Virtual Memory

PARITA PATEL

Virtual Memory

- Not a Real Memory
- Illusion of Physical Memory

In a memory hierarchy system, programs and data are first stored in auxiliary memory. Portions of a program or data are brought into main memory as they are needed by the CPU.

Virtual memory is a concept used in some large computer systems that permit the user to construct programs as though a large memory space were available, equal to the totality of auxiliary memory.

Virtual Memory

- Virtual memory is used to give programmers the illusion that they have a very large memory at their disposal, even though the computer actually has a relatively small main memory.
- A virtual memory system provides a mechanism for translating program-generated addresses into correct main memory locations.
- This is done dynamically, while programs are being executed in the CPU.
- The translation or mapping is handled automatically by the hardware by means of a mapping table.

Continue..

Let's assume that ,

Main Memory= 4GB

Program Size= 6 GB

CPU can execute this program but it is possible to store this program in the main memory?

NO.

There is no need to load a entire capacity of the program in main memory at once.

Relation between address and memory

An address used by a programmer will be called a **virtual address**, and the set of such addresses the **address space.**

An address in main memory is called a **location** or **physical address**.The set of such locations is called the **memory space**.

Thus the address space is the set of addresses generated by programs as they reference instructions and data; the memory space consists of the actual main memory locations directly addressable for processing.

Relation between address and memory

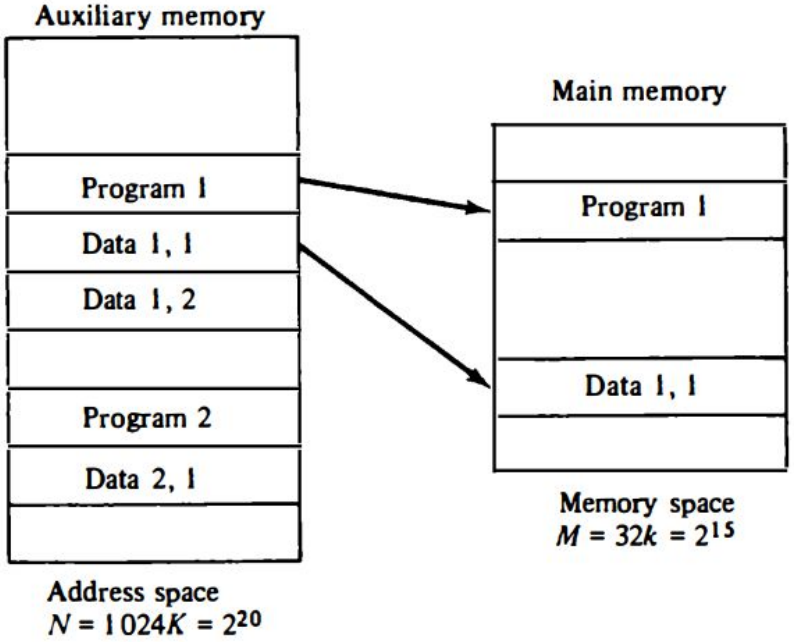


Figure 12-16 Relation between address and memory space in a virtual memory system.

Relation between address and memory

- Consider a computer with a ***main-memory capacity of 32K words*** (K = 1024). ***Fifteen bits*** are needed to specify a physical address in memory since **32K = 2^{15}**
- Suppose that the computer has available auxiliary memory for storing **2^{20} = 1024K words.**
- Thus auxiliary memory has a capacity for storing information equivalent to the capacity of 32 main memories.
- Denoting the **address space by N** and the **memory space by M**, we then have for this example **N = 1024K** and **M = 32K**.

Continue..

In a multiprogram computer system, programs and data are transferred to and from auxiliary memory and main memory based on demands imposed by the CPU.

Suppose that program 1 is currently being executed in the CPU. Program 1 and a portion of its associated data are moved from auxiliary memory into main memory as shown in Figure.

Address Mapping Using Pages

The address space and the memory space are each divided into groups of fixed size.

The **physical memory** is broken down into groups of equal size called **blocks**, which may range from ***64 to 4096 words*** each.

The term page refers to groups of address space of the same size.

For example, if a page or block consists of 1K words, then, using the previous example, address space is divided into 1024 pages and main memory is divided into 32 blocks.

Although both a page and a block are split into groups of 1K words, a page refers to the organization of address space, while a block refers to the organization of memory space.

Continue..

The programs are also considered to be split into pages.

Portions of programs are moved from auxiliary memory to main memory in records equal to the size of a page. The term "***page frame***" is sometimes used to denote a ***block.***

Consider a computer with an address space of 8K and a memory space of 4K. If we split each into groups of 1K words we obtain eight pages and four blocks as shown in Figure.

Continue..

