William Stallings
Computer Organization and Architecture
8th Edition

Chapter 4

Memória Cache

Características

- Posição
- Capacidade
- Unidade de Transferência
- Método de Acesso
- Performance
- Tipo Físico
- Características Físicas
- Organização

Posição

- CPU
- Interna
- Externa

Capacidade

- Tamanho
  - Bytes ou Words

Unidade de Transferência

- Interna
  - Normalmente definida pela capacidade do barramento
- Externa
  - Normalmente, um bloco que é bem maior que um word
- Endereçamento
  - Menor posição que pode ser endereçada unicamente
  - Cluster nos S.O. M$

Métodos de Acesso (1)

- Sequencial
  - Inicia na primeira posição da memória e faz a leitura em ordem
  - Tempo de acesso depende da posição do dado
  - Ex.: Fita
- Direto
  - Blocos individuais tem enderçamento único
  - Acesso é dado pulando a patir de endereços próximos, associado a uma busca
  - Tempo de acesso depende da posição atual e da posição anterior
  - Ex.: Disco rígido (HDD, SSD)

Métodos de Acesso (2)

- Aleatório (Random)
  - Endereços individuais identificam a posição exatamente
  - Tempo de acesso é independente da posição ou de acesso anterior
  - Ex.: RAM
- Associativo
  - Dado é localizado por comparação de conteúdo em um endereço da memória
  - Tempo de acesso é independente da posição ou de acesso anterior
  - Ex.: cache

Hierarquia de Memória

- Registradores
  - CPU
- Interna or Memória Principal
  - Pode incluir um ou mais níveis de cache
  - "RAM"
- Memória Externa
  - Armazenamento persistente

Performance

- Tempo de Acesso - Latência
  - Tempo entre solicitar o dado apresentado o endereço e receber o dado
- Tempo de Ciclo de Memória
  - Um tempo de recuperação pode ser necessário até a próxima solicitação
  - Tempo de Ciclo = Tempo de Acesso + Recuperação
- Taxa de Transferência
  - Taxa na qual o dado pode ser movido

Tipos Físicos

- Semicondutor
  - RAM
  - SSD
- Magnético
  - HDD
  - Fita
  - Disco Flexível
- Ótico
  - CD
  - DVD

Características Fisicas

- Decaimento
- Volatilidade
- Capacidade de Apagar
- Consumo de Energia

Organização

- Arranjo dos bits em bytes
- Nem sempre óbvio
- Ex.: Interpolado

Escolhas

- Quanto?
  - Capacidade
- Quão Rápido?
  - Tempo é dinheiro
- Quão caro?

Bits vs Bytes

- 1 Byte ➔ 8 bits
  - 00000000, 00000001, ..., 11111111

- 1 KB ➔ 1024 Bytes (1KB = 2^{10} Bytes)
- 1 MB ➔ 1024 KB (1MB = 2^{10} KB = 2^{20} Bytes)
- 1 GB ➔ 1024 MB (1GB = 2^{10} MB = 2^{20} KB = 2^{30} Bytes)
- 1 TB ➔ 1024 GB (1TB = 2^{10} GB = 2^{20} MB = 2^{30} KB = 2^{40} Bytes)

Portanto:

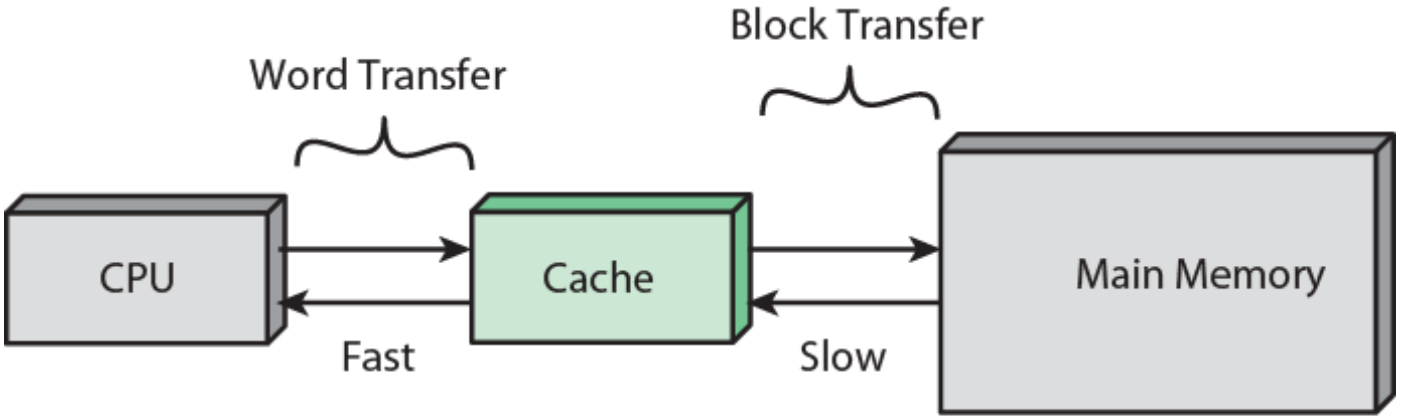
- 25 MB = ? bits
  - 25 * 1024 * 1024 * 8 bits = 209.715.100 bits

