Microprocessor & Computer Architecture (µpCA)

UE19CS252

Dr. D. C. Kiran

Department of
Computer Science and Engineering

Microprocessor & Computer Architecture (μpCA)

Unit 3: Memory

Dr. D. C. Kiran

Department of Computer Science and Engineering

Microprocessor & Computer Architecture (µpCA)

Syllabus

~~Unit 1: Basic Processor Architecture and Design~~

~~Unit 2: Pipelined Processor and Design~~

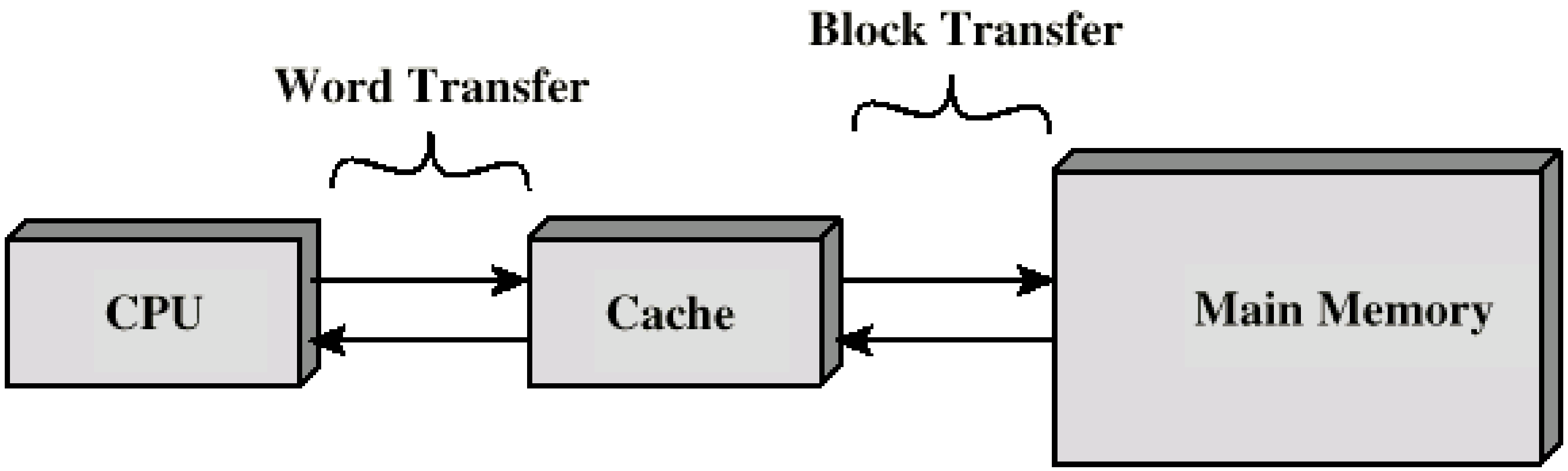
Unit 3: Memory

Unit 4: Input/Output Device Design

Unit 5: Advanced Architecture

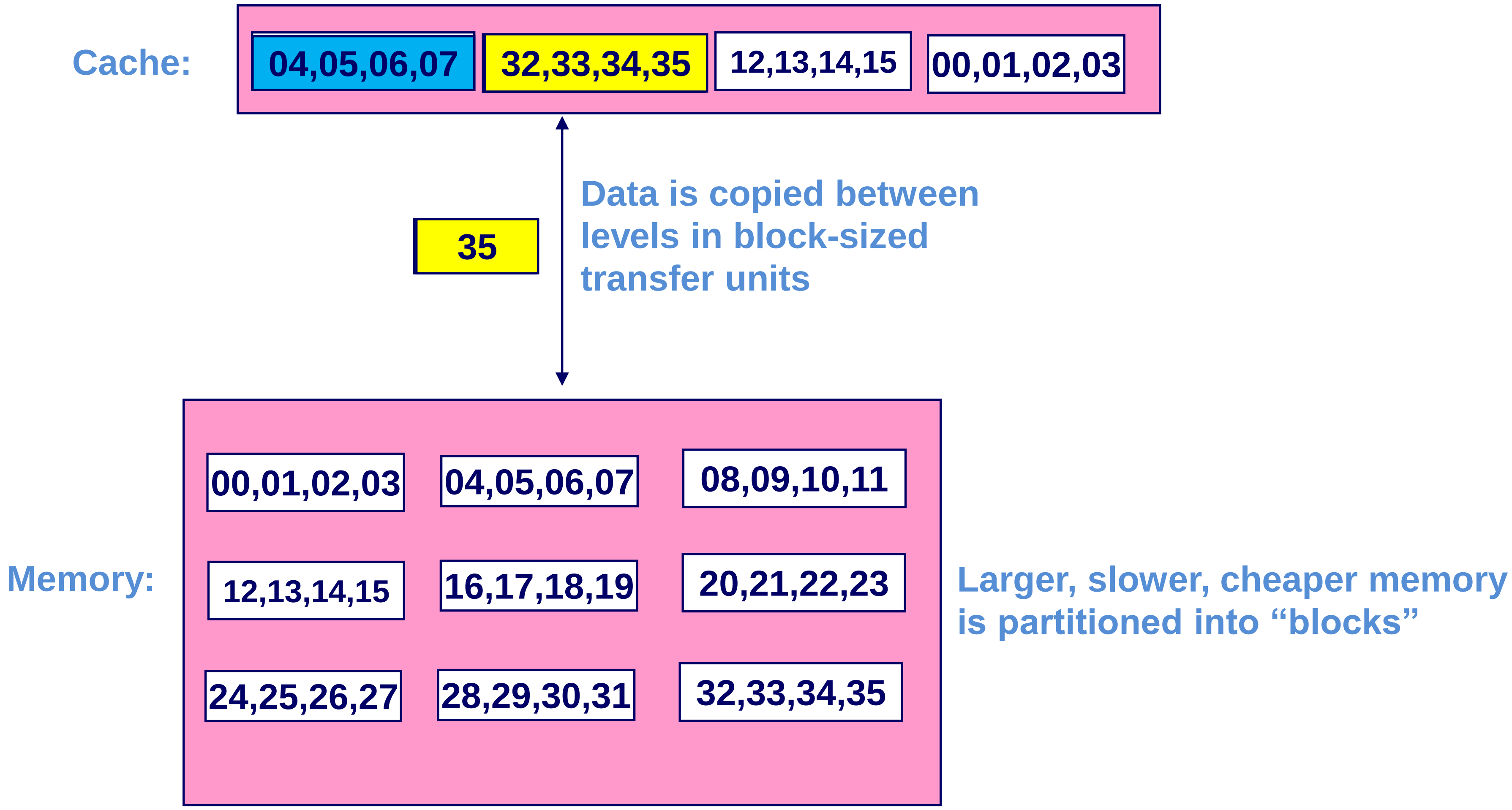
Cache

- Cache memory is an architectural arrangement which makes the main memory appear faster to the processor than it really is.



