Computer Systems Lecture 18

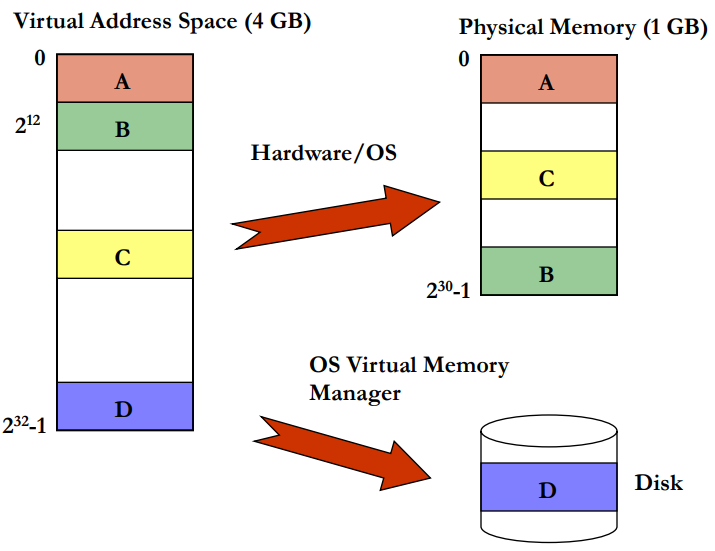
Virtual Memory: Motivation

Virtual memory addresses two main problems:

1. Capacity: how do we relieve the programmers/users from dealing with limited main memory?
   1. We want to allow for the physical memory to be smaller than the program’s address space (e.g. 32 bits -> 4GB; 64 bit -> A LOT MORE)
   2. We want to allow multiple programs to share the limited physical memory with no human intervention
2. Safety: how do we allow for safe and efficient sharing of memory among multiple programs?
   1. We want to prevent user programs from accessing the memory used by the OS
   2. We want strict control of access by each user program to the memory of other user programs.

Virtual Memory

* The basic idea: each program thinks it owns the entire memory -> the virtual address space
  + PC and load/store addresses are virtual addresses
* Actual main memory: physical address space
  + Virtual addresses are translated on the fly to physical addresses
  + Parts of virtual address space not recently used are stored on disk
* Address translation is done jointly by the OS and hardware

Address Translation for a Program

Physical Memory as a Cache for VM

* Virtual memory space can be larger than physical memory
  + Programmer always sees the full address space
* Physical memory is a cache for the virtual memory
  + Physical memory holds the currently used portions of a program’s code and data (exploits locality)
* Secondary storage (disk or SSD) ‘backs’ the physical memory
  + OS reserves a portion of the disk for swap space
  + OS swaps portions of each process’ code and data areas in and out of physical memory on demand (called paging)
  + Paging is transparent to the application ad the programmer

Paging

* A cache line or block of VM is called a page
  + Simply page or virtual page for virtual memory
  + Page frame or physical page for physical memory
* Typical page sizes are 4-16 KB (MB or GB in servers)
  + Large enough for efficient disk use and to keep translation tables (called page tables) small
* Mapping is done through a per-program page table
  + This allows control of which pages each program can access
  + Different programs can use the same virtual addresses