MODULE 5

Module V (8 Hours)

The Memory System – basic concepts, semiconductor RAM memories - organization – static and dynamic RAM, Structure of larger memories, semiconductor ROM memories, Speed, Size and cost, Cache memory – mapping functions – replacement algorithms , Virtual memory – paging and segmentation.

- Hamacher, Vranesic & Zaky, "Computer Organization" (5th Ed), McGraw Hill.

Read-Only Memory (ROM)

- Both static and dynamic RAM chips are volatile, which means that they retain information only while power is turned on.
- There are many applications requiring memory devices that retain the stored information when power is turned off.
- Many embedded applications do not use a hard disk and require non-volatile memories to store their software.
- Non-volatile memory's contents can be read in the same way as for their volatile counterparts.
- A special writing process is needed to place the information into a non-volatile memory.
- Since its normal operation involves only reading the stored data, a memory of this type is called a read-only memory (ROM).

Read-Only Memory (ROM)

- A logic value 0 is stored in the cell if the transistor is connected to ground at point P; otherwise, a 1 is stored.
- The bit line is connected through a resistor to the power supply.
- To read the state of the cell, the word line is activated to close the transistor switch.
- As a result, the voltage on the bit line drops to near zero if there is a connection between the transistor and ground.

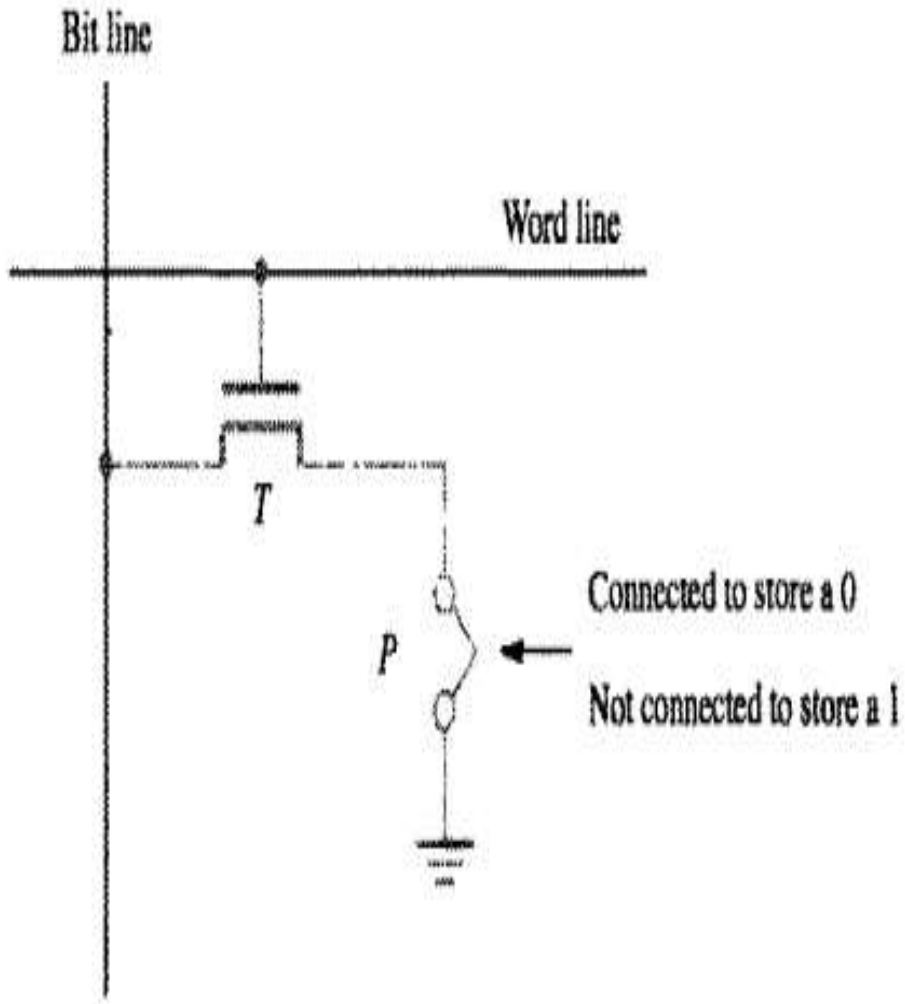

Figure 9 A ROM cell.

[Hamacher, Vranesic & Zaky, "Computer Organization" (5th Ed), McGraw Hill]

Read-Only Memory (ROM)

- If there is no connection to ground, the bit line remains at the high voltage level, indicating a 1.
- A sense circuit at the end of the bit line generates the proper output value.
- The state of the connection to ground in each cell is determined when the chip is manufactured, using a mask with a pattern that represents the information to be stored.