Microprocessor & Computer Architecture (μpCA)

General Cache Requirement



Data is copied between levels in block-sized transfer units

Larger, slower, cheaper memory is partitioned into “blocks”

Why Block Transfer?

Fast memory technology is more expensive per bit than slower memory

Solution: **90/10 rule** comes from empirical observation:
"A program spends 90% of its time in 10% of its code"

i.e The data or code accessed recently, may get accessed soon

Locality of Reference or Principle of Locality

Microprocessor & Computer Architecture (µpCA)

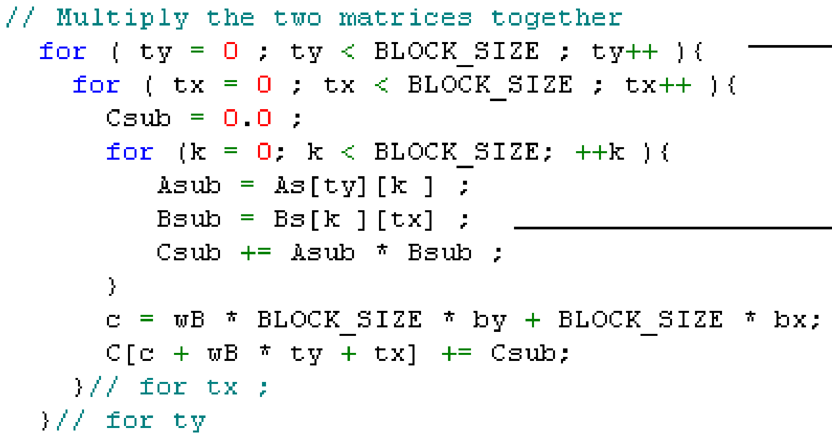
Locality of Reference

Temporal Locality:

if an item is referenced, it will tend to be referenced again soon.

