Lecture 28
July 16

Memory Caches

Write cache

- So far, we have only considered reads from memory.
- Memory writes can also be cached
- There are several ways this can be done:
    - Write to other cache & memory immediately
    - Write only to memory
    - Write only to cache
        - Update memory when that word is removed from cache
    - If several units are accessing memory, cache coherence becomes a problem
        - .e.g. multiple processors, processor & I/O subsystem
        - What if one processor writes & the other read before the cache does the write?

Disk Cache

- The relative speed difference between memory & disk is very large.
    - Caching data from the disk can make a huge speed difference
- Often hard disks have cache built in.
    - ~ 2MB in modern drives
    - the cache is handled by circuitry on the drive
    - this cache is invisible to the programmer
    - the operating system can also keep a disk cache in RAM
        - al disk access in a modern PC is done through the OS
    - if a request is made, the OS checks the cache in RAM
        - if not there ask the hard drive for it
- disk work differently from memory, so the cache can work differently
    - it’s easy to read a large chunk of adjacent data at once
    - so we can easily cache the next data from the disk
    - if that is accessed next, it will already be in the cache
- disk caches typically use LRU replacement
    - there is enough time to do it
- writing cache
    - if a program writes to the disk, it could be stored in the RAM cache
    - it can actually be written when the disk + CPU are free
    - if the write doesn’t make it to the disk, it could get lost
        - Power failure/ off switch
        - Disk eject
        - Crash

Virtual memory

- the problem: not enough space in memory to do what we won’t
- when programmer runs out of space in the register file, info is moved memory
    - done manually by the programmer
    - the same thing Can be done for memory
    - if space runs out move some stuff to HD until its needed
    - allows more data in “memory “ then there is actual RAM
    - done automatically by the hardware & OS
- modern computers can address a lot of memory
    - e.g. 32 bits addresses → 2^{32} bytes = 4 GB
    - most computers have a fraction of that
    - if some data is kept on the HD we could use more of the address space
    - the “virtual memory”