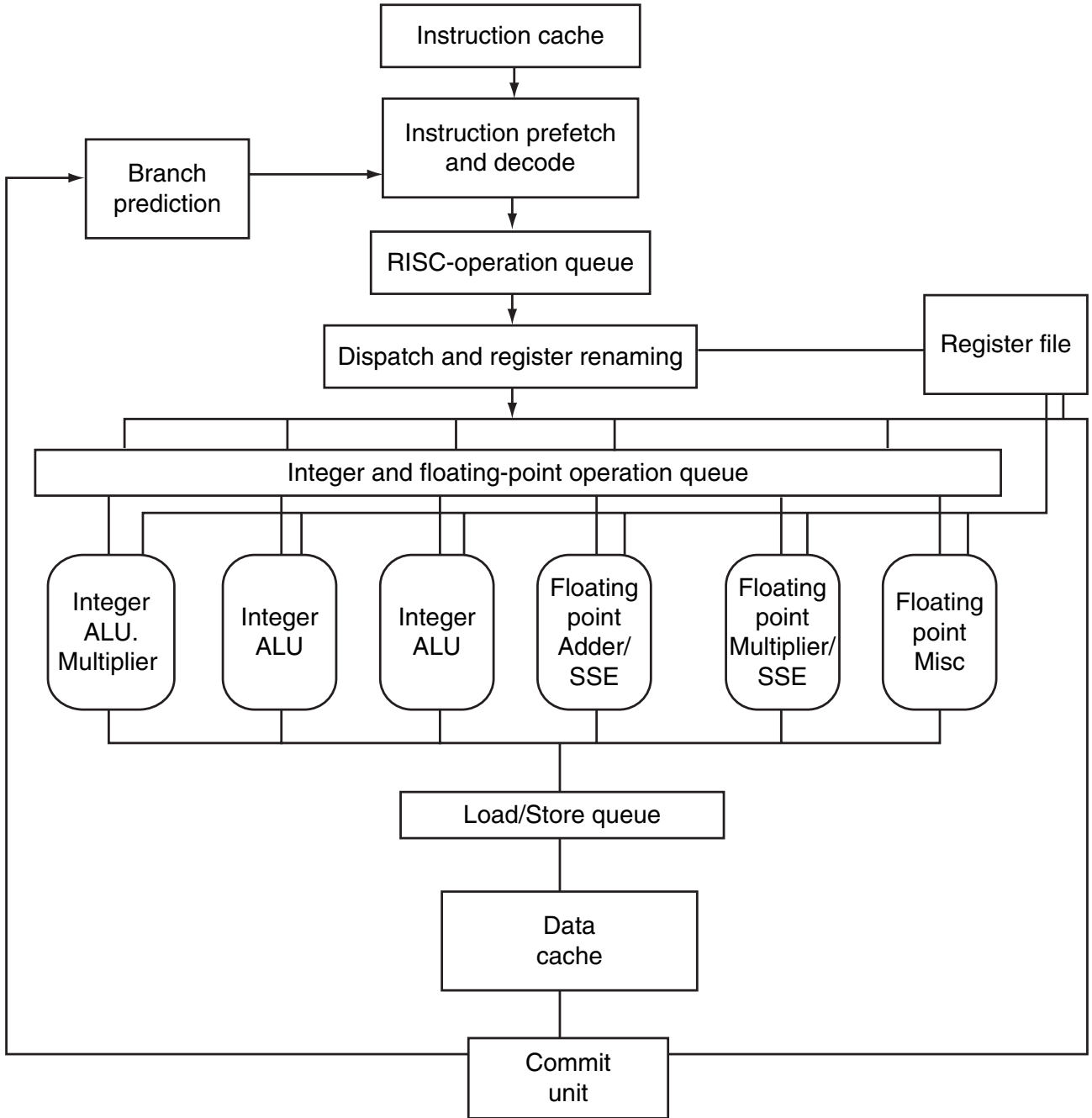
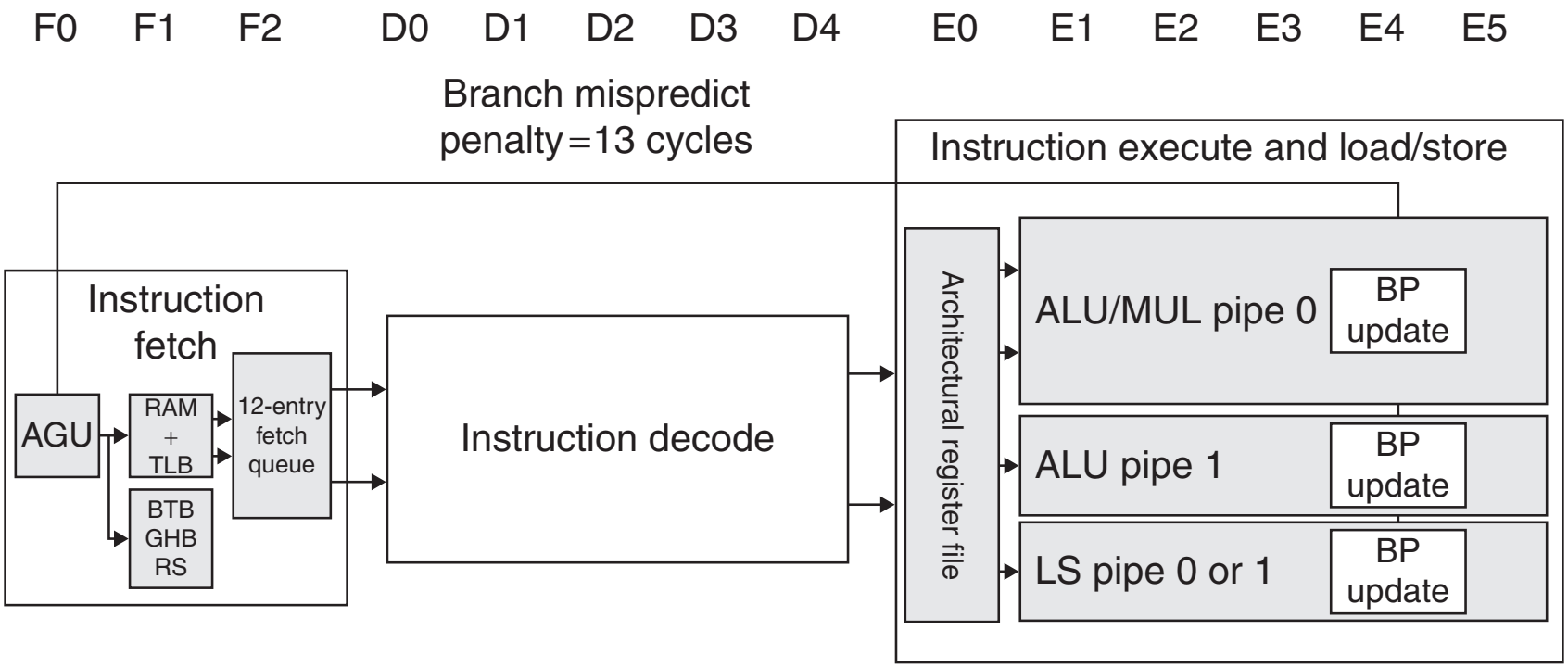
Microarquitectura AMD Opteron X4 (Barcelona)