Spatial Locality:

if an item is referenced, items whose addresses are close by will tend to be referenced soon.



```
// Multiply the two matrices together
  for ( ty = 0 ; ty < BLOCK_SIZE ; ty++ ){
    for ( tx = 0 ; tx < BLOCK_SIZE ; tx++ ){
      Csub = 0.0 ;
      for (k = 0; k < BLOCK_SIZE; ++k ){
         Asub = As[ty][k ] ;
         Bsub = Bs[k ][tx] ;
         Csub += Asub * Bsub ;
      }
      c = wB * BLOCK_SIZE * by + BLOCK_SIZE * bx;
      C[c + wB * ty + tx] += Csub;
    }// for tx ;
  }// for ty
```

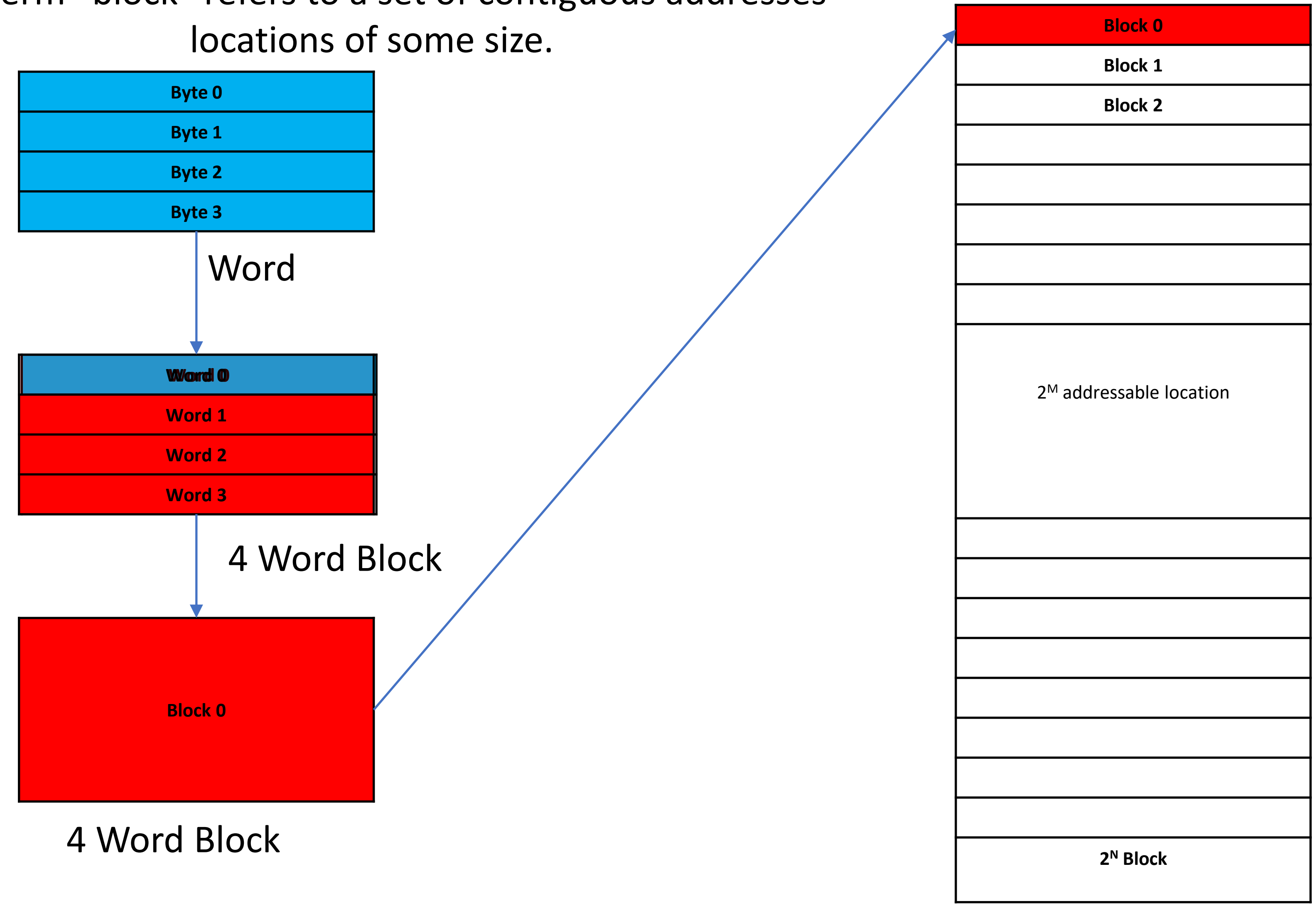
for-loop is temporal locality

array is spatial locality

Both properties hold for data and instructions

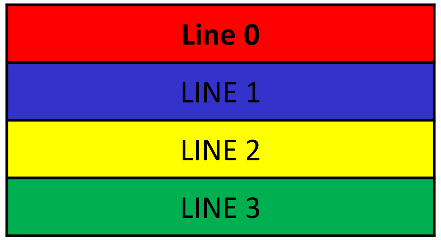
Block?

The term “block” refers to a set of contiguous addresses locations of some size.

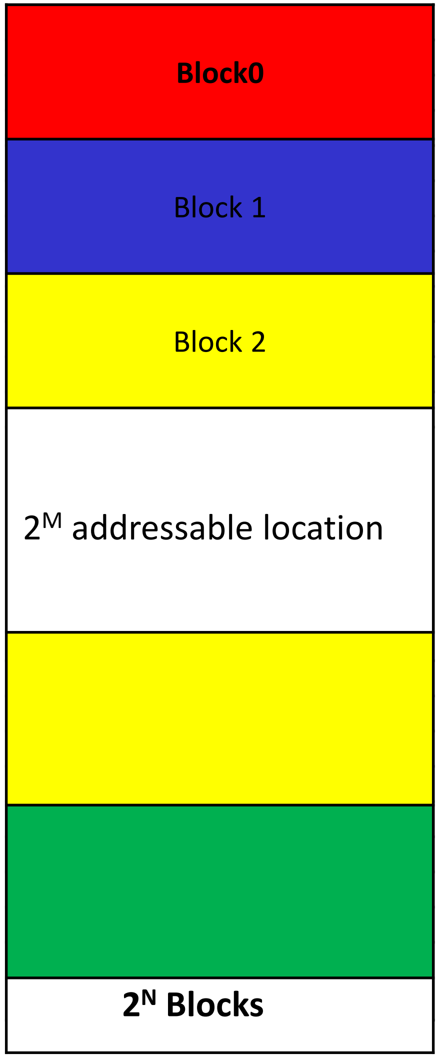


Block vs Line?

The term “block” refers to a set of contiguous addresses locations of some size.

