S. S Jain Subodh P.G. (Autonomous) College

SUBJECT - Computer Architecture

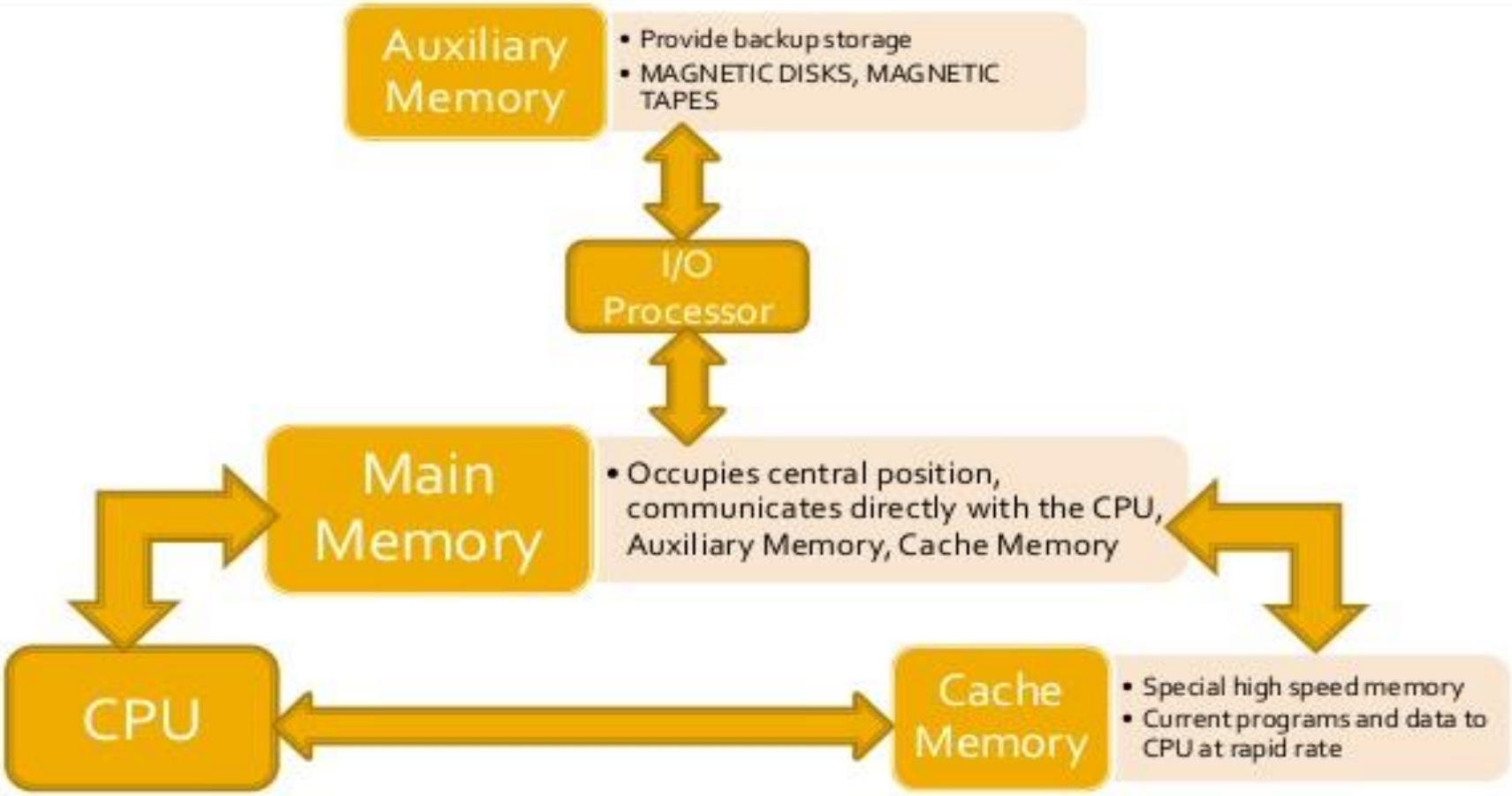
TITLE - Memory System Design

Created By:
Shalu J. Rajawat

Memory Organisation

- Memory Hierarchy
- Main Memory
- Auxiliary Memory
- Associative Memory
- Cache Memory
- Virtual Memory

Memory Hierarchy

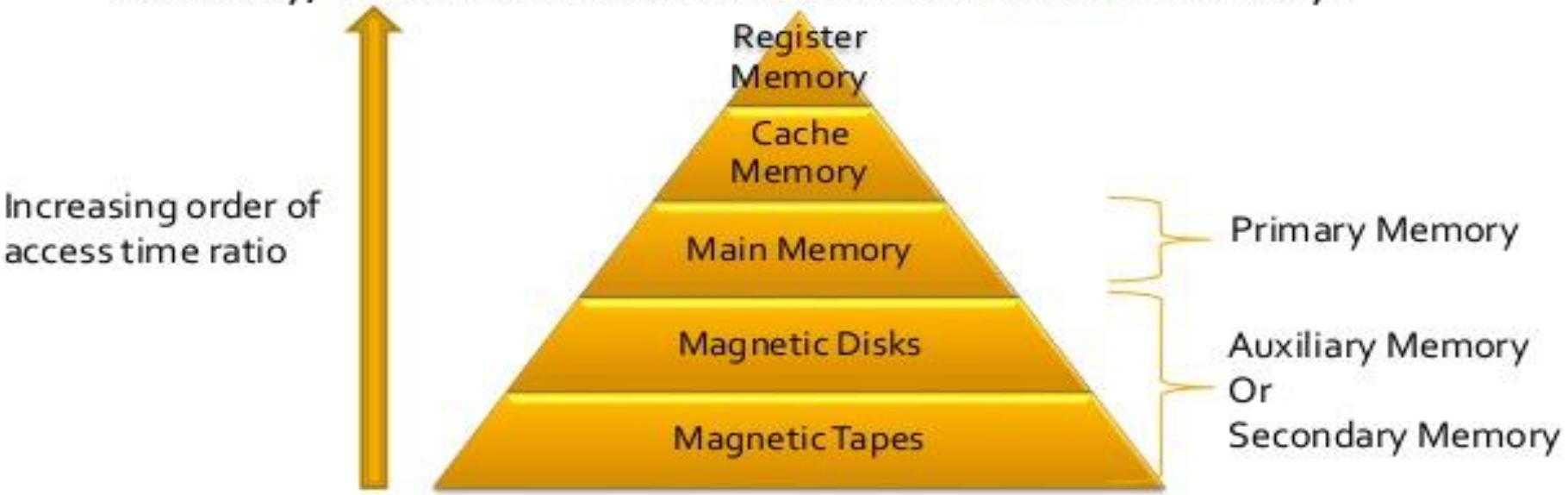


Memory Hierarchy

- CPU logic is usually faster than main memory access time, with the result that processing speed is limited primarily by the speed of main memory.
- The cache is used for storing segments of programs currently being executed in the CPU and temporary data frequently needed in the present calculations.
- The typical access time ratio between cache and main memory is about 1 to 7~10 .
- Auxiliary memory access time is usually 1000 times that of main memory.

Memory Hierarchy

- The memory hierarchy system consists of all storage devices employed in a computer system from the slow by high-capacity auxiliary memory to a relatively faster main memory, to an even smaller and faster cache memory.



