Computer Organization and Architecture

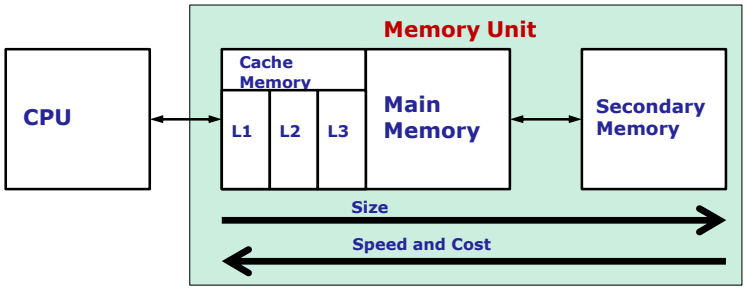
Memory Unit

Veena Thenkanidiyoor
National Institute of Technology Goa

1

1

Memory Hierarchy



- Processor processes instructions and data faster than it can be fetched from memory unit
- Memory access time is the bottleneck
- One way to reduce memory access time is to use faster memory
  - A small and faster memory bridge the gap between processor and main memory
- **Virtual memory**

2

2

Read-Only Memories

3

3

Read-Only Memories (ROMs)

- **SRAM and DRAM are volatile** i.e. they loose the stored information if power is turned off
  - Some applications need memory devices that retain the content when the power supply is turned off too
- Disk stores OS—when computer switched on, OS must be loaded onto memory from disk
  - Requires execution of a program that "boots" the operating system
  - Boot program is also very large– stored in disk
- Processor must execute some instructions that load the boot program into memory
- **Memory is volatile**– the processor would have no means of accessing these instructions
- **Provide a small amount of non-volatile memory**
  - **To hold instructions that loads the boot program from the disk**

4

4

Read-Only Memories (ROMs)

- Read-only memories are semiconductor, non-volatile memories
- Their normal operation involve only reading the stored data
- They are extensively used in embedded systems
- Different types of ROMs
  - Read Only Memory (ROM)
  - Programmable ROM (PROM)
    - Allows for loading data by the user, less expensive
  - Erasable, reprogrammable ROM (EPROM)
    - UV light is used for erasing the existing content
  - Electrically erasable reprogrammable ROM (EEPROM)

5

5

Flash Memory

- Approach similar to EEPROM
  - Small difference in how writing to be done on the memory
- Have greater density
  - Higher capacity, lower cost per bit
- Consumes less power
  - Suitable for portable devises
    - Hand-held computers, cell phones, digital cameras, and MP3 music players
- Flash cards
  - flash chips mounted on a small card to have a larger module
- Flash drives
  - Replace hard disk drives
  - Replaced floppy disks, CD ROMs etc

6

6

Virtual Memory

7

7

Virtual Memory

- Ideally, entire memory hierarchy would appear to the processor as a single memory unit
- In modern computer system, the physical main memory is not as large as the address space spanned by the address issued by the processor
- When a program (or process) does not completely fits into the main memory, parts of it will be there in secondary memory
- In modern computers, operating system moves the data automatically between main memory and secondary storage
- Programmer does not need to aware of the limitations imposed by the main memory

8

8

Virtual Memory Technique

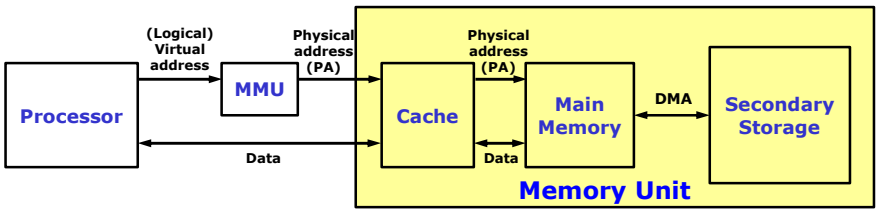
- Technique that automatically move program and data blocks into the physical main memory when they are required for execution
- Using virtual program concept, each program may use entire CPU local address space, at least up to secondary storage
- The address issued by the processor either for instruction or data are called virtual address or logical address
- These addresses are translated into physical memory addresses by a combination of hardware and software

9

9

Memory Management Unit (MMU)

- MMU translate the logical address into physical main memory address
- It is a part of the processor