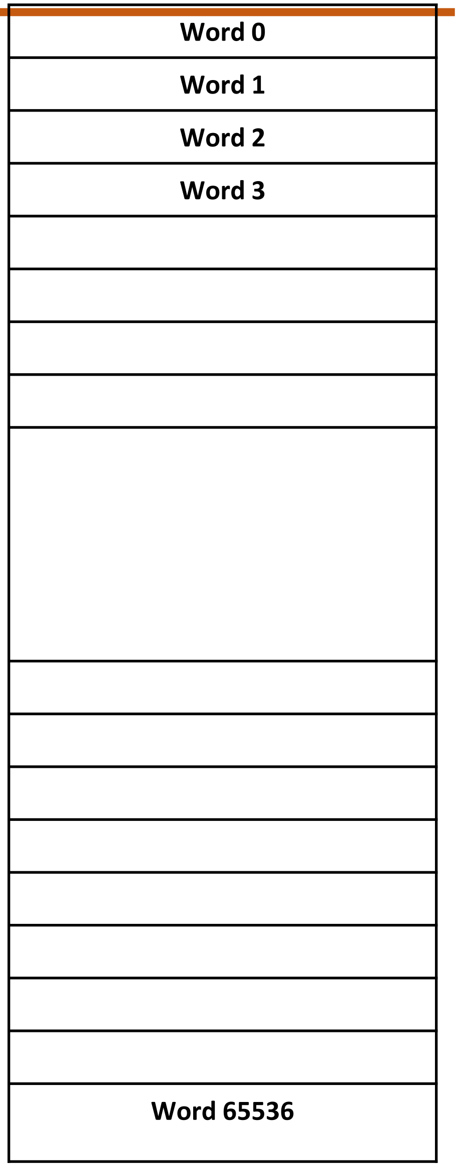


2^K Line



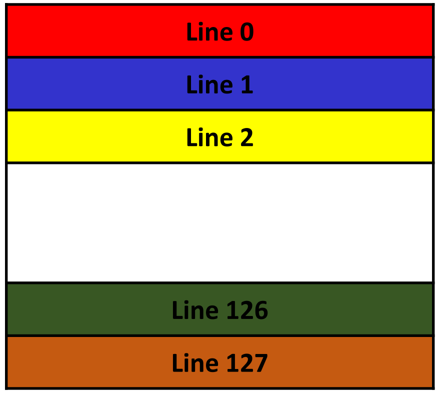
Block vs Line?

- **A simple processor example:**
  - Main memory is addressable by a 16-bit address.
  - Main memory has 65536 (64 k)words.
  - Main memory has 4096 (4 K) Blocks of 16 words each.
  - Consecutive addresses refer to consecutive words.
  - Cache consisting of 128 Lines of 16 words each.
  - Total size of cache is 2048 (2 K) words.

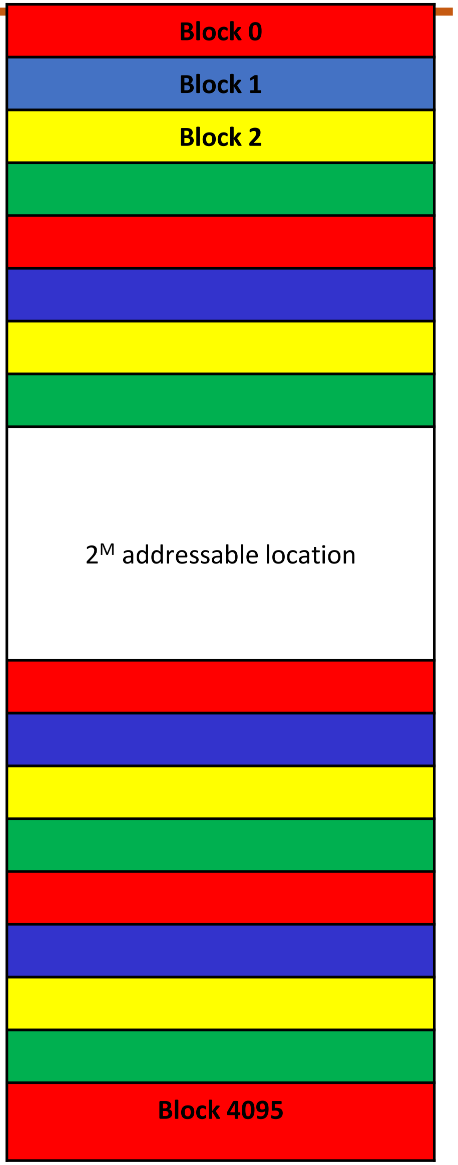


Block vs Line?

- **A simple processor example:**
  - Main memory is addressable by a 16-bit address.
  - Main memory has 65536 (64 k) words.
  - Main memory has 4096 (4 K) Blocks of 16 words each.
  - Consecutive addresses refer to consecutive words.
  - Cache consisting of 128 Lines of 16 words each.
  - Total size of cache is 2048 (2K) words.
  - 4096/128= 32 possible Blocks for one Line