- 257.698.037.760 bits
  - 257.698.037.760 bits / 8 = 32.212.254.720 Bytes
  - 32.212.254.720 Bytes / 1024 = 21.457.280 KB
  - 21.457.280 KB KB / 1024 = 30.720 MB
  - 30.720 MB / 1024 = 20 GB

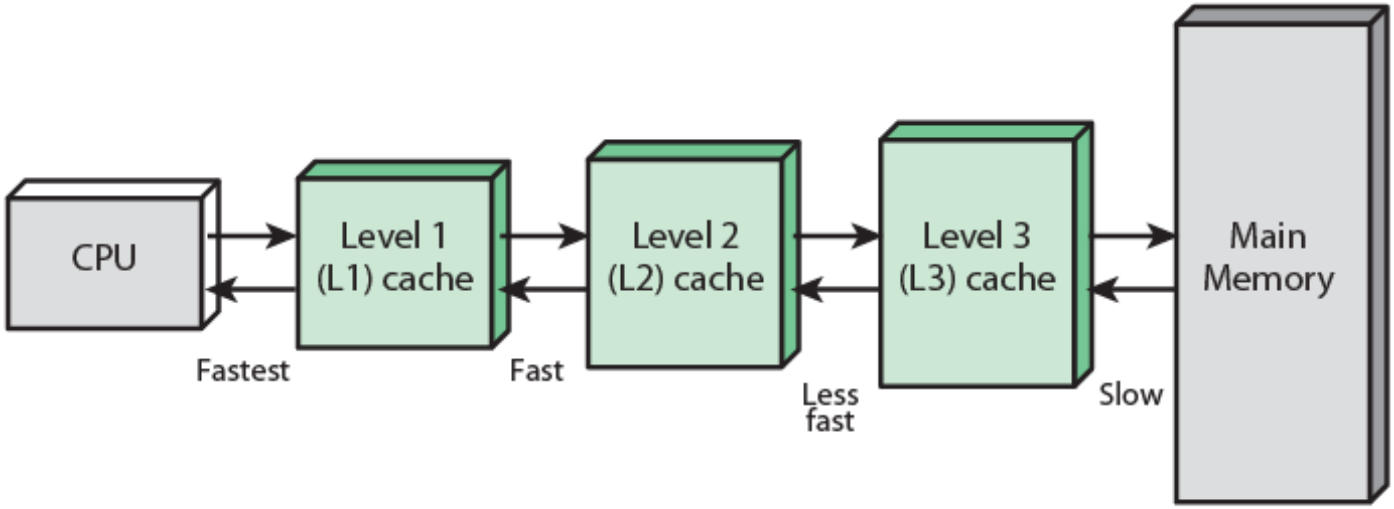
Cache

- Pequena quantidade de memória rápida
- Localizada entre a memória principal e a CPU
- Pode ser um modulo ou estar na CPU