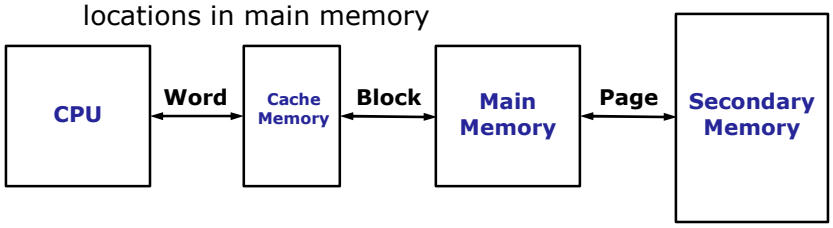
- If the data is not in main memory, MMU causes the operating system to bring data into memory from the disk
- Transfer of data between disk and main memory is performed using direct memory access (DMA) scheme

10

10

Address Translation

- The virtual memory address translation method is based on the concept of fixed length pages
- The address translations assumes that programs and data are composed of fixed size units called pages
- Unit of transfer between secondary memory and main memory is page
  - A page is a block of words that occupy contiguous locations in main memory

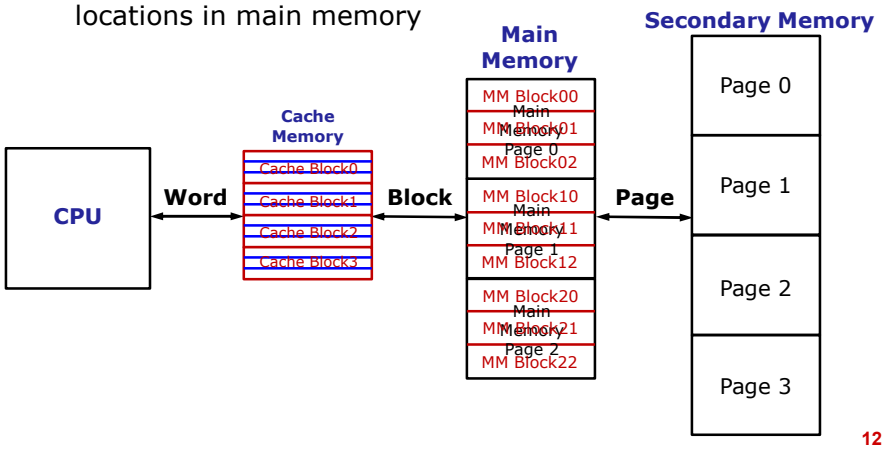


11

11

Address Translation

- The address translations assumes that programs and data are composed of fixed size units called pages
- Unit of transfer between secondary memory and main memory is page
  - A page is a block of words that occupy contiguous locations in main memory



12

12

Page

- The programs or data in the disk are seen by the virtual memory as a collection of pages
- This page is the basic unit of information that is moved between the main memory and the secondary memory
- Each page is of the size 2 KB to 16 KB
- Page should not be too small
  - Disk access time is much longer
  - It take considerable time to locate data in the disk
- Page should not be too large
  - Substantial portion of a page may not be used
- Demand paging: Pages are copied to main memory when requested

13

13

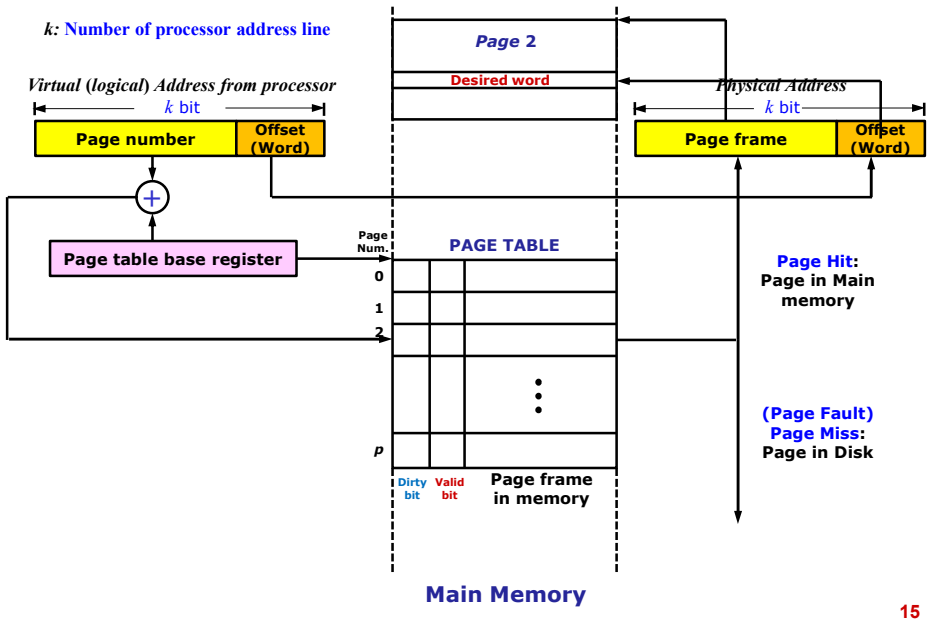
Parallels Between the Concepts of Cache and Virtual Memory

