
Master of Computer Applications

CAPOL403R01: Computer Organization & Architecture

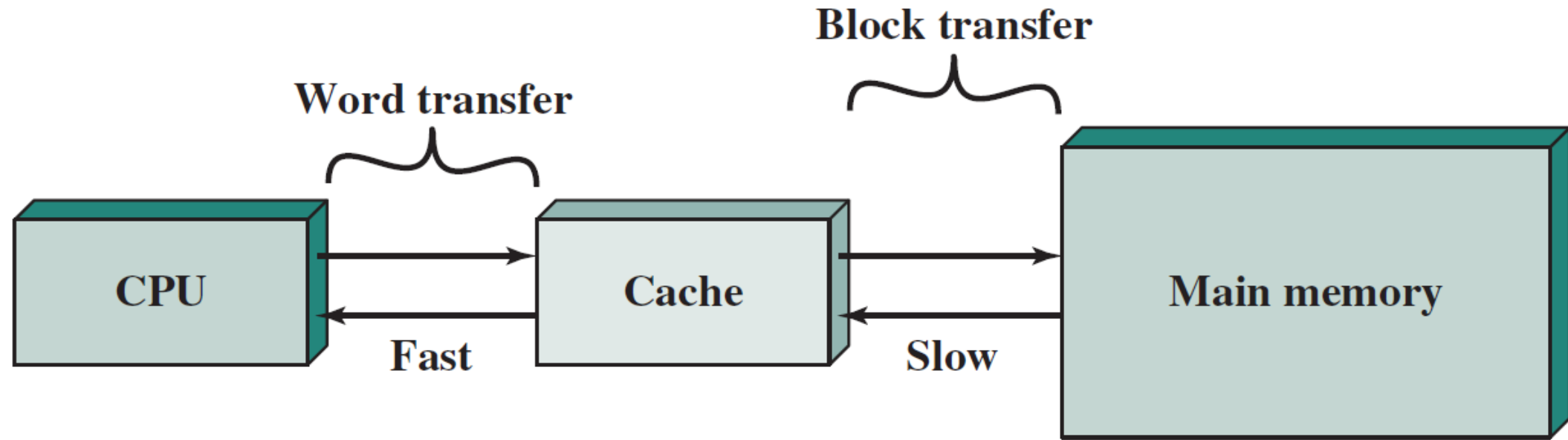
Unit III: Lecture 2
Cache Memory Principles

Dr. D. MURALIDHARAN
School of Computing
SASTRA Deemed to be University

Reference material

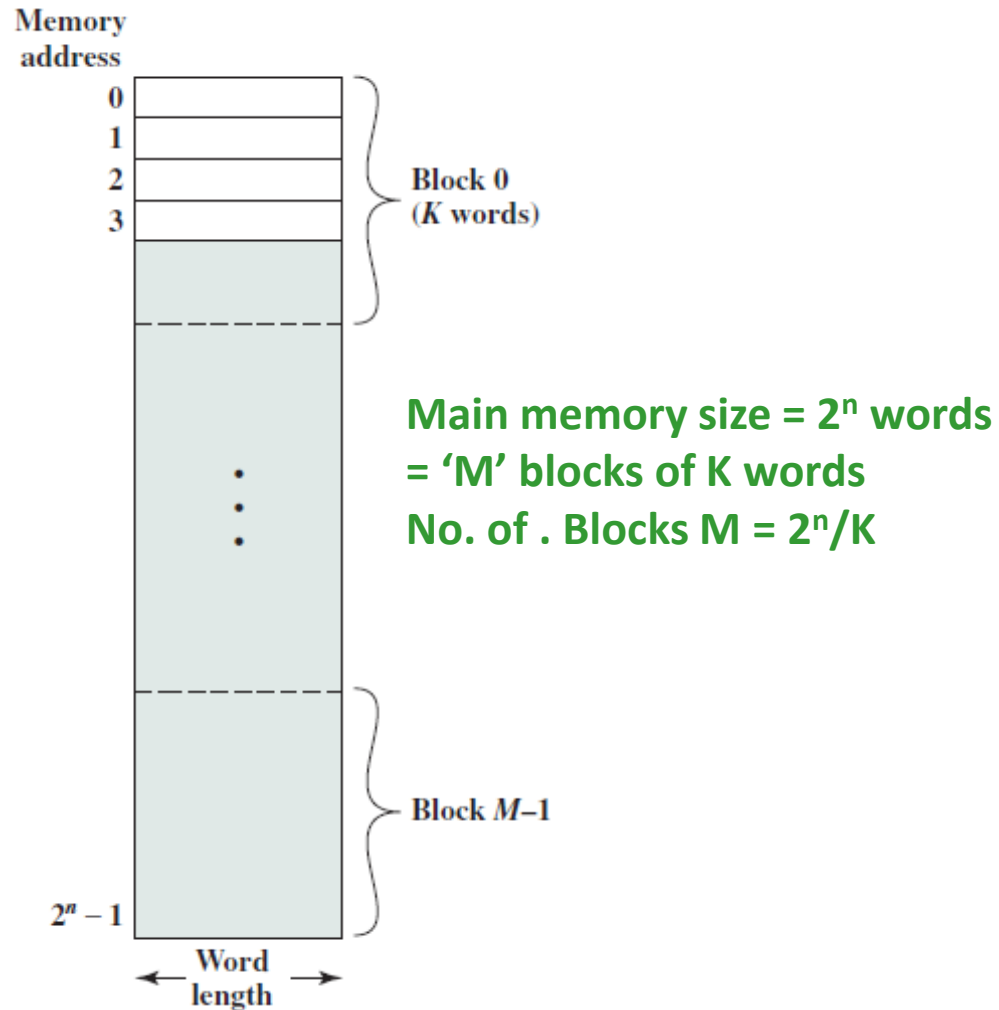
- Computer Organization and architecture – Designing for performance – Tenth Edition