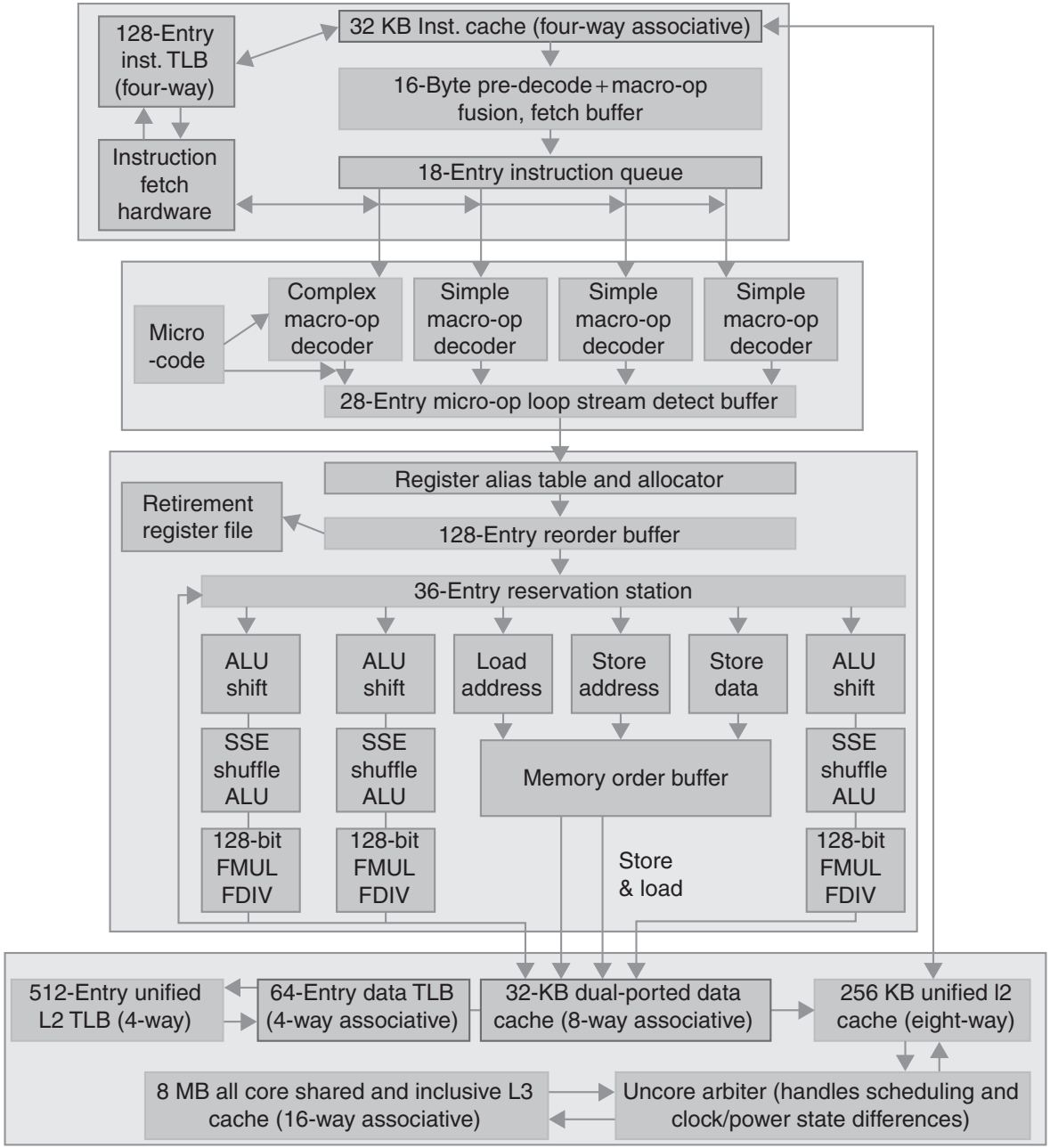
O ARM Cortex-A8 e o Intel Core i7-920 (Nehalem)

