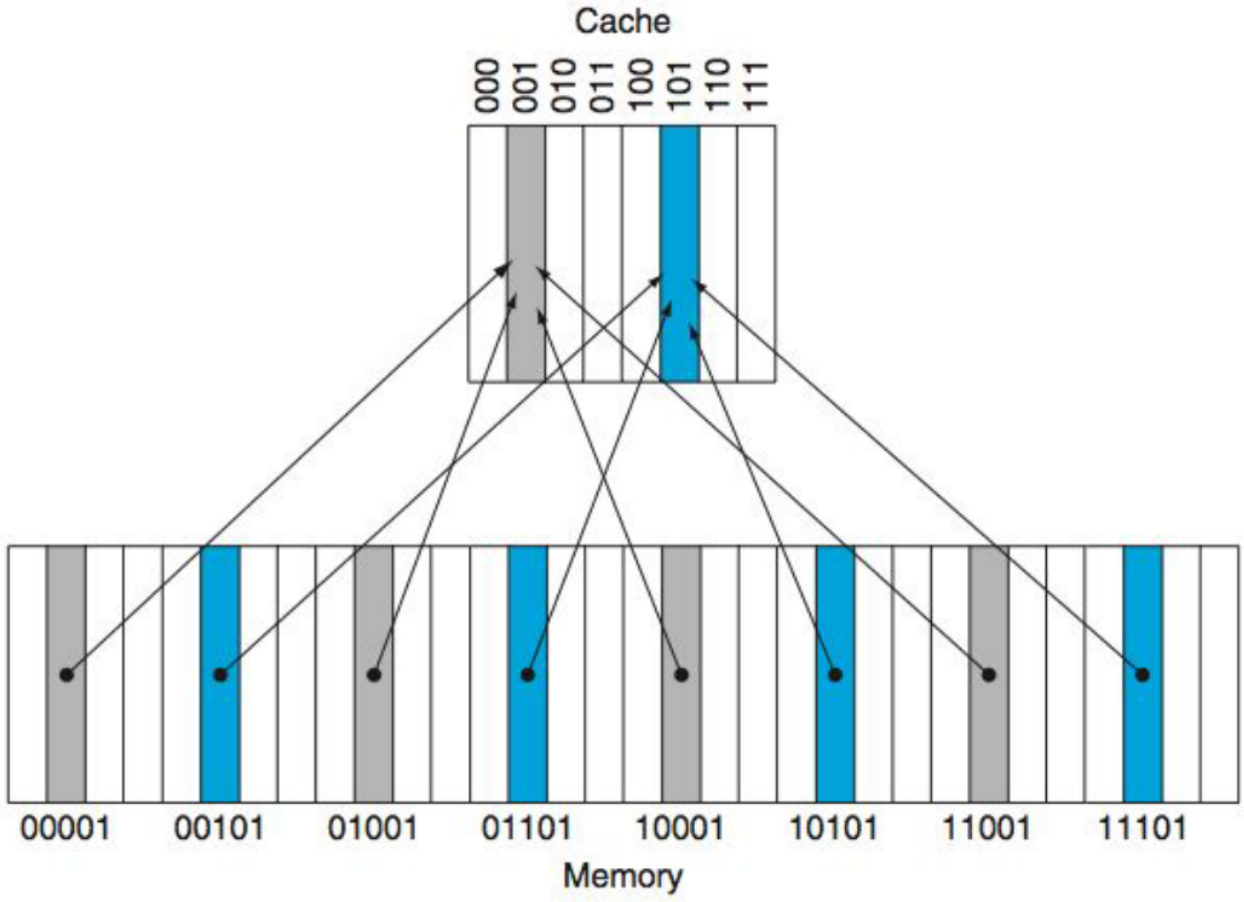

Figura: Cache diretamente mapeada. Fonte: PH 5.2

Memórias cache

O que acontece se o dado está na cache?

Memórias cache

O que acontece se o dado está na cache?
Processador pode usar o dado imediatamente (*hit*)

Memórias cache

O que acontece se o dado está na cache?
Processador pode usar o dado imediatamente (*hit*)

O que acontece se o dado não está na cache?

Memórias cache

O que acontece se o dado está na cache?
Processador pode usar o dado imediatamente (*hit*)

O que acontece se o dado não está na cache?
Unidade de gerência de memória precisa buscar o dado na RAM e trazer para a cache (*miss*).

Miss penalty: penalidade da falta na cache. Quanto tempo demora para trazer o dado?

Memórias cache e escrita

Como manter a consistência entre o que está na cache e o que está na RAM? Esse é um grande problema!

Memórias cache e escrita

Como manter a consistência entre o que está na cache e o que está na RAM? Esse é um grande problema!

Write-through: escreve na cache e no nível de cima.

Memórias cache e escrita

Como manter a consistência entre o que está na cache e o que está na RAM? Esse é um grande problema!