Cache

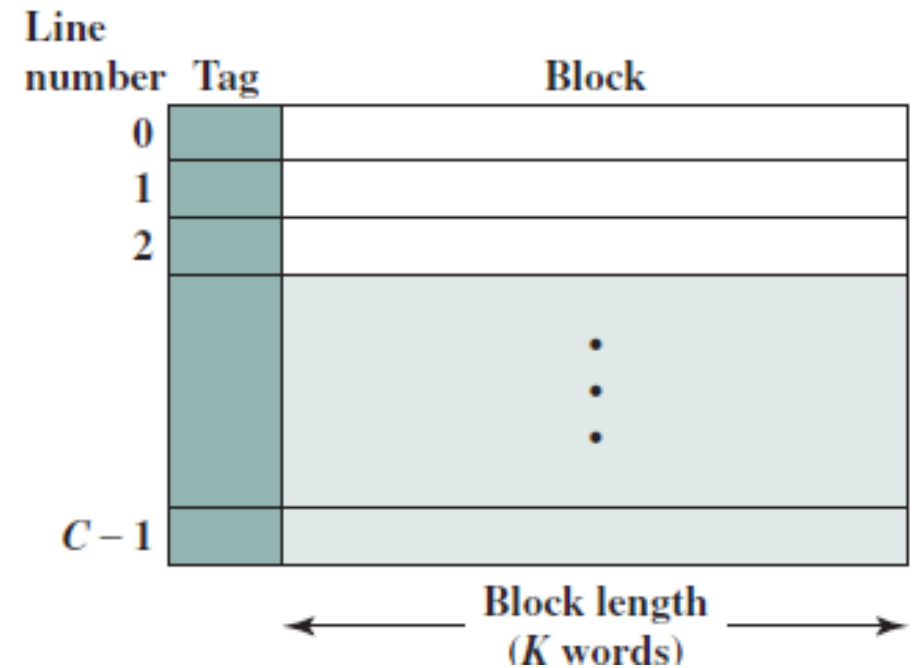


Cache...

