# COMPUTER ORGANIZATION (IS F242)

## LECT 36: CACHE MEMORY

# Cache Memory

- Motivation
- Cache Organization
- Cache Design
- Mapping Schemes
  - Direct Mapped Cache
  - Set Associative Cache
  - Fully Associative Cache

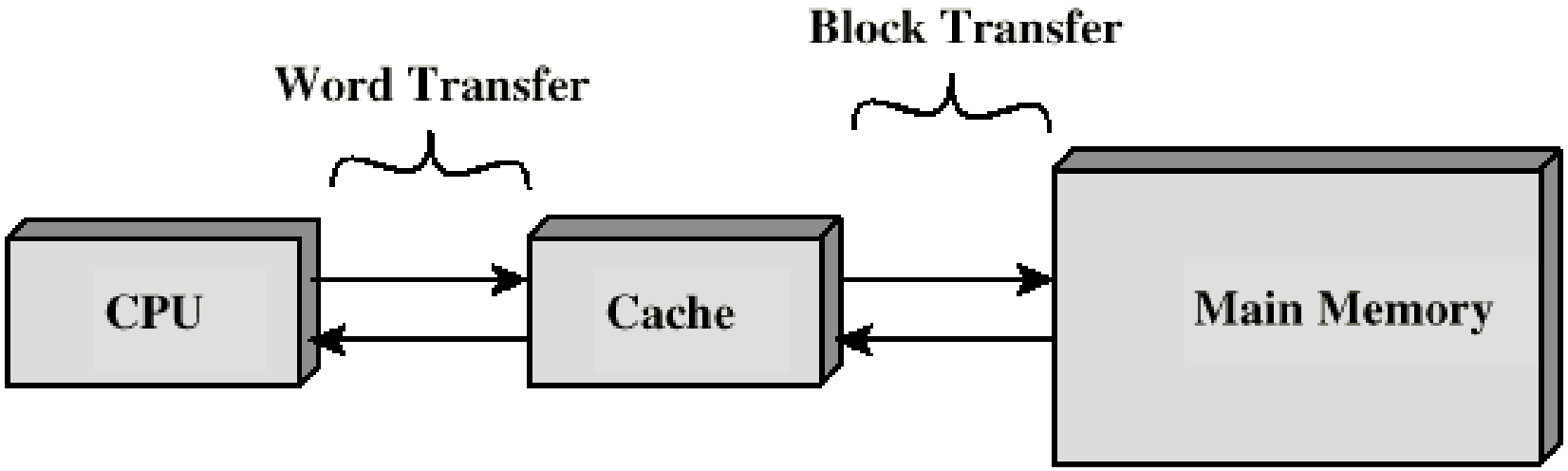
# Motivation

- Motivation
  - Large (cheap) memories (DRAM) are slow
  - Small (costly) memories (SRAM) are fast

- Make the average access time small
  - service most accesses from a small, fast memory
  - reduce the bandwidth required of the large memory