Cache e Memória Principal



(a) Single cache



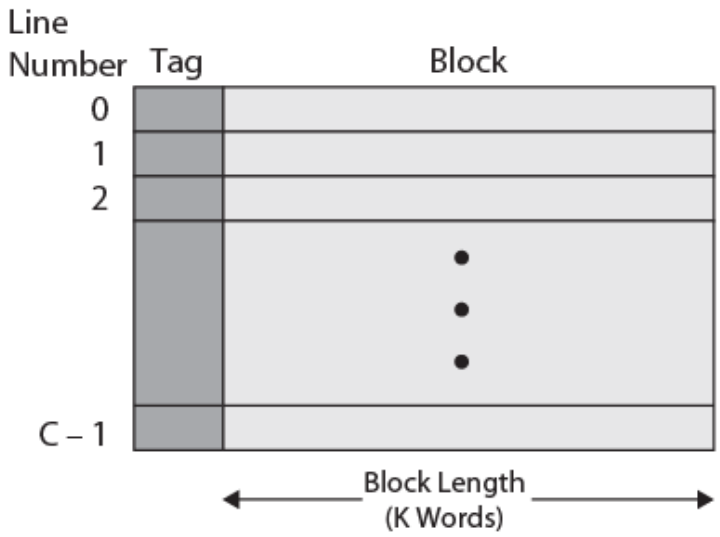
(b) Three-level cache organization

Cache

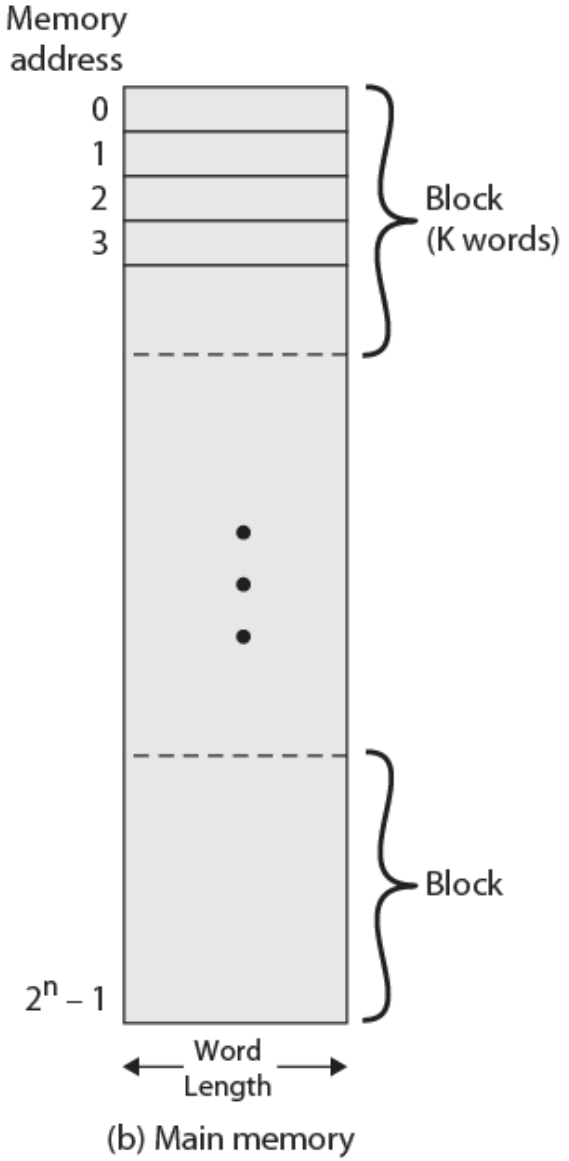

UMC
UM8881F
9642-EYT
M92201


SMART
IN332Q4AO84XG15
061995-T2
Made in USA

Estrutura de Cache/Memória Principal



(a) Cache

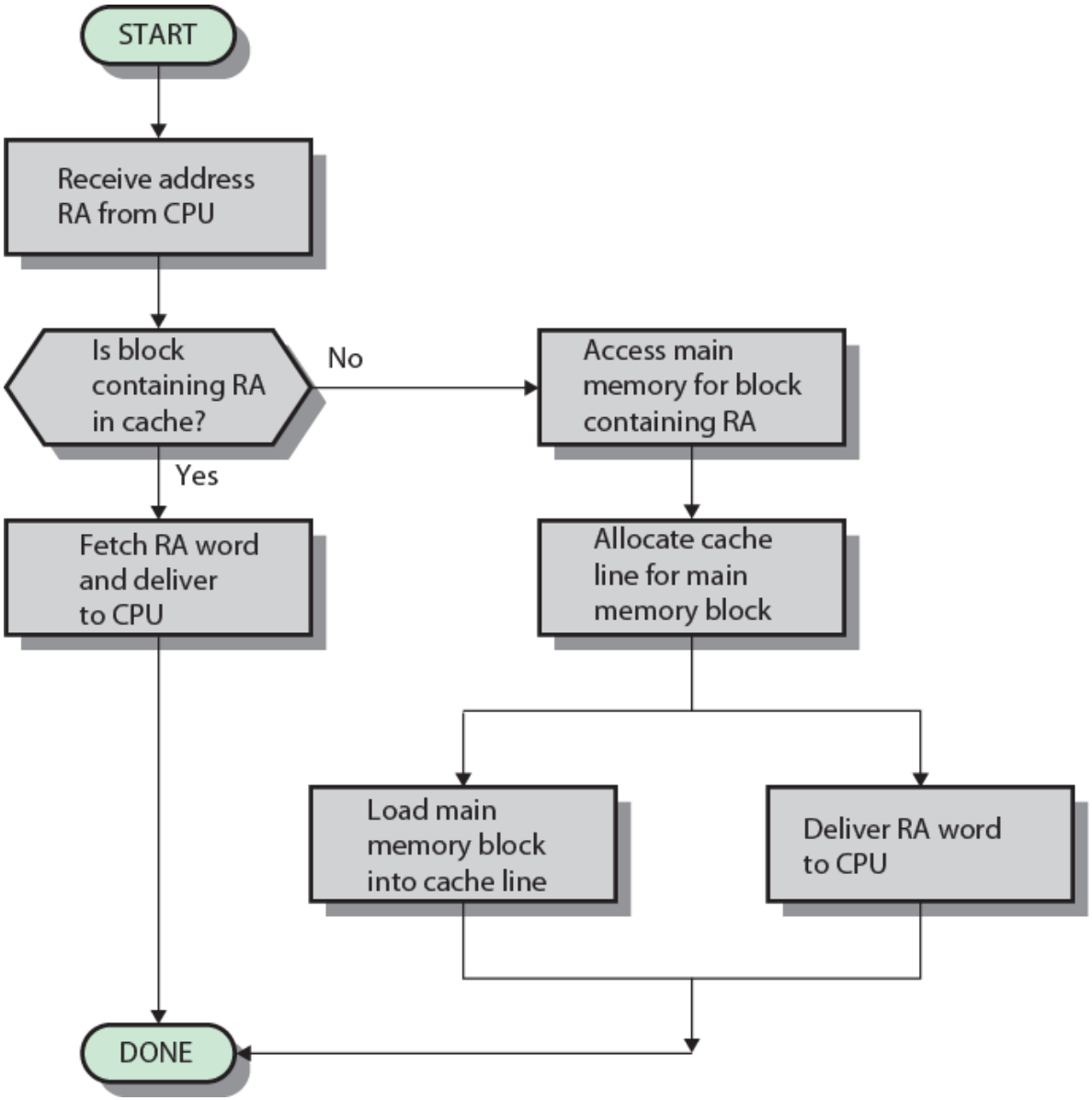


(b) Main memory

Operação de Cache

- CPU solicita conteúdo da memória
- Verifica na cache
- Se está presente, busca da cache (rápido)
- Se não está presente, faz a leitura do dado da memória principal, levando-o à cache
- Então envia o dado da cache para CPU
- Cache inclui tags que identificam qual endereço da memória principal está associado à cache

Operação da Cache - Fluxograma



Cache Design

- Endereçamento
- Tamanho
- Mapeamento
- Algoritmo de Substituição
- Políticas de Escrita
- Tamanho do Bloco
- Cache Multinível