Access Methods

- Each memory is a collection of various memory location. Accessing the memory means finding and reaching desired location and than reading information from memory location. The information from locations can be accessed as follows:

1. **Random access**
2. **Sequential access**
3. **Direct access**

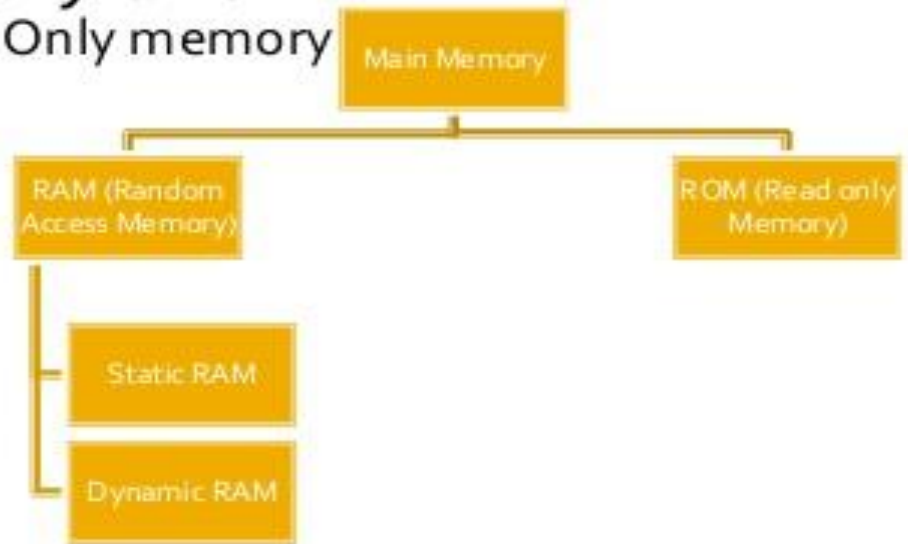
- **Random Access**: It is the access mode where each memory location has a unique address. Using these unique addresses each memory location can be addressed independently in any order in equal amount of time. Generally, main memories are random access memories(RAM).

Access Methods

- **Sequential Access**: If storage locations can be accessed only in a certain predetermined sequence, the access method is known as serial or sequential access.
  - Opposite of RAM: **Serial Access Memory** (SAM). SAM works very well for memory **buffers**, where the data is normally stored in the order in which it will be used (a good example is the texture buffer memory on a video card , magnetic tapes, etc.).
- **Direct Access**: In this access information is stored on tracks and each track has a separate read/write head. This features makes it a semi random mode which is generally used in magnetic disks.

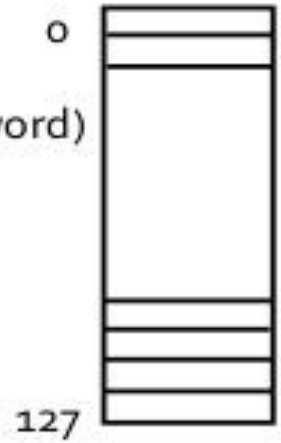
Main Memory

- Most of the main memory in a general purpose computer is made up of **RAM** integrated circuits chips, but a portion of the memory may be constructed with **ROM** chips.
- **RAM**– Random Access memory
  - Integrated RAM are available in two possible operating modes, ***Static*** *and* ***Dynamic.***
- **ROM**– Read Only memory

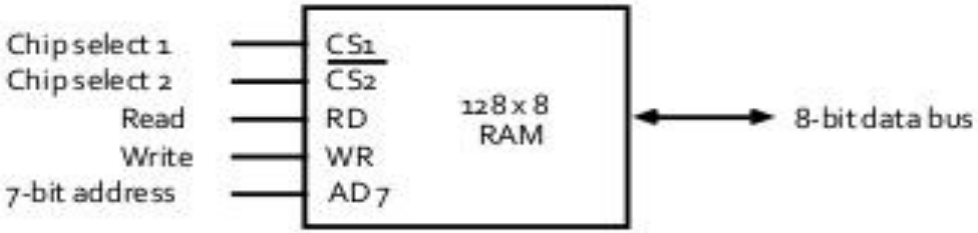


RANDOM ACCESS MEMORY (RAM)

- RAM is used for storing bulk of programs and data that is subject to change.



Typical RAM chip



CS_1	$\overline{CS_2}$	RD	WR	Memory function	State of data bus
0	0	x	x	Inhibit	High-impedence
0	1	x	x	Inhibit	High-impedence
1	0	0	0	Inhibit	High-impedence
1	0	0	1	Write	Input data to RAM
1	0	1	x	Read	Output data from RAM
1	1	x	x	Inhibit	High Impedence

Types of Random Access Memory (RAM)

- Static RAM (SRAM)
  - Each cell stores bit with a six-transistor circuit.
  - Retains value indefinitely, as long as it is kept powered.
  - Faster (8-16 times faster) and more expensive (8-16 times more expensive as well) than DRAM.
- Dynamic RAM (DRAM)
  - Each cell stores bit with a capacitor and transistor.
  - Value must be refreshed every 10-100 ms.
  - Slower and cheaper than SRAM. Has reduced power consumption, and a large storage capacity.

In contrast to , SRAM and DRAM:

- Non Volatile RAM (NVRAM)
  - retains its information when power is turned off (non volatile).
  - best-known form of NVRAM memory today is flash memory.

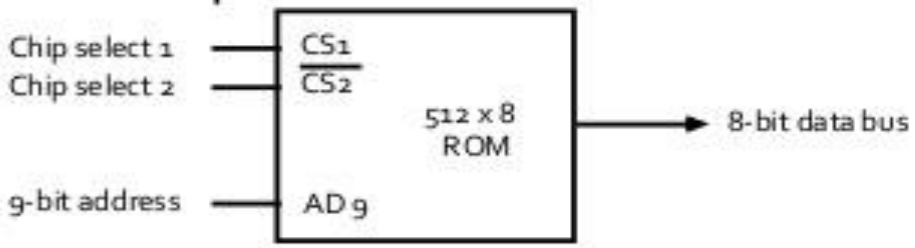
Virtually all desktop or server computers since 1975 used DRAMs for main memory and SRAMs for cache.

READ ONLY MEMORY (ROM)

- It is non-volatile memory, which retains the data even when power is removed from this memory. Programs and data that can not be altered are stored in ROM.
- ROM is used for storing programs that are **PERMANENTLY** resident in the computer and for tables of constants that do not change in value once the production of the computer is completed.
- The ROM portion of main memory is needed for storing an initial program called ***bootstrap loader***, which is to start the computer operating system when power is turned on.

READ ONLY MEMORY (ROM)

- Typical ROM chip:



- Since the ROM can only READ, the data bus can only be in output mode.
- No need of READ and WRITE control.
- Same sized RAM and ROM chip , it is possible to have more bits of ROM than of RAM , because the internal binary cells in ROM occupy less space than in RAM.

TYPES OF ROM

- The required paths in a ROM may be programmed in four different ways:

1. **Mask Programming:** It is done by the company during the fabrication process of the unit. The procedure for fabricating a ROM requires that the customer fills out the truth table he wishes the ROM to satisfy.

2. **Programmable Read only memory(PROM):**
PROM contain all the fuses intact giving all 1's in the bits of the stored words. A blown fuse defines binary 0 state and an intact fuse give a binary 1 state. This allows the user to program the PROM by using a special instruments called PROM programmer.