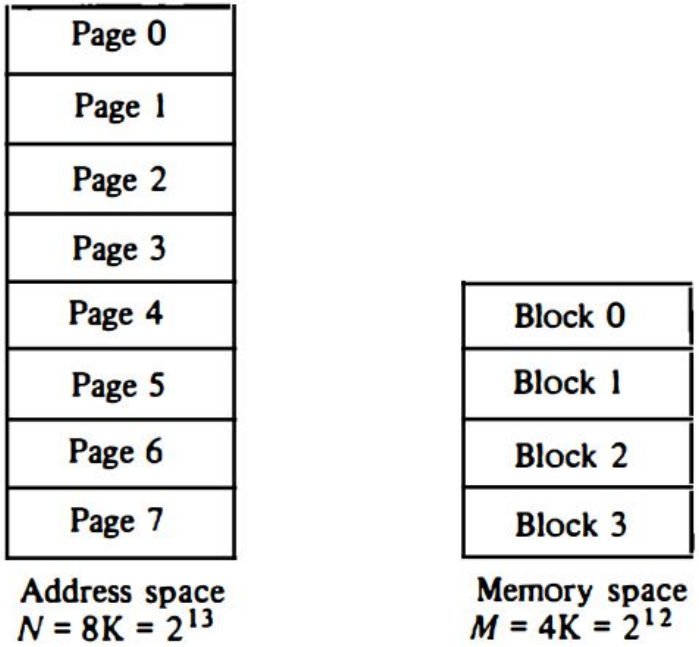
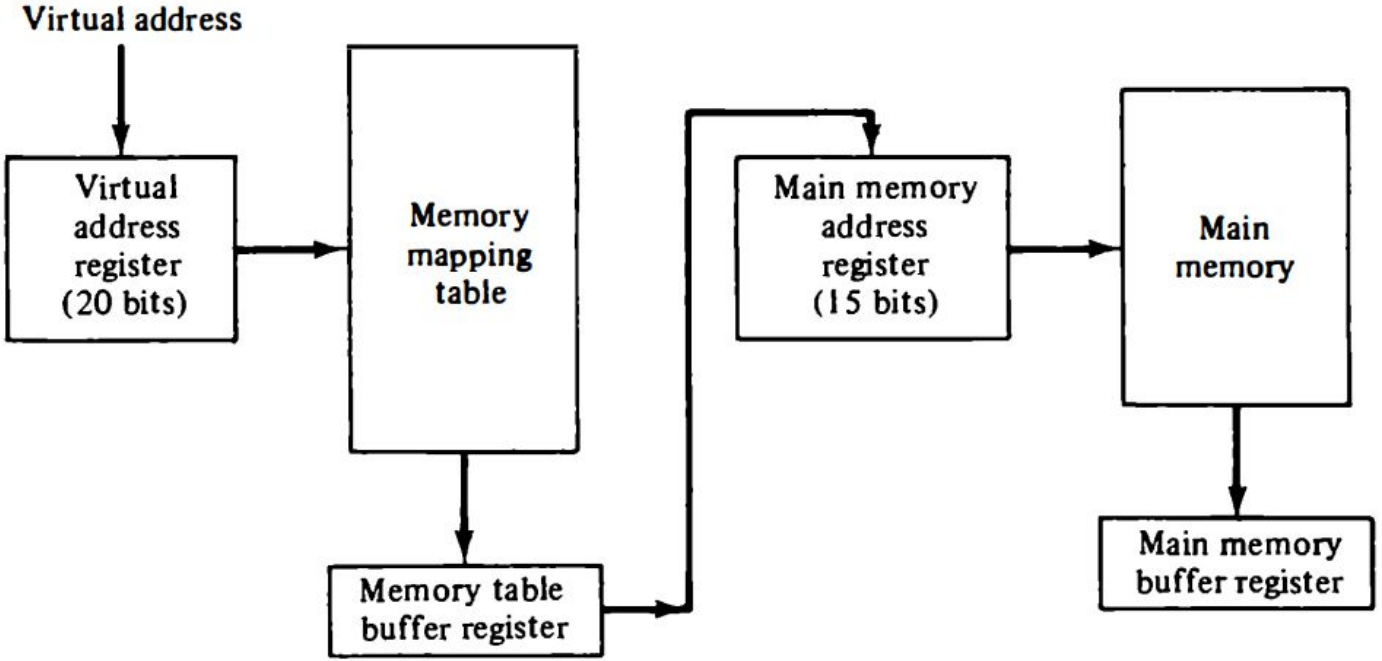


Figure 12-18 Address space and memory space split into groups of 1K words.