Main memory



Cache memory



Cache size = $C * K$ words

$C \ll M$

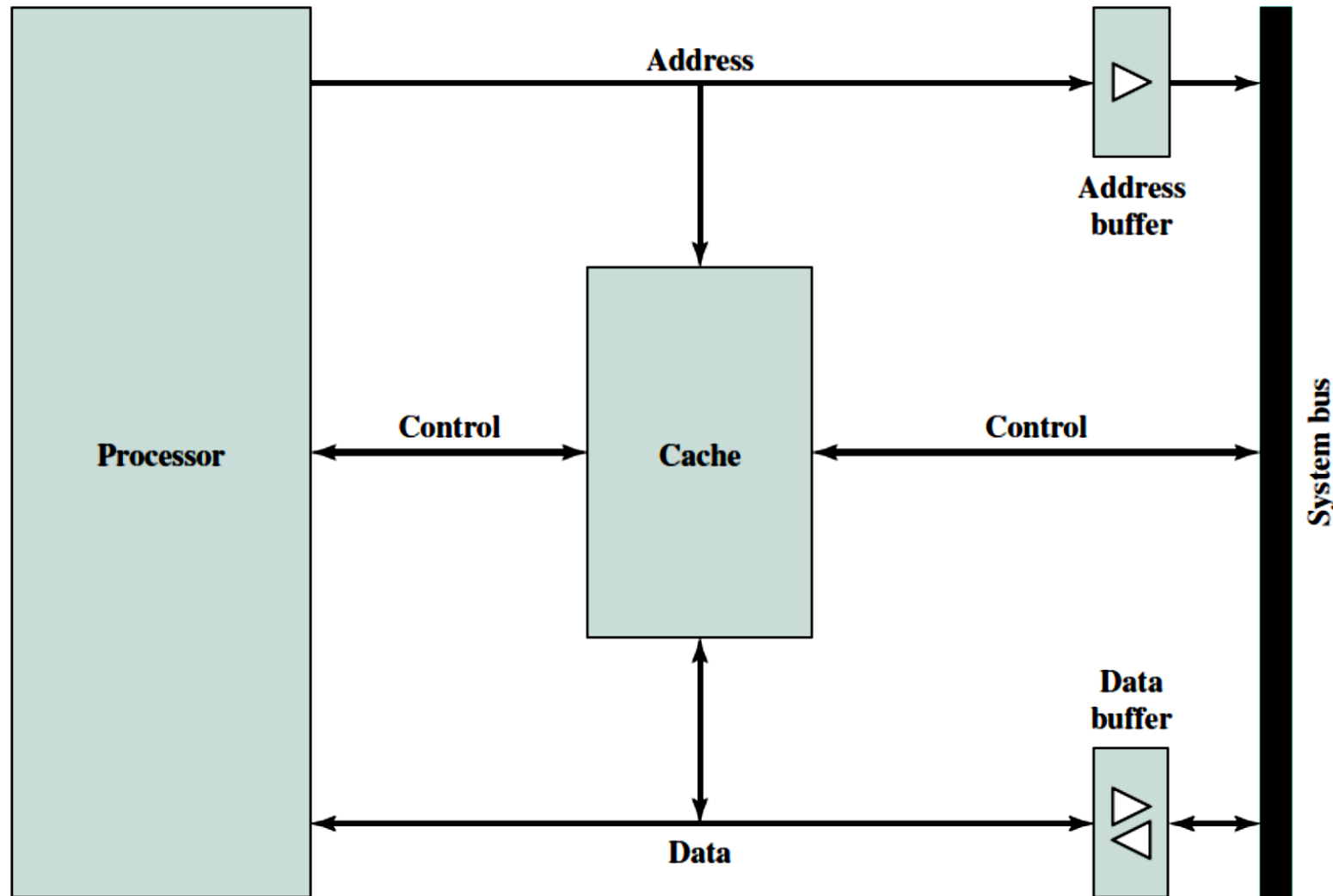
Cache...

- Main memory is divided into blocks
 - A block has 2^b words for some 'b' value
- Cache has 'C' lines
 - Cache size is ' $C \cdot 2^b$ ' words
- The cache contains copy of portions of main memory
 - It has C blocks of main memory
- When the processor attempts for a memory read, it first checks the cache
 - If it is in cache, the data will be accessed very quickly
 - If it is not, the block which has the data is moved to cache and it will be read
 - Data transfer between main memory and cache is in terms of blocks
 - Performance of processor $T = T_c + (1-H) \cdot T_m$
 - Where T_c is the cache access time, H is the hit ratio and T_m is the main memory access time