Write-through: escreve na cache e no nível de cima.

E se o dado não estiver na memória? *Miss*!

Memórias cache e escrita

Write buffer: armazena dado enquanto espera a RAM escrever. Processador continua executando.

Memórias cache e escrita

Write buffer: armazena dado enquanto espera a RAM escrever. Processador continua executando.

E se encher o buffer? Processador precisa esperar (*stall*)

Memórias cache e escrita

Write-back: escreve somente na cache e só atualiza na RAM quando for remover o bloco da cache. Bit *dirty*.

Memórias cache e escrita

Write-back: escreve somente na cache e só atualiza na RAM quando for remover o bloco da cache. Bit *dirty*.

Mais complexo!

Mapeamento associativo

Cache possui várias vias para cada índice. Cada endereço pode estar estar em qualquer via.

Mapeamento associativo

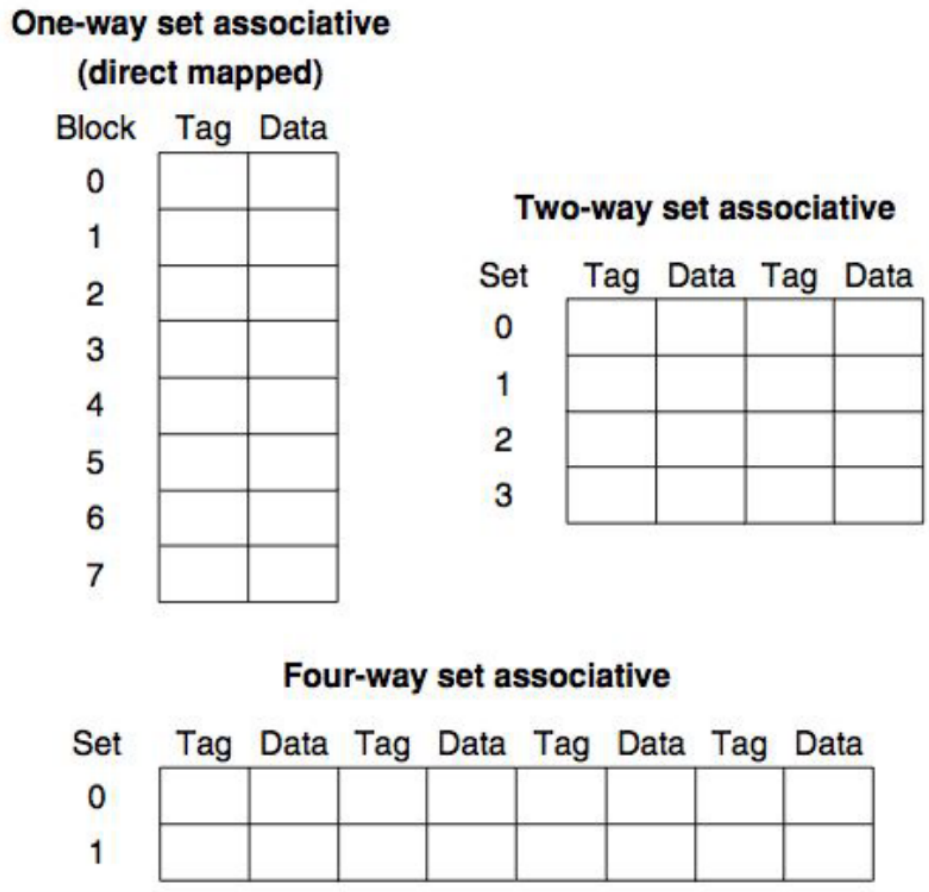

Figura: Cache e associatividade. Fonte: PH 5.3

Mapeamento associativo

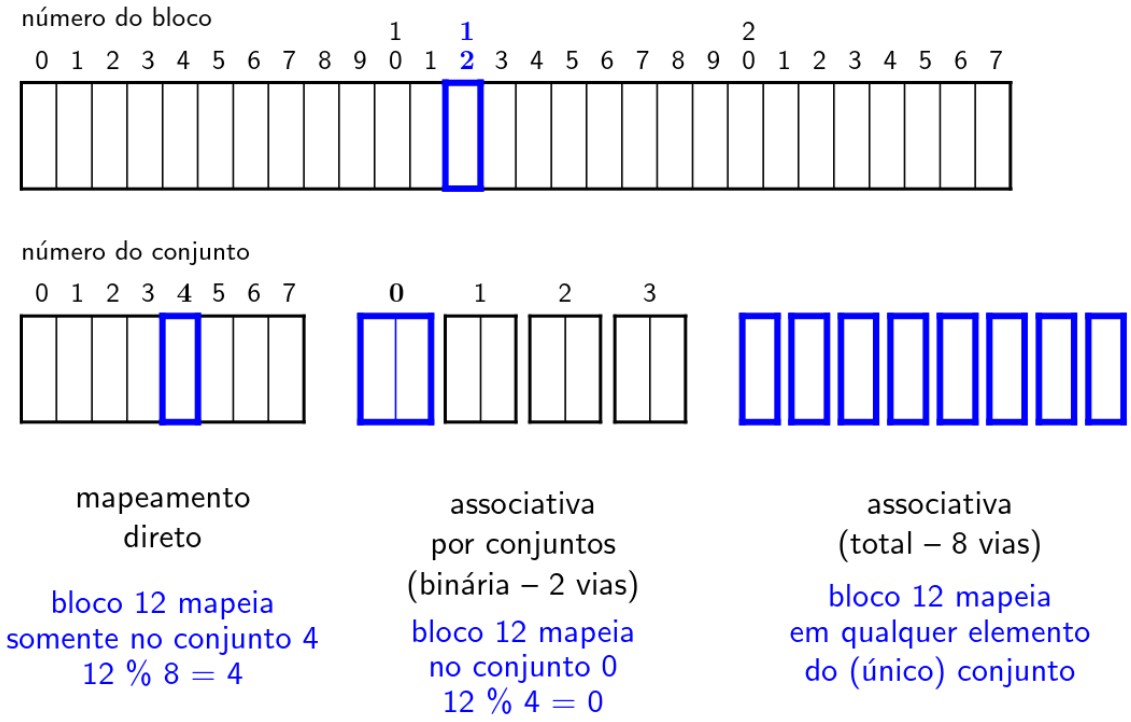

Figura: Cache e associatividade. Fonte: RH

Mapeamento associativo

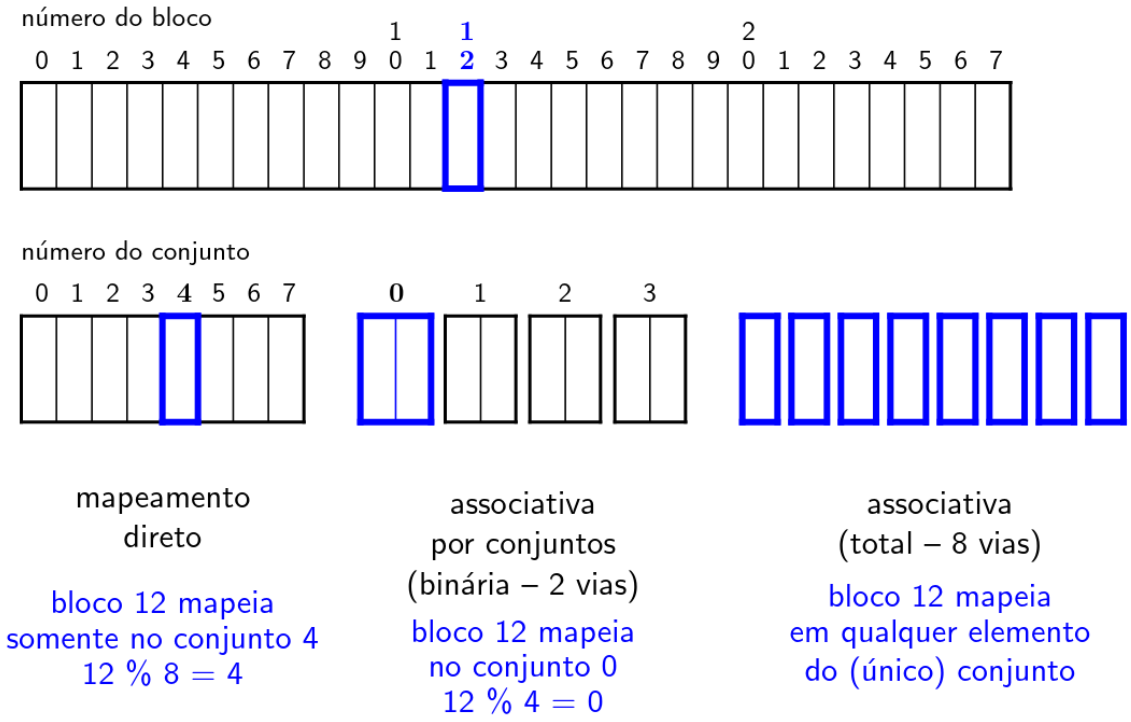

Figura: Cache e associatividade. Fonte: RH

Qual bloco remover?

INSTITUTO FEDERAL
Paraná
Campus Pinhais

Desempenho

Cada combinação de cache, processador e aplicação terá um desempenho.

Desempenho

Cada combinação de cache, processador e aplicação terá um desempenho.

Não existe solução perfeita! Os projetos de computadores pessoais geralmente chegam em uma hierarquia que funcione bem (não ótimo) para a maioria dos casos.