- Cache:
  - Bridges the speed gap between the processor and the main memory
  - It is implemented in hardware
- Virtual memory mechanism:
  - Bridges the size and speed gap between the main memory and secondary storage
  - It is usually implemented in part by software techniques
- Conceptually, cache techniques and main memory techniques are very similar
- They differ mainly in the details of their implementation

14

14

Virtual Memory Address Translation



15

15

Virtual Memory Address Translation

- The virtual address generated by the processor contain page number and offset (word) in the page
- To make sure that required page is in main memory, operating system create **page table for each process**
- This page table is kept in main memory
- Page table base register: Operating system keep the starting address of page table in it
- Page table hold the main memory location for each page
  - The area in main memory that can hold one page is called page frame
- Every entry in page table also include valid bit and dirty bit to describe the status of the page while it is in main memory

16

16

Virtual Memory Address Translation using Translation Lookaside Buffer

- Page table information is used by the MMU for every read and write access
- In order to speed up the address translation procedure, a small cache called Translation Lookaside Buffer (TLB) is incorporated in MMU
- It uses associative/set-associative mapping technique
- It hold a portion of page table corresponding to most recently accessed pages
- TLB holds only the page number and page frame number

17

17

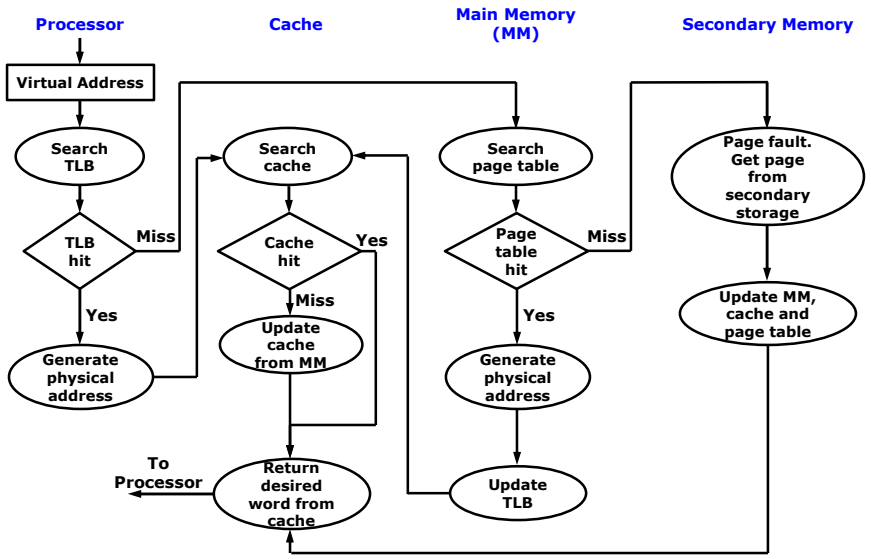
Virtual Memory Address Translation

- Page table information is used by the MMU for every read and write access
- Page fault: Whenever a requested page is not present in the main memory, page fault is said to have occurred
- When a page fault occurs, MMU asks operating system to intervene and raise an exception (interrupt)
  - Process in active get interrupted and control goes to operating system
  - Operating system then copies the requested page from disk to main memory
  - Then returns the control to the interrupted task
- During write operation pages get modified are indicated by dirty bit
- Modified page has to be written back to disk before removed from main memory
- Uses write back policy only

18

18

Operation of Memory Hierarchy and Virtual Memory Technique



19

19