Memory Management

Basic Hardware

A pair of base and limit registers define the logical address space

• base register holds the smallest legal physical memory address

• the limit register specifies the size of the range

Address binding of instructions and data to memory addresses can happen at three different stages

• Compile time: If memory location known a priori, absolute code can be generated; must recompile code if starting location changes

• Load time: Must generate relocatable code if memory location is not known at compile time

• Execution time: Binding delayed until run time if the process can be moved during its execution from one memory segment to another

• The concept of a logical address space that is bound to a separate physical address space is central to proper memory management

• Logical address – generated by the CPU; also referred to as virtual address

• Physical address – address seen by the memory unit

• Logical address space is the set of all logical addresses generated by a program

• Physical address space is the set of all physical addresses generated by a program

Memory-Management Unit (MMU)

The run-time mapping from virtual to physical addresses is done by a hardware device called the MMU

• The base register is now called are location register

• The value in the relocation register is added to every address generated by a user process at the time the address is sent to memory

• Ex: if the base is at 14000; an access to location 346 is mapped to location 14346

Dynamic Loading

• With dynamic loading, a routine is not loaded until it is called

• The advantage of dynamic loading is that a routine is loaded only when it is needed

Dynamic Linking

• Static linking – system libraries and program code combined by the loader into the binary program image

• Dynamically linked libraries are system libraries that are linked to user programs when the programs are run

Small piece of code, stub, used to locate the appropriate memory-resident library routine

Swapping

• A process can be swapped temporarily out of memory to a backing store, and then brought back into memory for continued execution

• Total physical memory space of processes can exceed physical memory
• Backing store – fast disk large enough to accommodate copies of all memory images for all users
• Roll out, roll in – swapping variant used for priority-based scheduling algorithms; lower-priority process is swapped out so higher-priority process can be loaded and executed

Memory allocation
Variable-partition scheme, the OS keeps a table indicating which parts of memory are available
• Hole – block of available memory; holes of various size are scattered throughout memory
• The ones most commonly used to select a free hole from the set of available holes
• First-fit - Allocate the first hole that is big enough
• Best-fit - Allocate the smallest hole that is big enough
• Worst-fit - Allocate the largest hole
Fragmentation

• As processes are loaded and removed from memory, the free memory space is broken into little pieces
• External Fragmentation – total memory space exists to satisfy a request, but it is not contiguous
• Reduce external fragmentation by compaction
• Internal Fragmentation – allocated memory may be slightly larger than requested memory; this size difference is memory internal to a partition, but not being used

Paging

• Physical address space of a process can be noncontiguous; process is allocated physical memory whenever the latter is available
• Avoids external fragmentation
• Avoids the need for compaction
• Divide physical memory into fixed-sized blocks called frames
• Size is power of 2, between 512 bytes and 16 Mbytes
• Divide logical memory into blocks of same size called pages
• To run a program of size N pages, need to find N free frames and load program
• Set up a page table to translate logical to physical addresses
Address generated by CPU is divided into:
• Page number (p) – used as an index into a page table which contains base address of each page in physical memory
• Page offset (d) – combined with base address to define the physical memory address that is sent to the memory unit

Page-table base register (PTBR) points to the page table
• Page-table length register (PTLR) indicates size of the page table

Memory Protection
if read-only or read-write access

Segment table – maps two-dimensional physical addresses; each
table entry has:
• base – contains the starting physical address where the segments reside in

memory

• limit – specifies the length of the segment

Segment-table base register (STBR) points to the segment table’

s

location in memory

• Segment-table length register (STLR) indicates number of segments

used by a program;

segment number s is legal if s < STLR

• Hierarchical Paging

• Hashed Page Tables

Hierarchical Page Tables

• Break up the logical address space into multiple page tables• Inverted Page Tables

A logical address (on 32-bit machine with 1K page size) is divided into:

• a page number consisting of 22 bits

• a page offset consisting of 10 bits

• Since the page table is paged, the page number is further divided

into:

• a 12-bit page number

• a 10-bit page offset

• where p1

is an index into the outer page table, and p2

is the

displacement within the page of the inner page table

• Known as forward-mapped page table

Hashed Page Tables

• For handling address spaces larger than 32 bits

• The virtual page number is hashed into a page table

• This page table contains a chain of elements hashing to the same location

• Each element contains

• (1) the virtual page number

• (2) the value of the mapped page frame

• (3) a pointer to the next element

• Virtual page numbers are compared in this chain searching for a match

• If a match is found, the corresponding physical frame is extracted

Inverted Page Table

• Rather than each process having a page table and keeping track of all possible logical pages, track all physical pages

• One entry for each real page of memory

• Entry consists of the virtual address of the page stored in that real memory location, with information about the process that owns that page

• Inverted Page Table - has one entry for each real page (or frame) of memory