TYPES OF ROM

3. **Erasable PROM(EPROM):** In a PROM once fixed pattern is permanent and can not be altered. The EPROM can be restructured to the initial state even through it has been programmed previously. When EPROM is placed under a special ultra-violet light for a given period of time all the data are erased. After erase, the EPROM returns to it initial state and can be programmed to a new set of values.

4. **Electrically Erasable PROM(EEPROM):** It is similar to EPROM except that the previously programmed connections can be erased with an electrical signal instead of ultra violet light. The advantage is that device can be erased without removing it from it's socket.

Auxiliary Memory

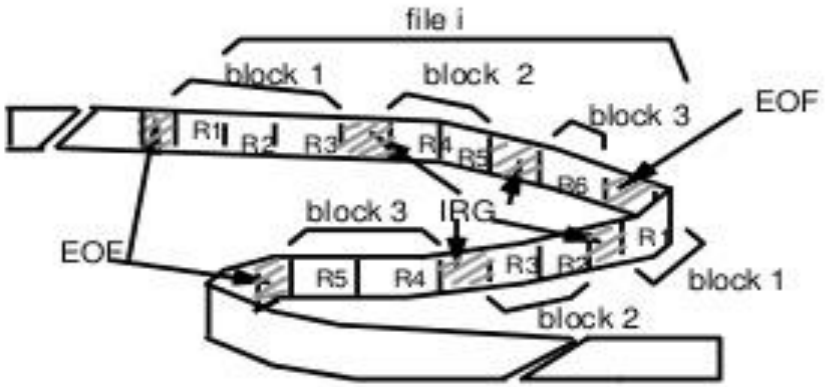
- Also called as Secondary Memory, used to store large chunks of data at a lesser cost per byte than a primary memory for backup.
- It does not lose the data when the device is powered down—it is non-volatile.
- It is not directly accessible by the CPU, they are accessed via the input/output channels.
- The most common form of auxiliary memory devices used in consumer systems is flash memory, optical discs, and magnetic disks, magnetic tapes.

Types of Auxiliary Memory

- *Flash memory:* An electronic non-volatile computer storage device that can be electrically erased and reprogrammed, and works without any moving parts. Examples of this are **USB flash drives** and **solid state drives**.
- *Optical disc:* Its a storage medium from which data is read and to which it is written by lasers. There are three basic types of optical disks: CD-ROM (read-only), WORM (write-once read-many) & EO (erasable optical disks).

Types of Auxiliary Memory

- *Magnetic tapes:* A magnetic tape consists of electric, mechanical and electronic components to provide the parts and control mechanism for a magnetic tape unit.
- The tape itself is a strip of plastic coated with a magnetic recording medium. Bits are recorded as magnetic spots on tape along several tracks called **RECORDS.**
- Each record on tape has an identification bit pattern at the beg. and the end.



- R/W heads are mounted in each track so that data can be recorded and read as a sequence of characters.
- Can be stopped , started to move forward, or in reverse, or can be rewound, but cannot be stopped fast enough between individual characters.

Types of Auxiliary Memory

Magnetic Disk:

- A magnetic disk is a circular plate constructed of metal or plastic coated with magnetized material.
- Both sides of the disk are used and several disks may be stacked on one spindle with read/write heads available on each surface.
- Bits are stored in magnetized surface in spots along concentric circles called tracks. Tracks are commonly divided into sections called sectors.
- Disk that are permanently attached and cannot removed by occasional user are called hard disks.

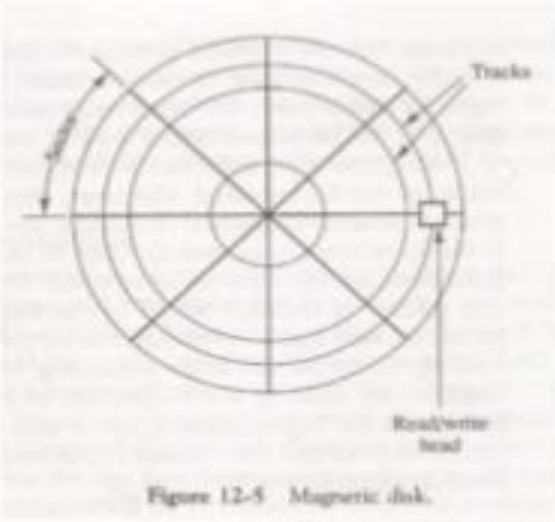


Figure 12-5 Magnetic disk.

ASSOCIATIVE MEMORY

- A memory unit accessed by contents is called an associative memory or content addressable memory(CAM).
- This type of memory is accessed simultaneously and in parallel on the basis of data content rather than by specific address or location.

READ/WRITE OPERATION IN CAM

- **Write operation**:
  - When a word is written in in an associative memory, no address is given.
  - The memory is capable of finding an unused location to store the word.
- **Read operation**:
  - When a word is to be read from an associative memory, the contents of the word, or a part of the word is specified.
  - The memory locates all the words which match the specified content and marks them for reading.

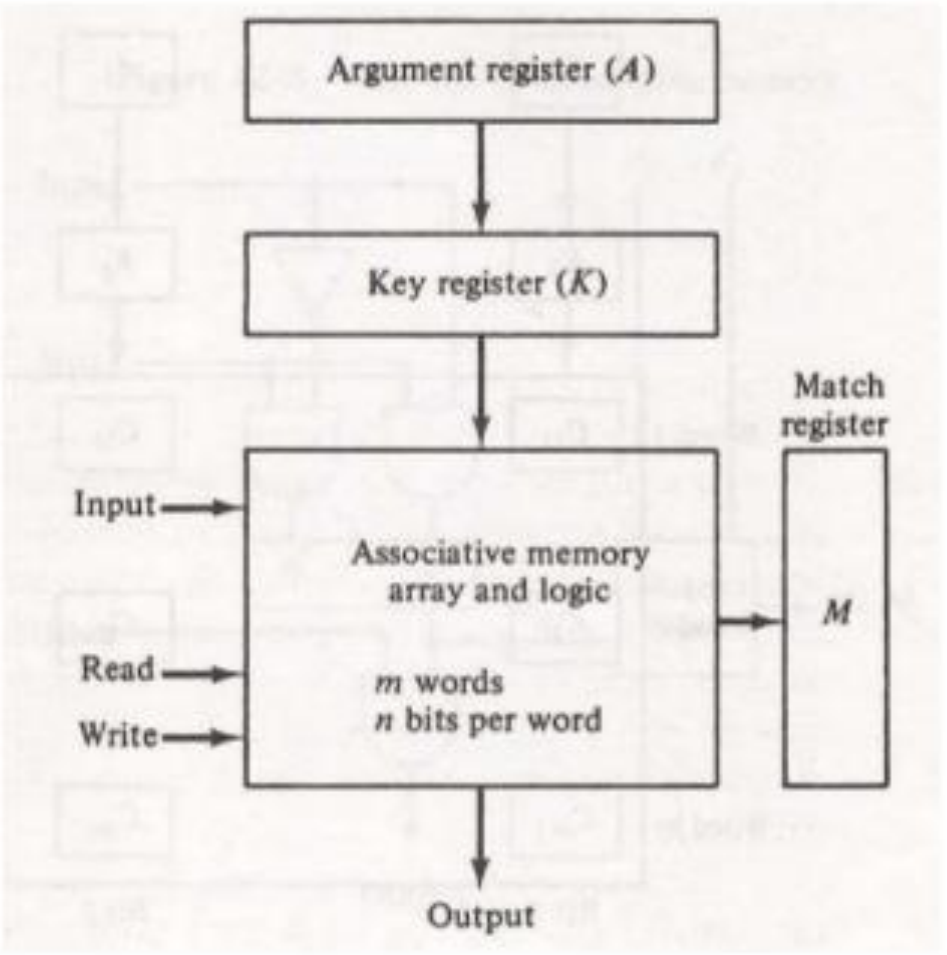
HARDWARE ORGANISATION

Argument register(A): It contains the word to be searched. It has *n* bits(one for each bit of the word).