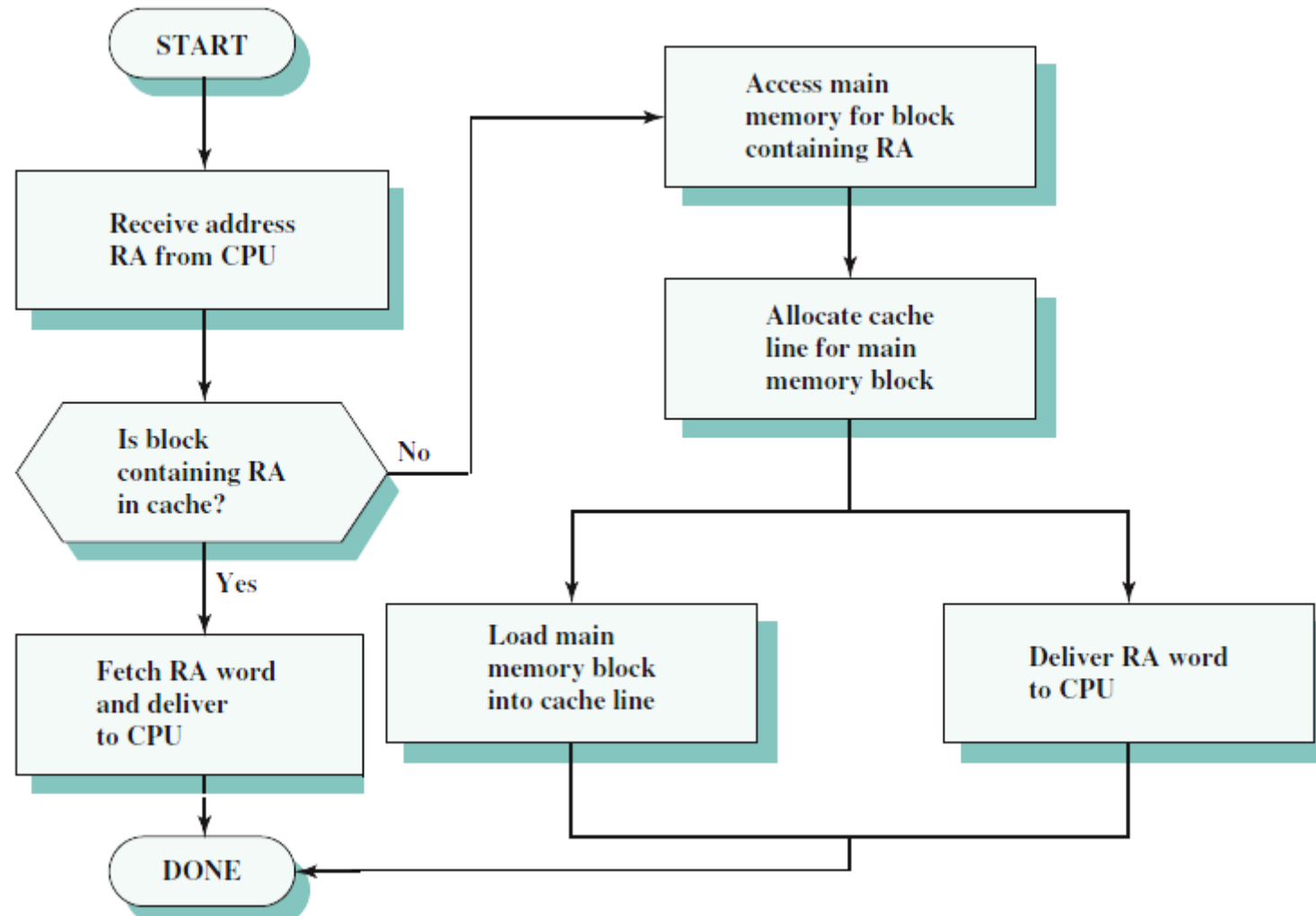
Cache...

- Number of cache lines are smaller than the number of block numbers
 - $C \ll M$
- Hence, more than one blocks are mapped on a single line
- To identify which block is in the cache line, tag bits are used.
 - Tag bits are added to each line
- When cache address is modified, the updated word may not be on main memory
 - Control bits are added to each line

Typical Cache Organization



Read operation



Elements of Cache Design

Cache Size: Do we need a big cache or a small one?

Line Size: How big it could be?

**Replacement algorithms:
Which line has to be replaced
when cache is full?**

**Number of Caches:
One or two?
Unified or Split?**

**Writing Policy:
When to update the main
memory?**

**Mapping Functions:
How to decide the cache line for a
particular block?**

Thank you