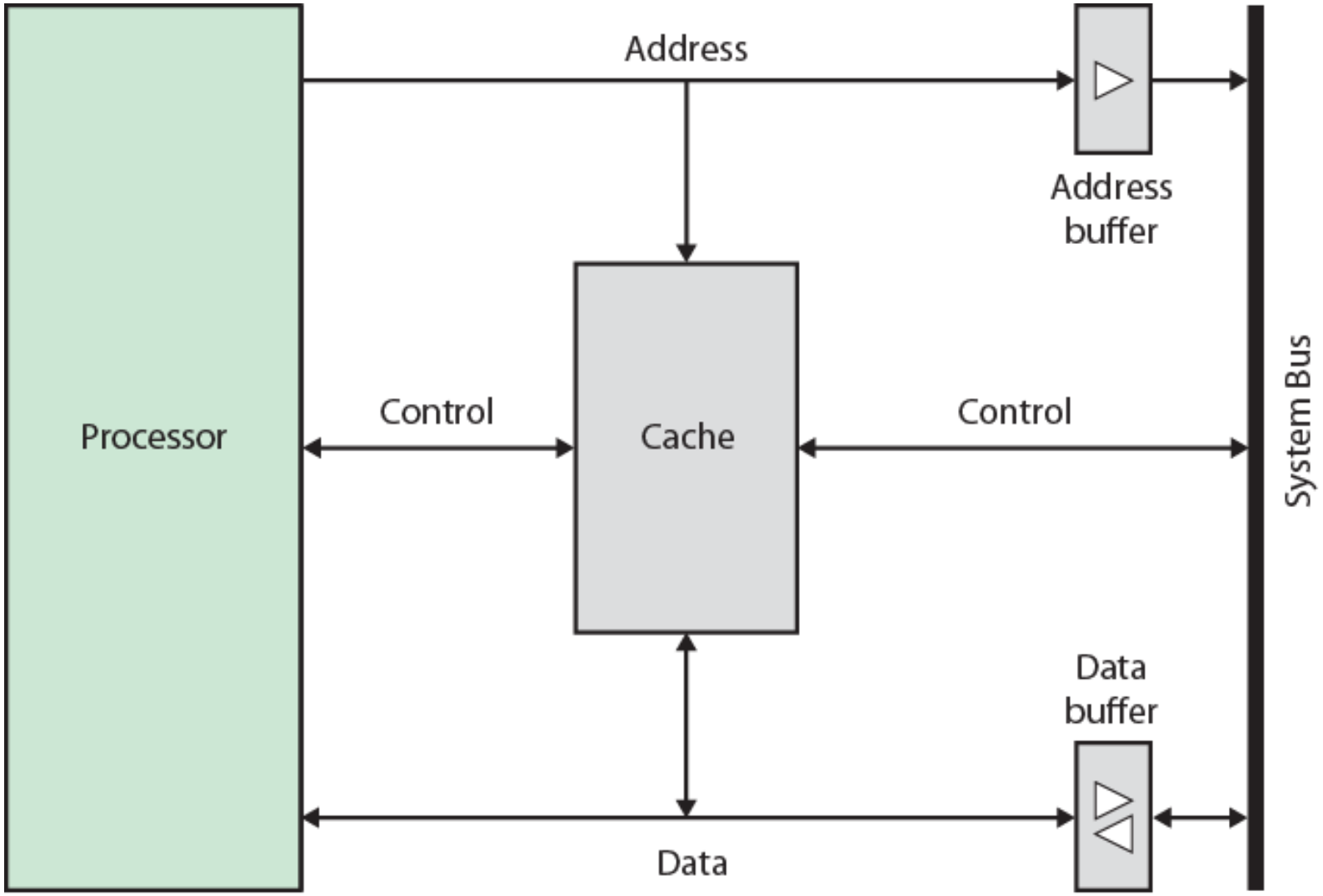
Endereçamento da Cache

- Onde fica a memória cache ?
  - Entre a unidade de processamento e a unidade de gerenciamento de memória virtual (Memory Management Unit – MMU)
  - Entre MMU e memória principal
- Cache armazena dados referenciando endereçamento da memória principal

Tamanho

- Custo
  - Muita Cache é caro
- Desempenho
  - Muita cache é rápido (até um certo ponto)
  - Buscar dados na cache demandam tempo