Key Register(K): It provides mask for choosing a particular field or key in the argument word. It also has *n* bits.

Associative memory array: It contains the words which are to be compared with the argument word.

Match Register(M): It has m bits, one bit corresponding to each word in the memory array . After the matching process, the bits corresponding to matching words in match register are set to 1.



MATCHING PROCESS

- The entire argument word is compared with each memory word, if the key register contains all 1's. Otherwise, only those bits in the argument that have 1's in their corresponding position of the key register are compared.
- Thus the key provides a mask or identifying piece of information which specifies how the reference to memory is made.
- To illustrate with a numerical example, suppose that the argument register A and the key register K have the bit configuration as shown below.
- Only the three left most bits of A are compared with the memory words because K has 1's in these three positions only.

A	101 111100	
K	111 000000	
Word 1	100 111100	no match
Word 2	101 000001	match

DISADVANTAGE

- An associative memory is more expensive than a random access memory because each cell must have an extra storage capability as well as logic circuits for matching its content with an external argument.
- For this reason, associative memories are used in applications where the search time is very critical and must be very short.

CACHE MEMORY

- If the active portions of the program and data are placed in a fast small memory, the **average memory access time** can be reduced.
- Thus reducing the **total execution time** of the program
- Such a fast small memory is referred to as cache memory
- The cache is the fastest component in the memory hierarchy and approaches the speed of CPU component

Basic Operations of Cache

- When CPU needs to access memory, the cache is examined.
- If the word is found in the cache, it is read from the cache memory.
- If the word addressed by the CPU is not found in the cache, the main memory is accessed to read the word.
- A block of words containing the one just accessed is then transferred from main memory to cache memory.
- If the cache is full, then a block equivalent to the size of the used word is replaced according to the replacement algorithm being used.

Hit Ratio

- When the CPU refers to memory and finds the word in cache, it is said to produce a **hit**
- Otherwise, it is a **miss**

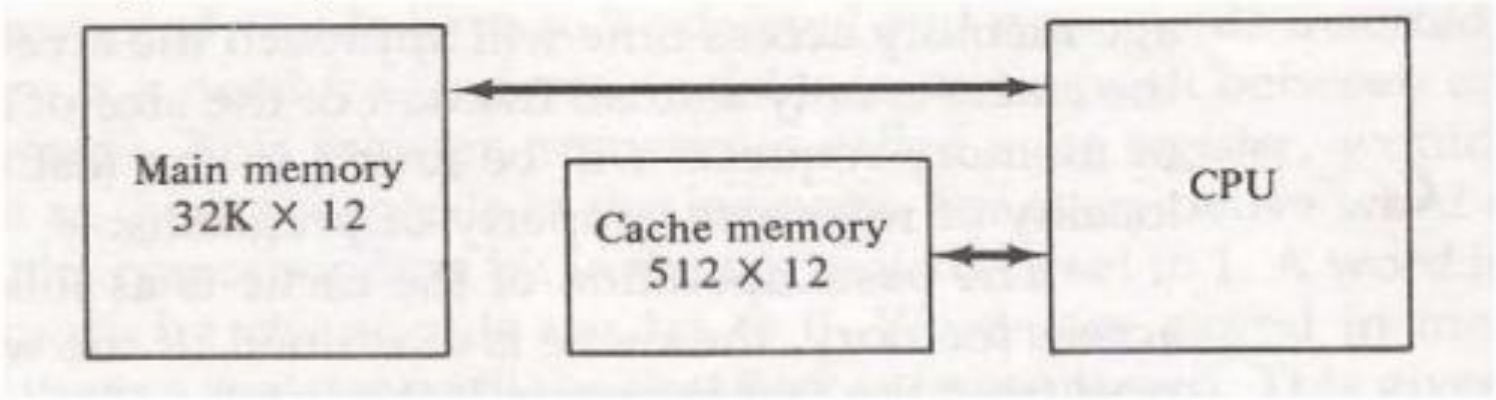
- The performance of cache memory is frequently measured in terms of a quantity called **hit ratio**

Hit ratio = hit / (hit+miss)

Mapping Process

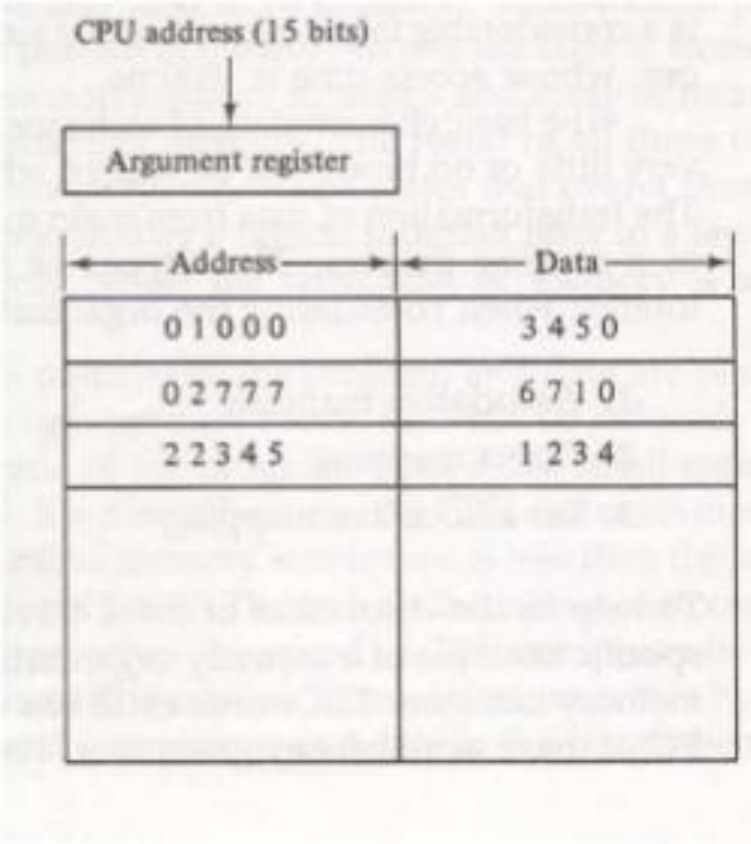
- The transformation of data from main memory to cache memory is referred to as a **mapping** process, there are three types of mapping:
  - Associative mapping
  - Direct mapping
  - Set-associative mapping

- To help understand the mapping procedure, we have the following example:



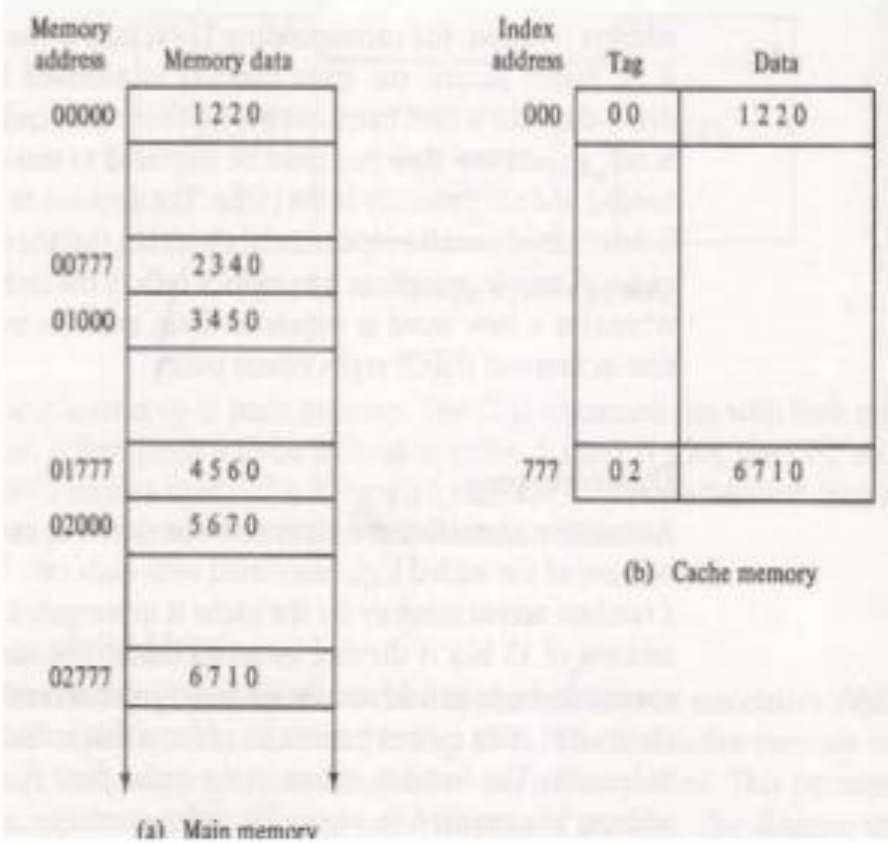
Associative Mapping

- The fastest and most flexible cache organization uses an associative memory.
- The associative memory stores both the address and data of the memory word.
- This permits any location in cache to store a word from main memory.
- The address value of 15 bits is shown as a five-digit **octal** number and its corresponding 12-bit word is shown as a four-digit octal number



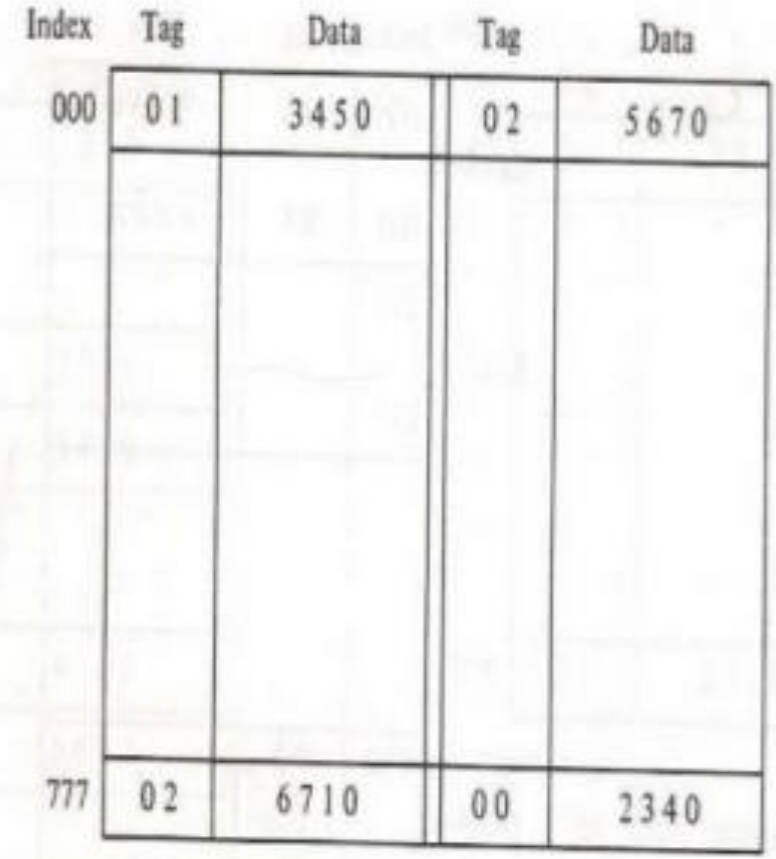
Direct Mapping

- Associative memory is expensive compared to RAM.
- In general case, there are 2^k words in cache memory and 2^n words in main memory (in our case, k=9, n=15).
- The n bit memory address is divided into two fields: k-bits for the **index** and n-k bits for the **tag** field.



Set – Associative Mapping

- The disadvantage of direct mapping is that two words with the same index in their address but with different tag values cannot reside in cache memory at the same time.
- Set-Associative Mapping is an improvement over the direct-mapping in that each word of cache can store two or more word of memory under the same index address.