# Cache Memory

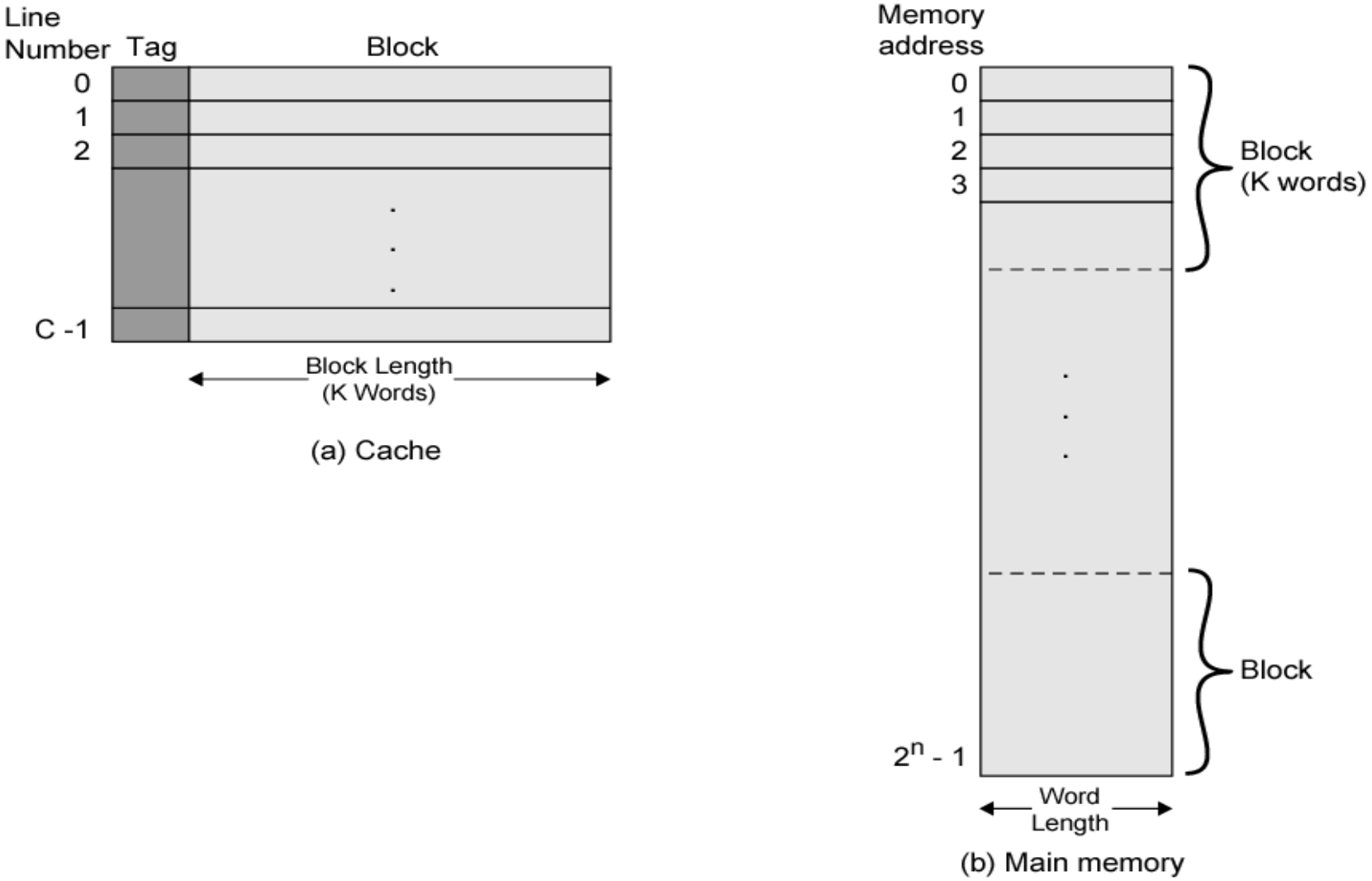
- Small amount of fast memory (SRAM)
- Sits between normal main memory and CPU
- May be located on CPU chip or module
- Transparent to OS