Organização Típica de Cache



Disponibilidade Histórica de Cache

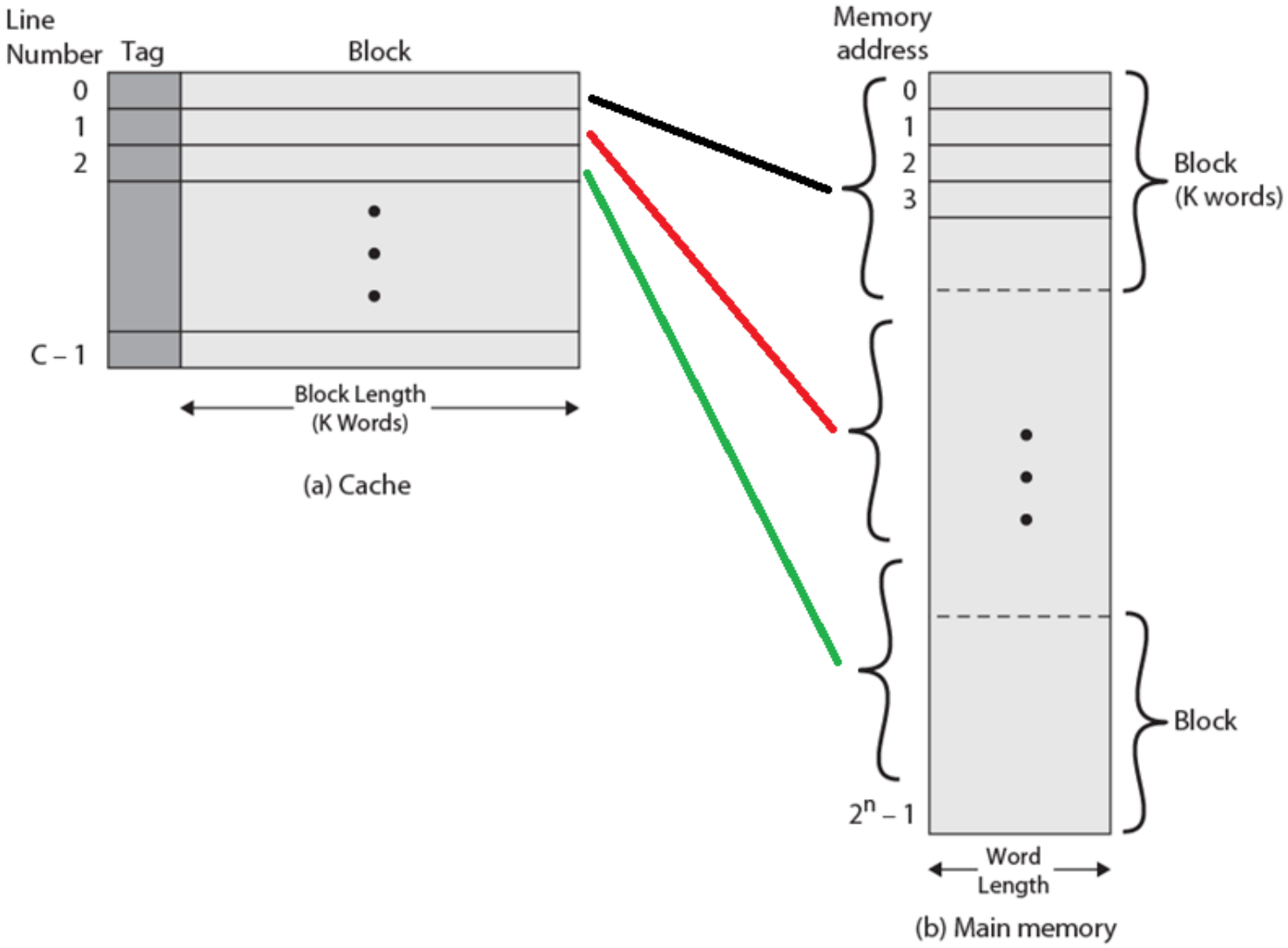
Processador	Tipo de Computador	Ano de Introdução	L1 cache	L2 cache	L3 cache
IBM 360/85	Mainframe	1968	16 to 32 KB	—	—
PDP-11/70	Minicomputer	1975	1 KB	—	—
VAX 11/780	Minicomputer	1978	16 KB	—	—
IBM 3033	Mainframe	1978	64 KB	—	—
IBM 3090	Mainframe	1985	128 to 256 KB	—	—
Intel 80486	PC	1989	8 KB	—	—
Pentium	PC	1993	8 KB/8 KB	256 to 512 KB	—
PowerPC 601	PC	1993	32 KB	—	—
PowerPC 620	PC	1996	32 KB/32 KB	—	—
PowerPC G4	PC/server	1999	32 KB/32 KB	256 KB to 1 MB	2 MB
IBM S/390 G4	Mainframe	1997	32 KB	256 KB	2 MB
IBM S/390 G6	Mainframe	1999	256 KB	8 MB	—
Pentium 4	PC/server	2000	8 KB/8 KB	256 KB	—
IBM SP	High-end server/ supercomputer	2000	64 KB/32 KB	8 MB	—
CRAY MTA[b]	Supercomputer	2000	8 KB	2 MB	—
Itanium	PC/server	2001	16 KB/16 KB	96 KB	4 MB
SGI Origin 2001	High-end server	2001	32 KB/32 KB	4 MB	—
Itanium 2	PC/server	2002	32 KB	256 KB	6 MB
IBM POWER5	High-end server	2003	64 KB	1.9 MB	36 MB
CRAY XD-1	Supercomputer	2004	64 KB/64 KB	1MB	—

Mapeamento Direto

- Um bloco da memória principal é carregado em uma linha específica da cache
- Uma linha é examinada para a busca do dados
- Busca na Cache torna-se rápida
- Pode haver subutilização da Cache