Index	Tag	Data	Tag	Data
000	01	3450	02	5670
777	02	6710	00	2340

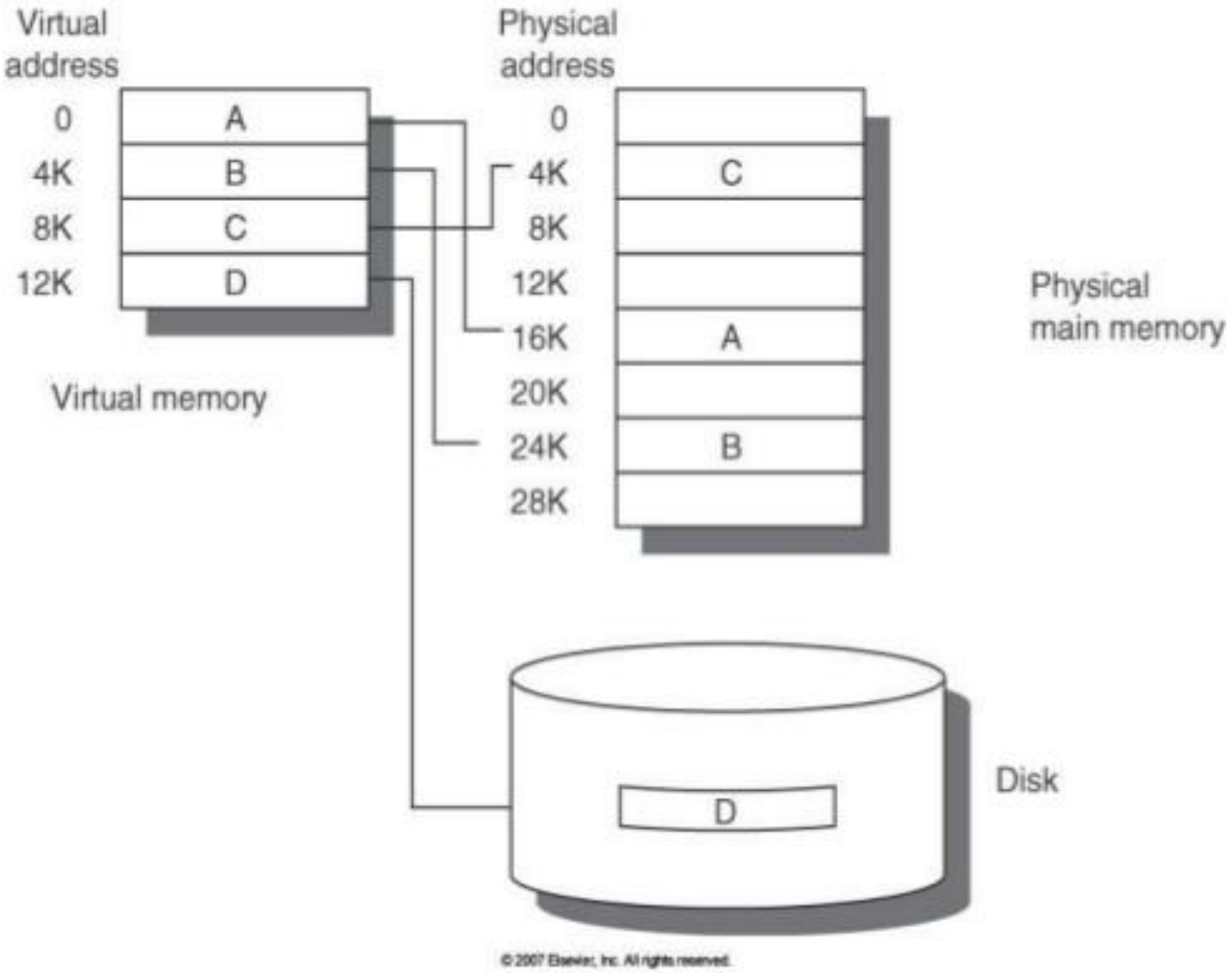
Replacement Algorithms

- Optimal replacement algorithm – find the block for replacement that has minimum chance to be referenced next time.

- Two algorithms:
  - FIFO: Selects the item which has been in the set the longest.
  - LRU: Selects the item which has been least recently used by the CPU.

Virtual Memory

- Virtual memory is a common part of operating system on desktop computers.
- The term Virtual Memory refers to something which appears to be present but actually is not.
- This technique allows users to use more memory for a program than the real memory of a computer.


Virtual address
0
4K
8K
12K
A
B
C
D
Virtual memory
Physical address
0
4K
8K
12K
16K
20K
24K
28K
C
A
B
Physical main memory
D
Disk
© 2007 Elsevier, Inc. All rights reserved.

Need Of Virtual Memory

- Virtual Memory is a imaginary memory which we assume or use, when we have a material that exceeds our memory at that time.

- Virtual Memory is temporary memory which is used along with the ram of the system.

Advantages

- Allows Processes whose aggregate memory requirement is greater than the amount of physical memory, as infrequently used pages can reside on the disk.
- Virtual memory allows speed gain when only a particular segment of the program is required for the execution of the program.
- This concept is very helpful in implementing multiprogramming environment.

DISADVANTAGES

- Applications run rather slower when they are using virtual memory.
- It takes more time to switch between applications.
- Reduces system stability.

MEMORY MANAGEMENT

PAGING

New Concept!!

- Logical address space of a process can be noncontiguous; process is allocated physical memory whenever that memory is available and the program needs it.
- Divide **physical** memory into fixed-sized blocks called **frames** (size is power of 2, between 512 bytes and 8192 bytes).
- Divide **logical** memory into blocks of same size called **pages**.
- Keep track of all free frames.
- 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.
- Internal fragmentation.

MEMORY MANAGEMENT

PAGING

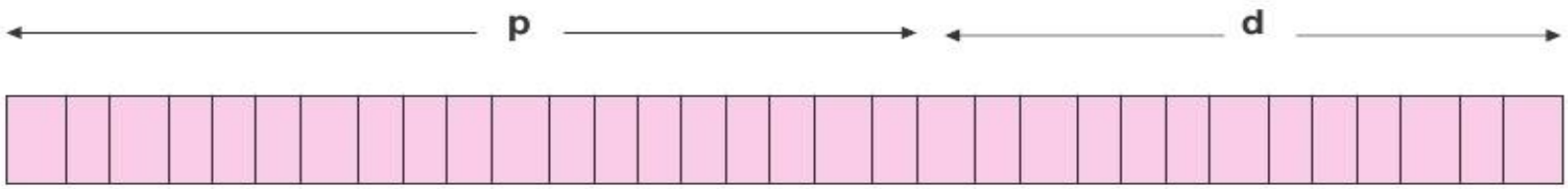
Address Translation Scheme

Address generated by the 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.

4096 bytes = 2^12 – it requires 12 bits to contain the Page offset



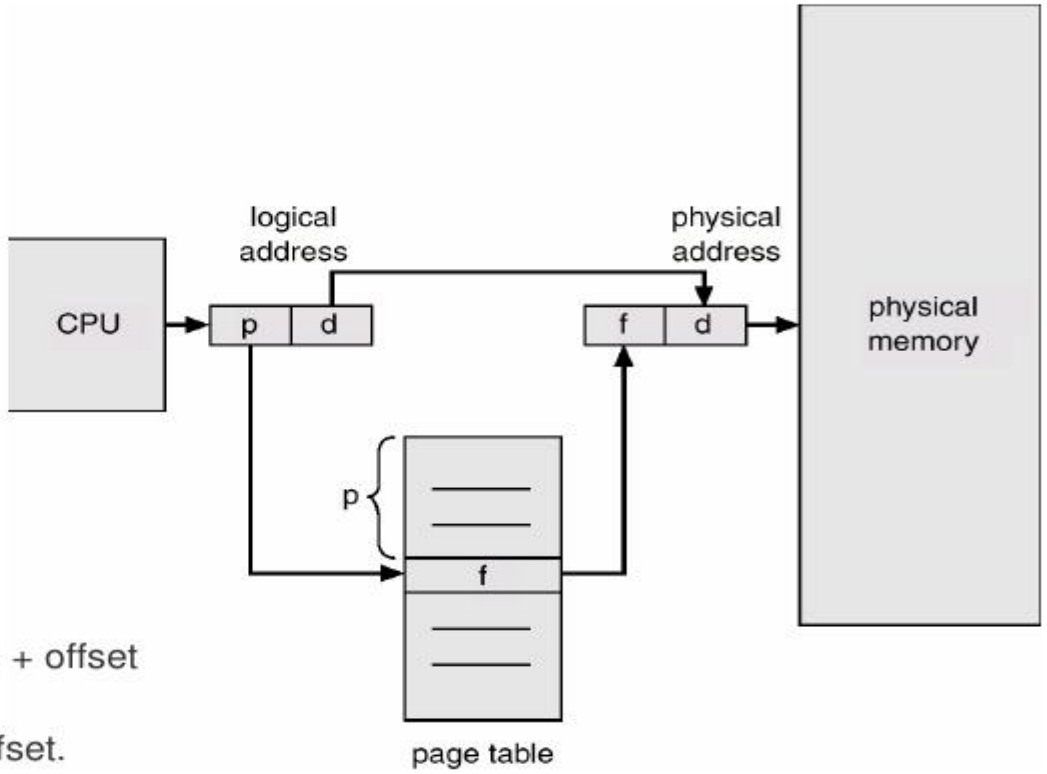
MEMORY MANAGEMENT

PAGING

Permits a program's memory to be physically noncontiguous so it can be allocated from wherever available. This avoids fragmentation and compaction.