Processor	ARM A8	Intel Core i7 920
Market	Personal Mobile Device	Server, Cloud
Thermal design power	2 Watts	130 Watts
Clock rate	1 GHz	2.66 GHz
Cores/Chip	1	4
Floating point?	No	Yes
Multiple Issue?	Dynamic	Dynamic
Peak instructions/clock cycle	2	4
Pipeline Stages	14	14
Pipeline schedule	Static In-order	Dynamic Out-of-order with Speculation
Branch prediction	2-level	2-level
1st level caches / core	32 KiB I, 32 KiB D	32 KiB I, 32 KiB D
2nd level cache / core	128–1024 KiB	256 KiB
3rd level cache (shared)	–	2–8 MiB

Pipeline do ARM Cortex-A8



Pipeline do Intel Core i7-920 (Nehalem)



Bottlenecks

Factores que influenciam negativamente o aproveitamento do ILP

- Instruções que não é possível traduzir para poucas operações RISC (acontece nas arquitecturas CISC, como a x86)
- Saltos condicionais difíceis de prever, originando desperdícios de tempo por erros na especulação
- Dependências longas
- Acessos à memória

Organização da memória

Memória

Problemas

Problema 1

Latência da memória RAM típica: 50 ns

Relógio de processador com frequência de 1 GHz: T = 1 ns

Um acesso à memória leva 50 ciclos

Problema 2

A memória tem uma dimensão limitada

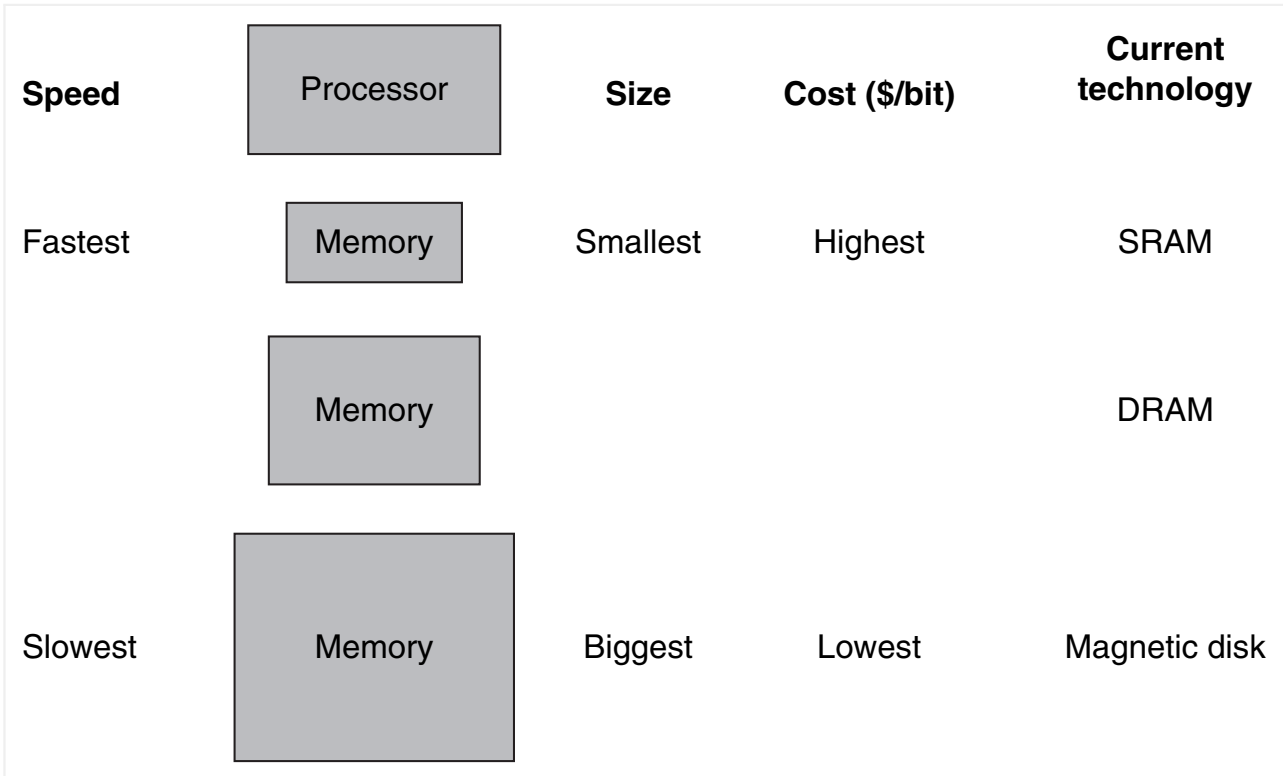
Problema 3

O computador precisa de memória

Hierarquia de memória

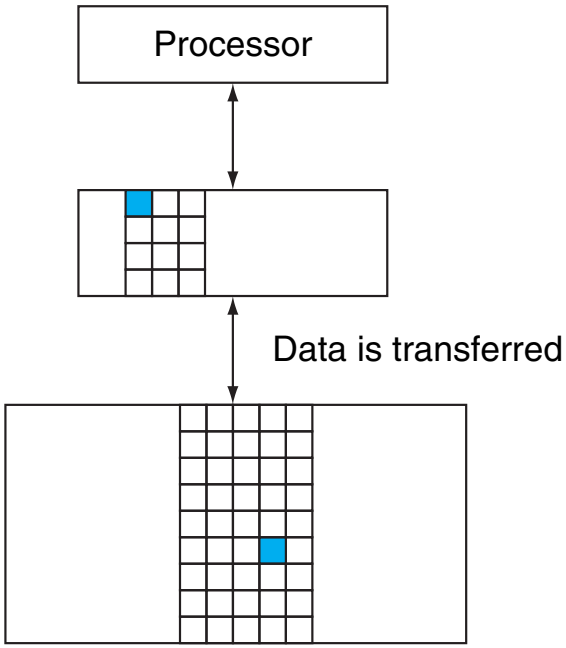
Tecnologia	Tempo de acesso	Preço/GB
SRAM	0.5 – 2.5 ns	\$500 – \$1000
DRAM	50 – 70 ns	\$10 – \$20
Flash	5 000 – 50 000 ns	\$0.40 – \$1
Disco magnético	5 000 000 – 20 000 000 ns	\$0.05 – \$0.10