Mapeamento Direto da Cache para a Memória Principal

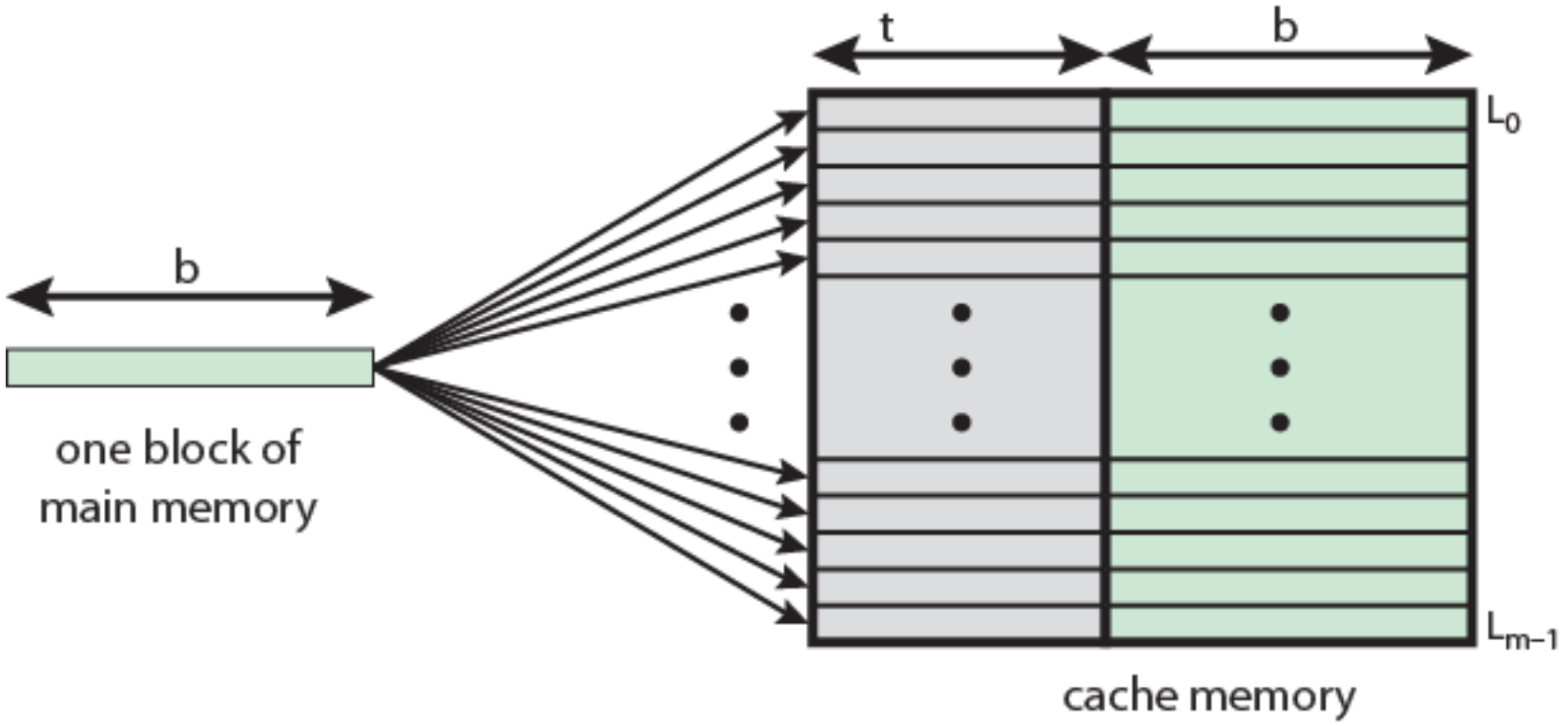
Direct



Mapeamento Associativo

- Um bloco da memória principal pode ser carregado em qualquer linha da cache
- Tag apenas identifica o bloco da memória
- Todas as linhas da Cache são examinadas na busca
- A busca se torna de baixo desempenho
- A utilização da Cache é total