Frames = physical blocks
Pages = logical blocks

Size of frames/pages is defined by hardware (power of 2 to ease calculations)



HARDWARE
An address is determined by:

page number (index into table) + offset
---> mapping into --->
base address (from table) + offset.

MEMORY MANAGEMENT

PAGING

Paging Example - 32-byte memory with 4-byte pages

0 a 1 b 2 c 3 d
4 e 5 f 6 g 7 h
8 i 9 j 10 k 11 l
12 m 13 n 14 o 15 p

0	5
1	6
2	1
3	2

Page Table

0	
4	i j k l
8	m n o p
12	
16	
20	a b c d
24	e f g h
28	

Physical Memory

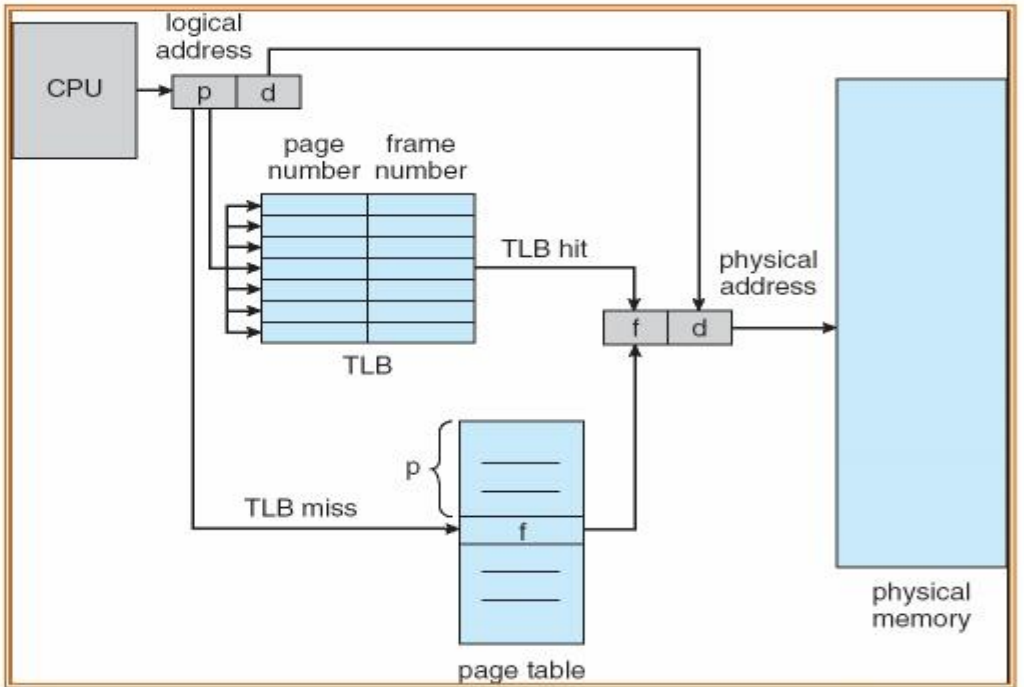
MEMORY MANAGEMENT

PAGING

- A 32 bit machine can address 4 gigabytes which is 4 million pages (at 1024 bytes/page). WHO says how big a page is, anyway?
- Could use dedicated registers (OK only with small tables.)
- Could use a register pointing to table in memory (slow access.)
- Cache or associative memory
- (TLB = Translation Lookaside Buffer):
- simultaneous search is fast and uses only a few registers.

IMPLEMENTATION OF THE PAGE TABLE

TLB = Translation Lookaside Buffer



MEMORY MANAGEMENT PAGING

IMPLEMENTATION OF THE PAGE TABLE

Issues include:

key and value
hit rate 90 - 98% with 100 registers
add entry if not found

Effective access time = %fast * time_fast + %slow * time_slow

Relevant times:

2 nanoseconds to search associative memory – the TLB.
20 nanoseconds to access processor cache and bring it into TLB for next time.

Calculate time of access:

hit = 1 search + 1 memory reference
miss = 1 search + 1 mem reference(of page table) + 1 mem reference.

MEMORY MANAGEMENT

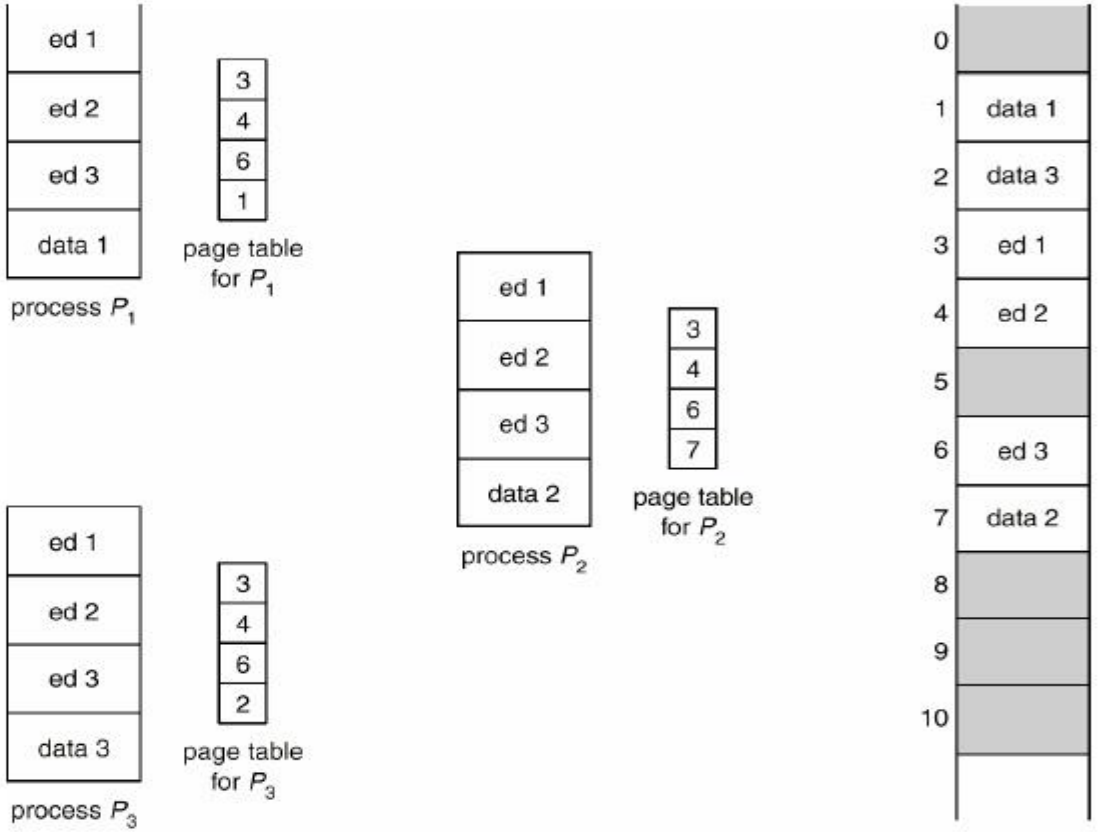
PAGING

SHARED PAGES

Data occupying one physical page, but pointed to by multiple logical pages.