Memória mais perto do processador é a mais rápida



Memória mais rápida é mais cara e, por isso, mais pequena

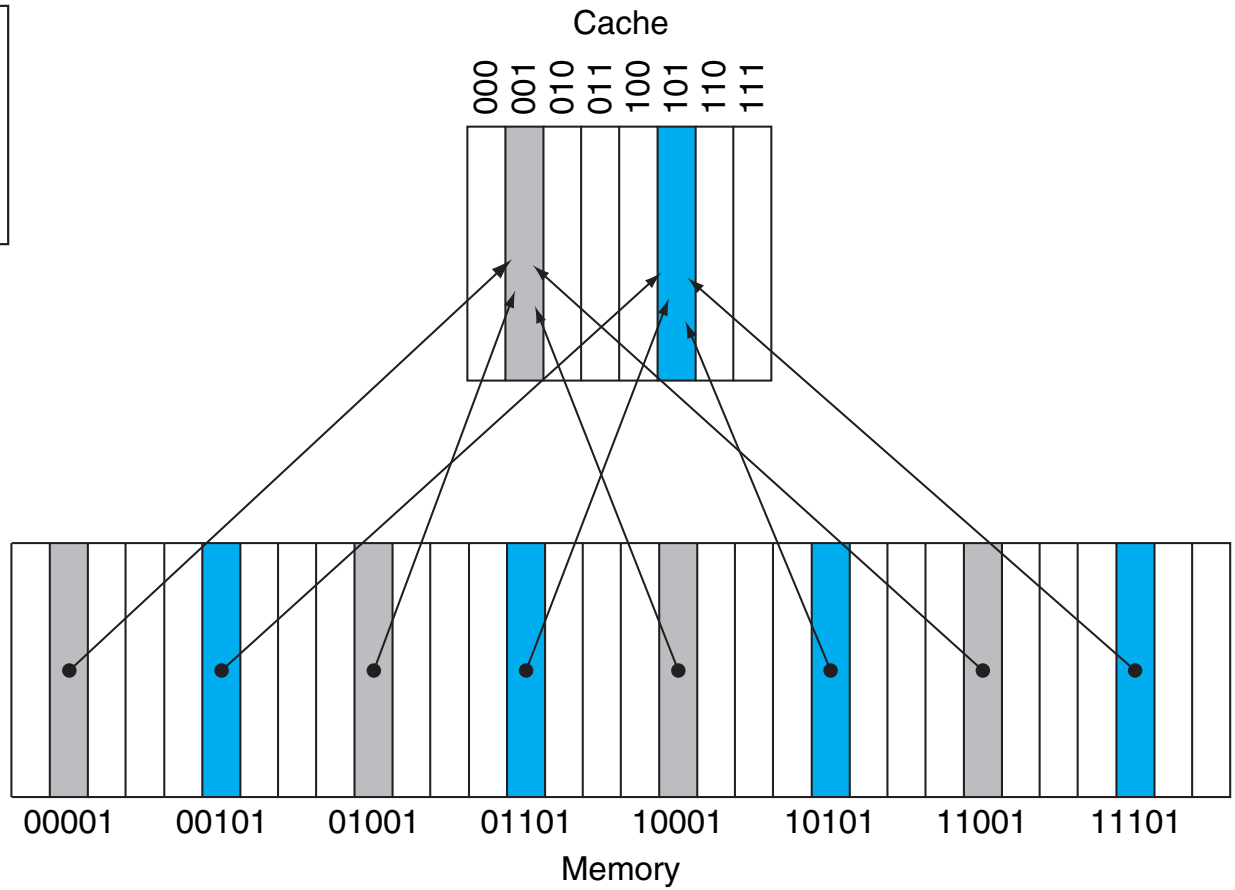
Cache



A informação é copiada para a memória do nível superior, que é mais pequena

Essa memória é uma cache

Os conteúdos de posições de memória distintas vão ter de ocupar a mesma posição (em momentos diferentes)



Hit and miss

Hit O conteúdo da posição de memória acedida está na cache

Hit time Tempo (geralmente, medido em ciclos de relógio) que demora o acesso ao conteúdo de uma posição de memória na cache

Miss O conteúdo da posição de memória acedida não está na cache

Miss penalty Tempo (também medido em ciclos) que demora transferir o conteúdo de uma posição de memória de um nível inferior da hierarquia para o nível acima (podendo substituir o seu anterior conteúdo) e tê-lo disponível para utilização

Hit rate Fracção das posições de memória acedidas cujo conteúdo foi encontrado na cache

Miss rate $= 1 -$ *hit rate*

Associatividade da cache

Para uma cache com 8 posições

One-way set associative

(direct mapped)

1 posição/conjunto
(8 conjuntos)

Block	Tag	Data
0		
1		
2		
3		
4		
5		
6		
7		

Two-way set associative

Set	Tag	Data	Tag	Data
0				
1				
2				
3				

2 posições/conjunto
(4 conjuntos)

Four-way set associative

4 posições/conjunto
(2 conjuntos)

Set	Tag	Data	Tag	Data	Tag	Data	Tag	Data
0								
1								

8 posições/conjunto
(1 conjunto)

Eight-way set associative (fully associative)