Mapeamento Associativo da Cache para a Memória Principal

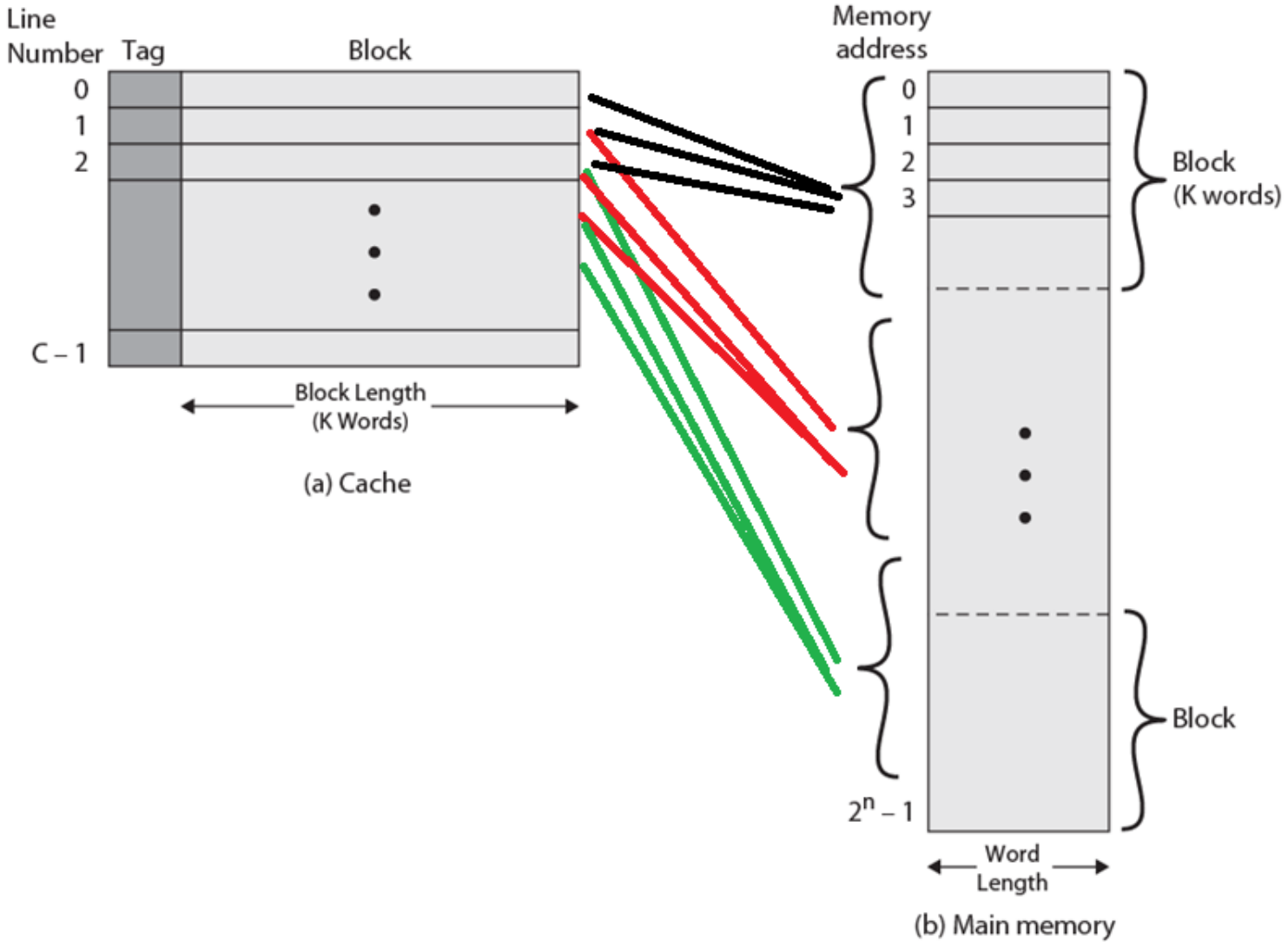


Mapeamento Set Associative

- Um bloco da memória principal pode ser carregado em algumas linhas da cache
- Não é o melhor desempenho possível
- Não é a melhor utilização possível
- Utilização otimizada da cache

Mapeamento Set Associative da Cache para a Memória Principal

Set Associative



Algoritmos de Substituição (1)
Mapeamento Direto

- Sem escolha
- Cada bloco só estará em uma linha específica
- Substitui-se na linha

Algoritmos de Substituição (2) Mapeamento Associativo e Set Associative

- Algoritmo implementado diretamente no hardware
- Random – Aleatório
- First in first out (FIFO) – Primeiro que entra é o primeiro que sai
- Least Recently used (LRU) – Menos Recentemente Usado
  - Substitui o bloco presente na cache há mais tempo
- Least frequently used (LFU) – Menos frequentemente usado
  - Substitui o bloco com menos requisições

Políticas de Escrita

- Ao atualizar o dado na cache, este deve ser atualizado, imediatamente, na memória principal
- Multicore podem ter caches independentes
- E/S podem acessar a memória diretamente

Caches Multinível

- Podem estar todos na CPU ou alguns níveis na CPU e outros Fora
  - L1 no chip, L2 fora
  - L1 e L2 no chip, L3 fora
  - L1, L2 e L3 no chip

Cache unificado versus compartilhado

- Um nível de cache para dados e instruções (ou um apenas para dados e um apenas para instruções)
- Vantagens do cache unificado
  - Taxa de acertos mais alta
    - Balanceamento de carga entre busca de dados e instruções
    - Apenas um cache para projetar e implementar
- Vantagens do cache compartilhado
  - Melhor balanceamento de espaço livre
  - Uso otimizado em cada núecleo de processamento

Cache Compartilhado

- Intel Smart Cache

Pentium 4 Cache

- 80386 – sem cache no chip
- 80486 – 8k set associative
- Pentium (todas as versões) – L1 cache dividido
  - Dados e Instruções
- Pentium III – L3 cache possível fora da CPU
- Pentium 4
  - L1 cache
    - 8k bytes
    - set associative
  - L2 cache
    - Alimentando ambos L1 caches
    - set associative
  - L3 cache on chip

Pentium 4 Diagrama