Useful for common code - must be write protected. (NO write-able data mixed with code.)

Extremely useful for read/write communication between processes.



MEMORY MANAGEMENT

PAGING

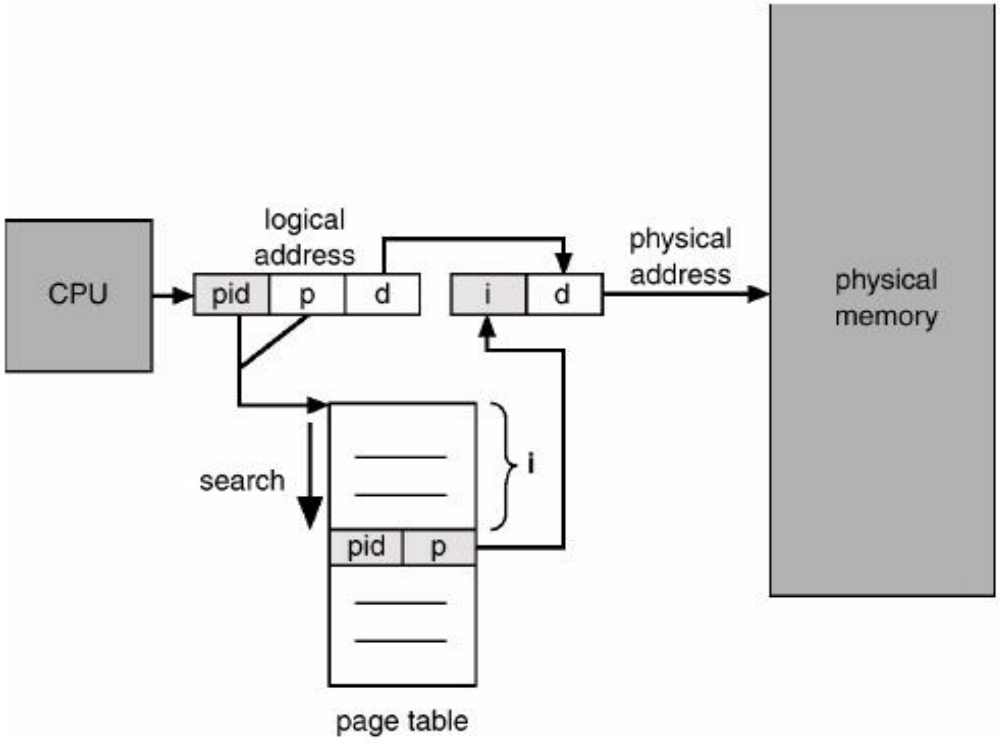
INVERTED PAGE TABLE:

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.

Essential when you need to do work on the page and must find out what process owns it.

Use hash table to limit the search to one - or at most a few - page table entries.



MEMORY MANAGEMENT PAGING

PROTECTION:

- Bits associated with page tables.
- Can have read, write, execute, valid bits.
- Valid bit says page isn't in address space.
- Write to a write-protected page causes a fault. Touching an invalid page causes a fault.

ADDRESS MAPPING:

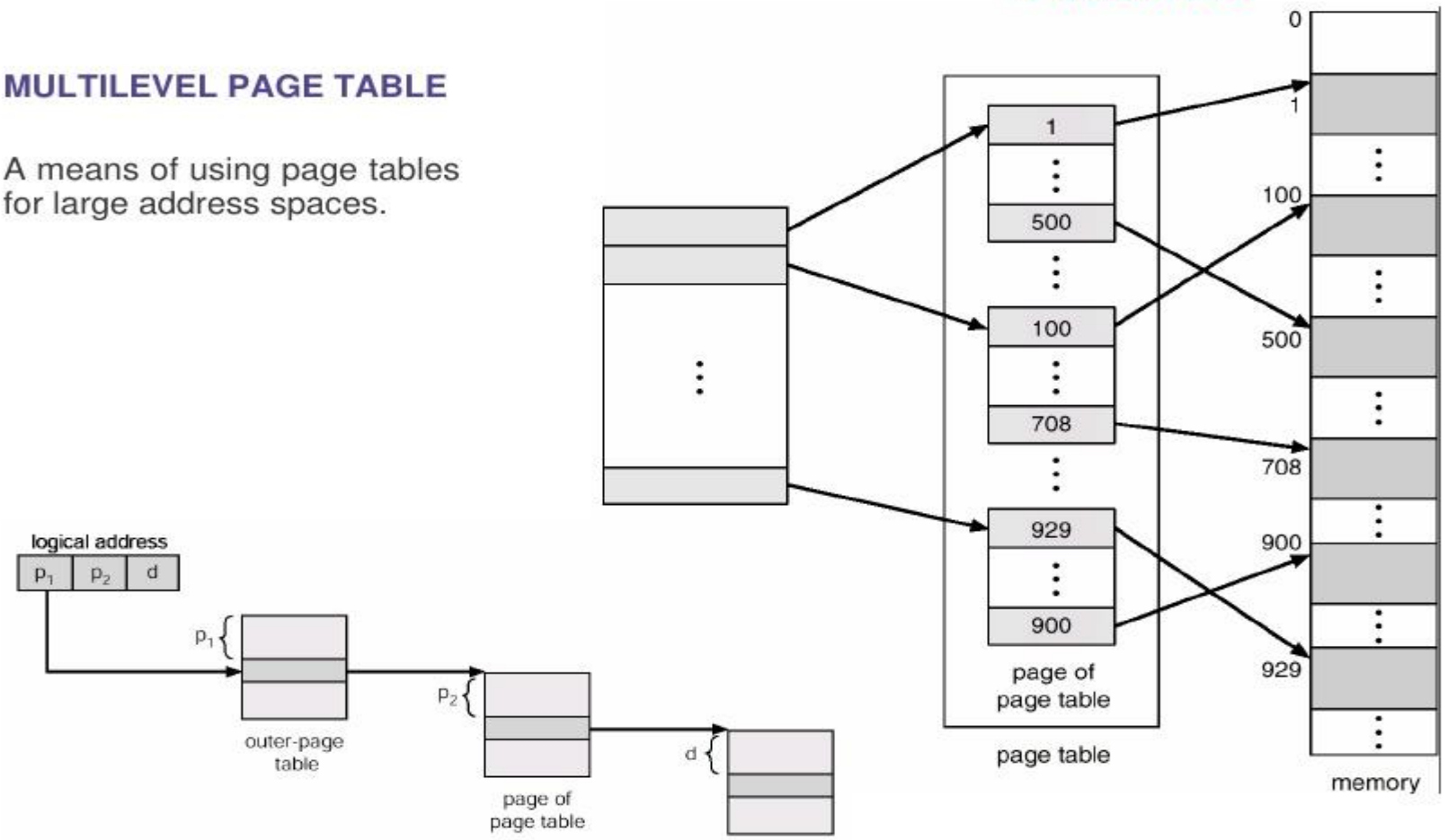
- Allows physical memory larger than logical memory.
- Useful on 32 bit machines with more than 32-bit addressable words of memory.
- The operating system keeps a frame containing descriptions of physical pages; if allocated, then to which logical page in which process.

MEMORY MANAGEMENT

PAGING

MULTILEVEL PAGE TABLE

A means of using page tables for large address spaces.



THANKS