# Cache Memory Basics

Read pp. 289-305

# Cache Memory Basics

- Cache memory is a small and fast memory between CPU and main memory
- Computer program has many routines, which are executed repeatedly – locality of reference.
- Cache memory operation should be transparent to CPU
  - CPU provides standard Read and Write control lines and address
  - Only difference to the CPU is the access (cycle) time
    - Data and instructions in cache – fast
    - Data and instruction not in cache – 10 times slower (one order of magnitude)
- Need a word from memory
  - In cache – called hit
  - Not in cache – called miss

# Cache Memory Basics (Continued)

- Is cache worthwhile?
  - If cache is small you miss most of the time. That is not good because cache costs overhead. - Hope most time hits – using advanced replacement mechanisms
- Locality of reference in computer program
  - Property of computer program: most of execution time spent on routines in which many instructions executed repeatedly (such as loop)
  - Manifested in 2 ways
    - Temporal – recently executed instructions likely to be executed again
    - Spatial – instruction close to current instruction likely to be executed again
  - Solutions – continued on the next page

# Cache Memory Basics (Continued)

- Solutions
    - Temporal  - bring instructions into cache when first needed and hopefully remain there until needed again
    - Spatial – do not just bring in one instruction at a time but a block of instructions at a time (read whole block)
- Read operations (instruction or data)
    - Hit – read into CPU
    - Miss – block containing required word read into cache
        - After the entire block is loaded, the requested word forwarded to CPU
        - The requested word is forwarded to CPU as soon as it is in cache – called load -through

# Cache Memory Basics (Continued)

- Write operations
  - Write –through: write immediately to both cache and main memory
    - Cache has true picture of memory
    - Slow
    - May have to write the single word several times
  - Write-back: write only to cache and mark word as updated (dirty or modified bit), and write back to main memory late
    - Faster
    - May result in unnecessary write operations because when a cache block is written back to the main memory all words of the block are written back