Tag	Data	Tag	Data	Tag	Data	Tag	Data	Tag	Data	Tag	Data	Tag	Data	Tag	Data

Associatividade da cache

O que significa

Direct mapped

Colocação numa posição fixa da cache

2-way set associative

Colocação em qualquer uma de um conjunto de 2 posições da cache

4-way set associative

Colocação em qualquer uma de um conjunto de 4 posições da cache

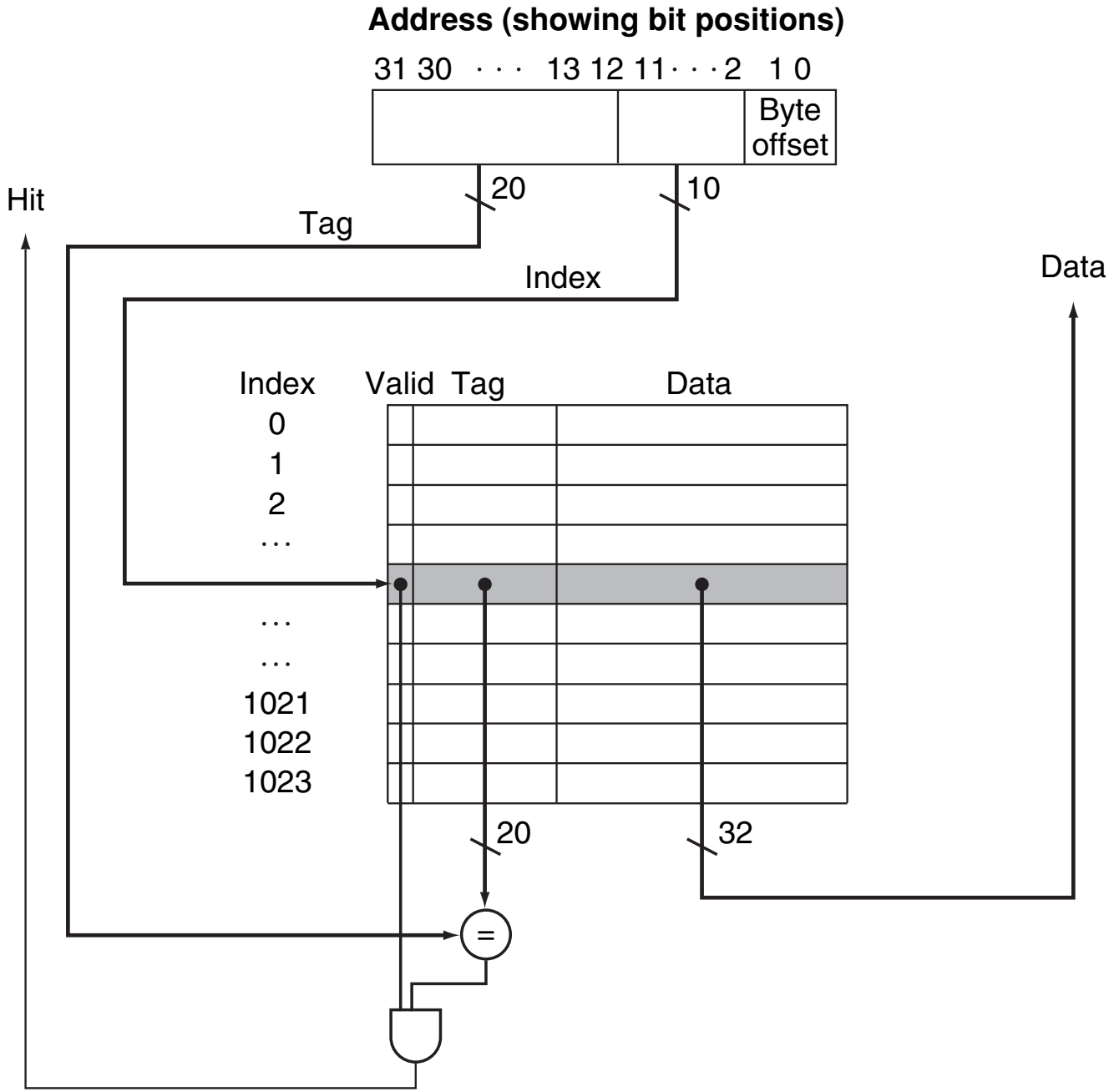
n-way set associative

Colocação em qualquer uma de um conjunto de n posições da cache

Fully associative

Colocação em qualquer posição da cache

Implementação de uma cache *direct-mapped*



Implementação de uma cache *4-way set associative*