128 Block Cache



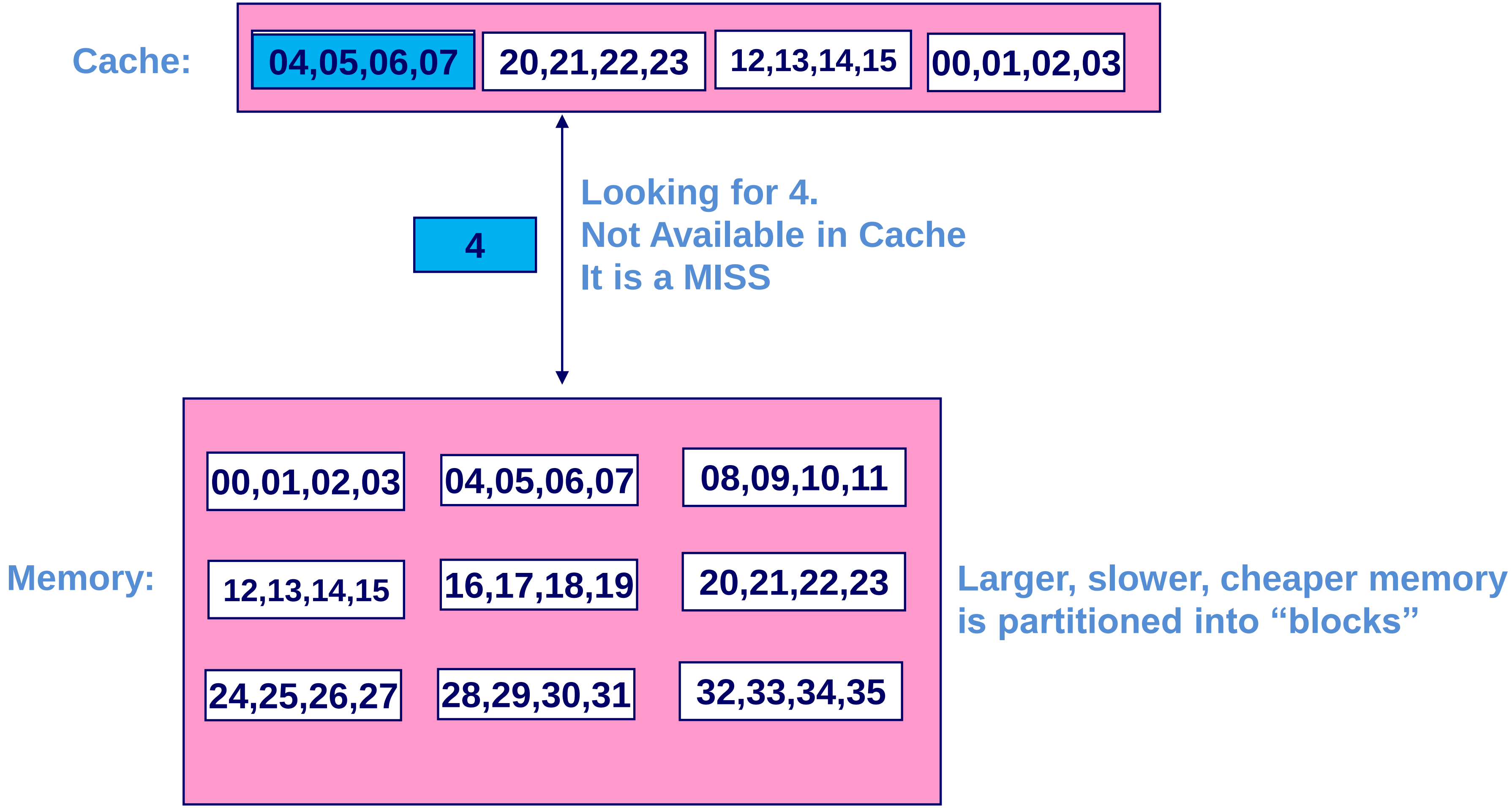
To Note and Few Terminology

- Cache Lines vs Blocks
- Block Size
- Line Size
- # Cache lines <<< # Blocks
- Block Transfer

- Cache Hit
  - Read Hit
  - Write Hit
- Cache Miss
  - Read Miss
  - Write Miss

Microprocessor & Computer Architecture (µpCA)

Rear or Write MISS



Cache:

04,05,06,07	20,21,22,23	12,13,14,15	00,01,02,03

4

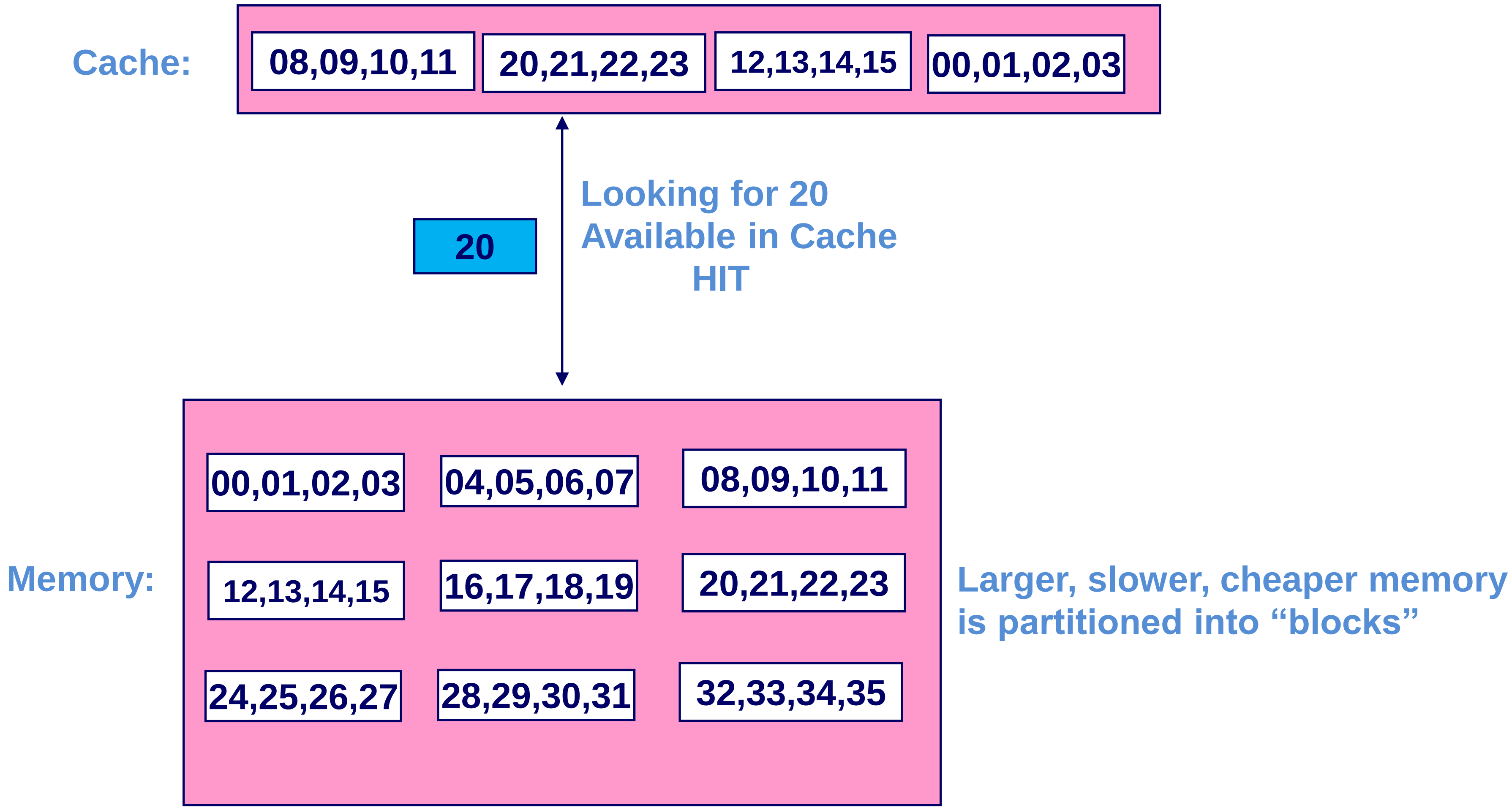
Looking for 4.
Not Available in Cache
It is a MISS

Memory:

00,01,02,03	04,05,06,07	08,09,10,11
12,13,14,15	16,17,18,19	20,21,22,23
24,25,26,27	28,29,30,31	32,33,34,35

Larger, slower, cheaper memory
is partitioned into “blocks”

Rear or Write HIT



Cache:

08,09,10,11 | 20,21,22,23 | 12,13,14,15 | 00,01,02,03

20

Looking for 20
Available in Cache
HIT

Memory:

00,01,02,03 | 04,05,06,07 | 08,09,10,11

12,13,14,15 | 16,17,18,19 | 20,21,22,23

24,25,26,27 | 28,29,30,31 | 32,33,34,35

Larger, slower, cheaper memory is partitioned into “blocks”

Four Questions in Cache Design

Cache Design is controlled by Four Questions:

Q1: Where can a block be placed in the cache?

- Block Placement

Q2: How is a block found if it is in the cache?

- Block Identification.

Q3: Which block should be replaced on a miss?

- Block Replacement.

Q4: What happens on a write ?

- Write Strategy.

Next Session

Cache Design Principles

Next Session

Cache Design Principles

PES UNIVERSITY ONLINE

THANK YOU

Dr. D. C. Kiran

Department of Computer Science and Engineering

dckiran@pes.edu

9829935135