# Why Cache

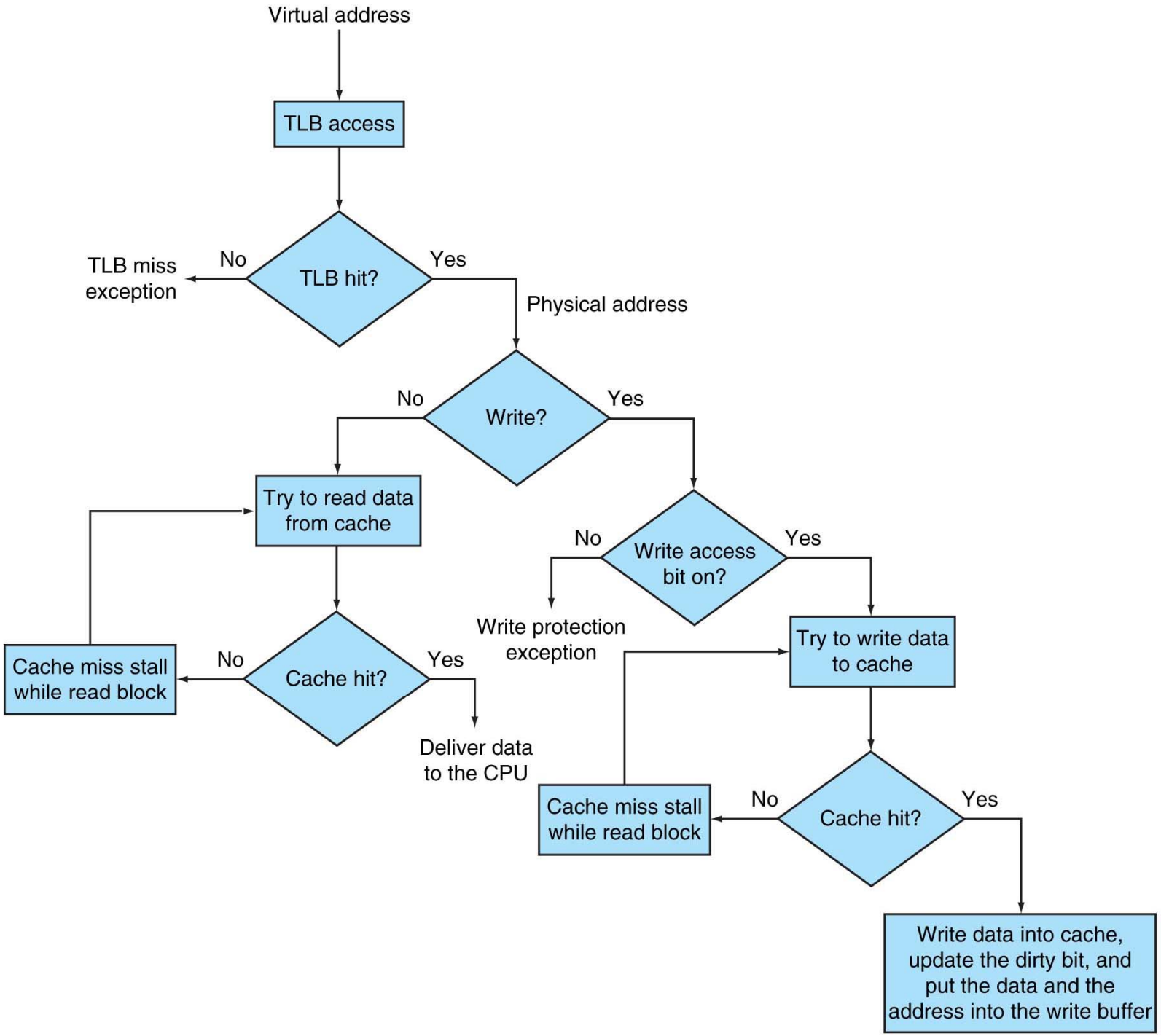
- Increase Speed
  - More cache is faster (up to a point)
    - Searching cache for data takes time
- Cost
  - More cache is expensive

# Cache/Main Memory Structure



# Cache operation - overview

- CPU requests contents of memory location
- Checks cache for data
- If present, get from cache (fast) **-- HIT**
- If not present, read required block from main memory to cache **-- MISS**
- Then deliver from cache to CPU
- Cache includes tags to identify which block of main memory is in each cache slot


Virtual address
TLB access
TLB hit?
No
TLB miss exception
Yes
Physical address
Write?
No
Yes
Try to read data from cache
Cache hit?
No
Cache miss stall while read block
Yes
Deliver data to the CPU
Write access bit on?
No
Write protection exception
Yes
Try to write data to cache
Cache hit?
No
Cache miss stall while read block
Yes
Write data into cache, update the dirty bit, and put the data and the address into the write buffer