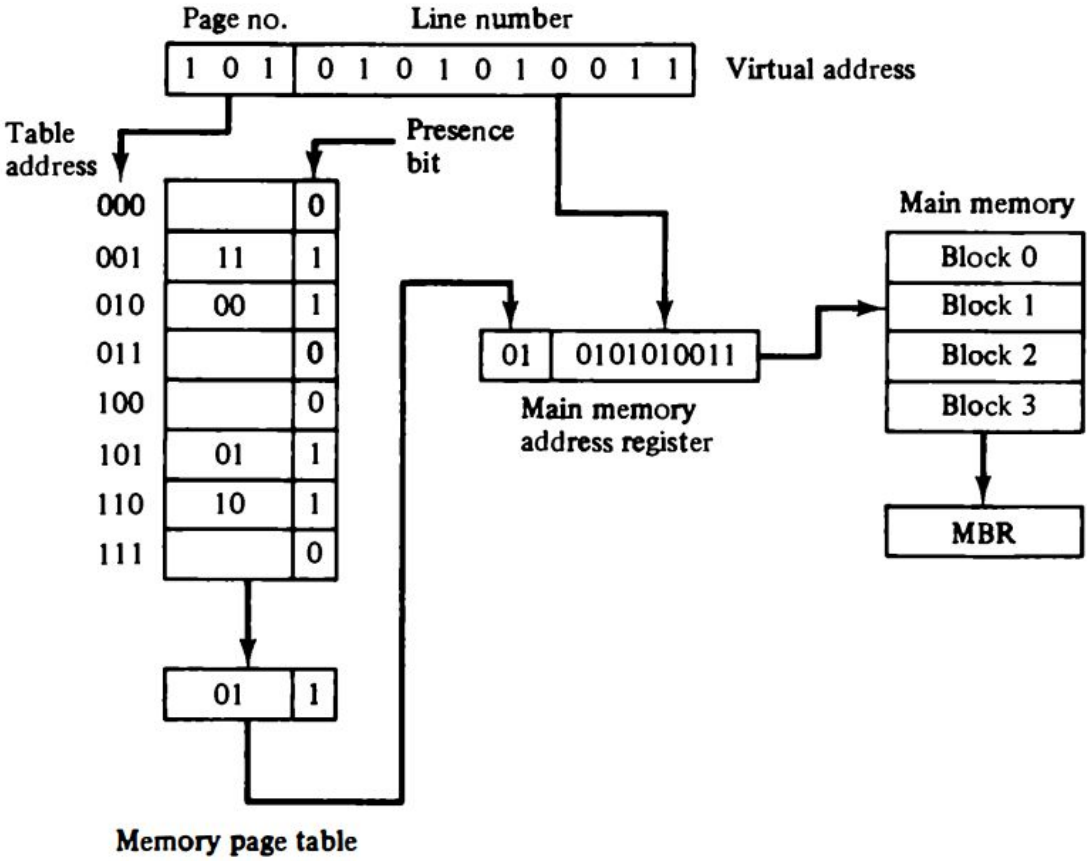
Memory table for mapping a virtual address

Figure 12-17 Memory table for mapping a virtual address.



Memory table in a paged system.

Figure 12-19 Memory table in a paged system.



Page Replacement

A virtual memory system is a combination of hardware and software techniques. The memory management software system handles all the software operations for the efficient utilization of memory space. It must decide

(1) which page in main memory ought to be removed to make room for a new Page,

(2) when a new page is to be transferred from auxiliary memory to main memory, and

(3) where the page is to be placed in main memory.

The hardware mapping mechanism and the memory management software together constitute the architecture of a virtual memory.

Page Fault

When a program starts execution, one or more pages are transferred into main memory and the page table is set to indicate their position. The program is executed from main memory until it attempts to reference a page that is still in auxiliary memory. This condition is called ***page fault***.

When page fault occurs, the execution of the present program is suspended until the required page is brought into main memory.

Since loading a page from auxiliary memory to main memory is basically an VO operation, the operating system assigns this task to the VO processor.

Continue..

When a page fault occurs in a virtual memory system, it signifies that the page referenced by the CPU is not in main memory.

A new page is then transferred from auxiliary memory to main memory. If main memory is full, it would be necessary to remove a page from a memory block to make room for the new page.

The policy for choosing pages to remove is determined from the replacement algorithm that is used.

The goal of a replacement policy is to try to remove the page least likely to be referenced in the immediate future.

Continue..

Two of the most common replacement algorithms used are

FIFO (First-in, First-Out)

LRU (Least Recently used)

FIFO (First-in, First-Out)

The FIFO algorithm selects for replacement the page that has been in memory the longest time.

Each time a page is loaded into memory, its identification number is pushed into a FIFO stack.

FIFO will be full whenever memory has no more empty blocks. When a new page must be loaded, the page least recently brought in is removed. The page to be removed is easily determined because its identification number is at the top of the FIFO stack.

The FIFO replacement policy has the advantage of being easy to implement. It has the disadvantage that under certain circumstances pages are removed and loaded from memory too frequently.

LRU (Least Recently used)

The LRU policy is more difficult to implement but has been more attractive on the assumption that the least recently used page is a better candidate for removal than the least recently loaded page as in FIFO.

The LRU algorithm can be implemented by associating a counter with every page that is in main memory. When a page is referenced, its associated counter is set to zero.

Continue..

At fixed intervals of time, the counters associated with all pages presently in memory are incremented by 1. The least recently used page is the page with the highest count. The counters are often called aging registers, as their count indicates their age, that is, how long ago their associated pages have been referenced.

Differences Between Physical and Virtual Memory

	Physical Memory	Virtual Memory
Definition	It is the actual RAM	It is a memory management technique
Approach	It uses swapping	It uses paging
Accessibility	It can access CPU	It cannot access the CPU directly
Size	It is limited to the size of the RAM	It is limited by the size of the disk
Speed	It is faster than virtual memory	It is slow compared to physical memory