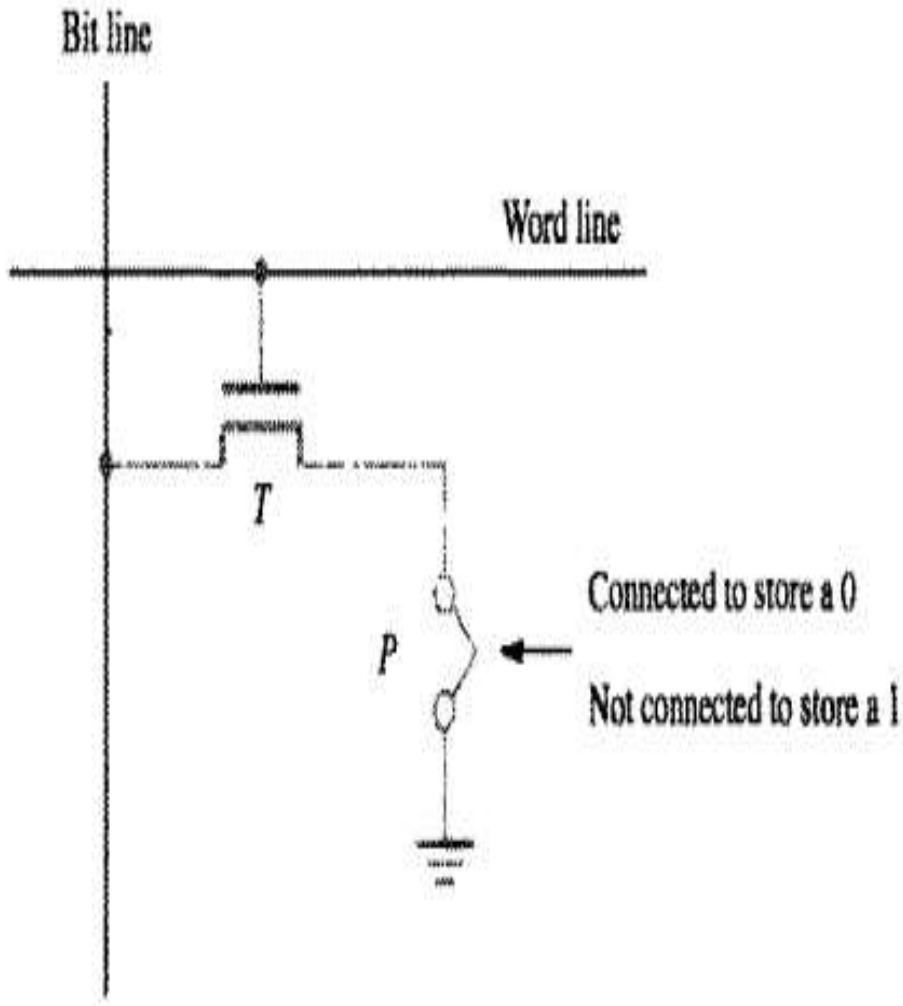

Figure 9 A ROM cell.

[Hamacher, Vranesic & Zaky, "Computer Organization" (5th Ed), McGraw Hill]

Read-Only Memory (ROM)

PROM

- Some ROM designs allow the data to be loaded by the user and thus providing a programmable ROM (PROM).
- Programmability is achieved by inserting a fuse at point P, in Figure.
- Before it is programmed, the memory contains all 0s. The user can insert 1s at the required locations by burning out the fuses at these locations using high-current pulses.
- Of course, this process is irreversible.
- PROMs provide flexibility and convenience not available with ROMs.
- The cost of preparing the masks needed for storing a particular information pattern makes ROMs cost effective only in large volumes.

Read-Only Memory (ROM)

EPROM

- Another type of ROM allows the stored data to be erased and new data to be written into it.
- Such an erasable, reprogrammable ROM is usually called an EPROM.
- It provides considerable flexibility during the development phase of digital systems.
- Since EPROMs are capable of retaining stored information for a long time, they can be used in place of ROMs or PROMs while software is being developed.
- In this way, memory changes and updates can be easily made.
- An EPROM cell has a structure similar to the ROM cell.
- However, the connection to ground at point P is made through a special transistor.

Read-Only Memory (ROM)

EEPROM

- An EPROM must be physically removed from the circuit for reprogramming. Also, the stored information cannot be erased selectively. The entire contents of the chip are erased when exposed to ultraviolet light.
- Another type of erasable PROM can be programmed, erased, and reprogrammed electrically. Such a chip is called an electrically erasable PROM, or EEPROM.
- It does not have to be removed for erasure.
- Moreover, it is possible to erase the cell contents selectively.
- One disadvantage of EEPROMs is that different voltages are needed for erasing, writing, and reading the stored data, which increases circuit complexity.

Read-Only Memory (ROM)

Flash Memory

- A flash cell is based on a single transistor controlled by trapped charge, much like an EEPROM cell.
- Also like an EEPROM, it is possible to read the contents of a single cell.
- The key difference is that, in a flash device, it is only possible to write an entire block of cells.
- Prior to writing, the previous contents of the block are erased.
- Flash devices have greater density, which leads to higher capacity and a lower cost per bit.
- They require a single power supply voltage, and consume less power in their operation.

Read-Only Memory (ROM)

Flash Cards

- One way of constructing a larger module is to mount flash chips on a small card.
- Such flash cards have a standard interface that makes them usable in a variety of products.
- A card is simply plugged into a conveniently accessible slot.
- Flash cards with a USB interface are widely used and are commonly known as memory keys

Read-Only Memory (ROM)

Flash Drives

- Larger flash memory modules have been developed to replace hard disk drives, and hence are called flash drives.
- However, the storage capacity of flash drives is significantly lower than hard disks.
- Currently, the capacity of flash drives is on the order of 64 to 128 Gbytes. In contrast, hard disks have capacities exceeding a terabyte.
- Also, disk drives have a very low cost per bit.

SPEED, SIZE AND COST

- An ideal memory would be fast, large, and inexpensive.
- It is clear that a very fast memory can be implemented using static RAM chips.
- But, these chips are not suitable for implementing large memories, because their basic cells are larger and consume more power than dynamic RAM cells.
- Although dynamic memory units with gigabyte capacities can be implemented at a reasonable cost, the affordable size is still small compared to the demands of large programs with voluminous data.
- A solution is provided by using secondary storage, mainly magnetic disks, to provide the required memory space.
- Disks are available at a reasonable cost, and they are used extensively in computer systems. However, they are much slower than semiconductor memory units.

SPEED, SIZE AND COST

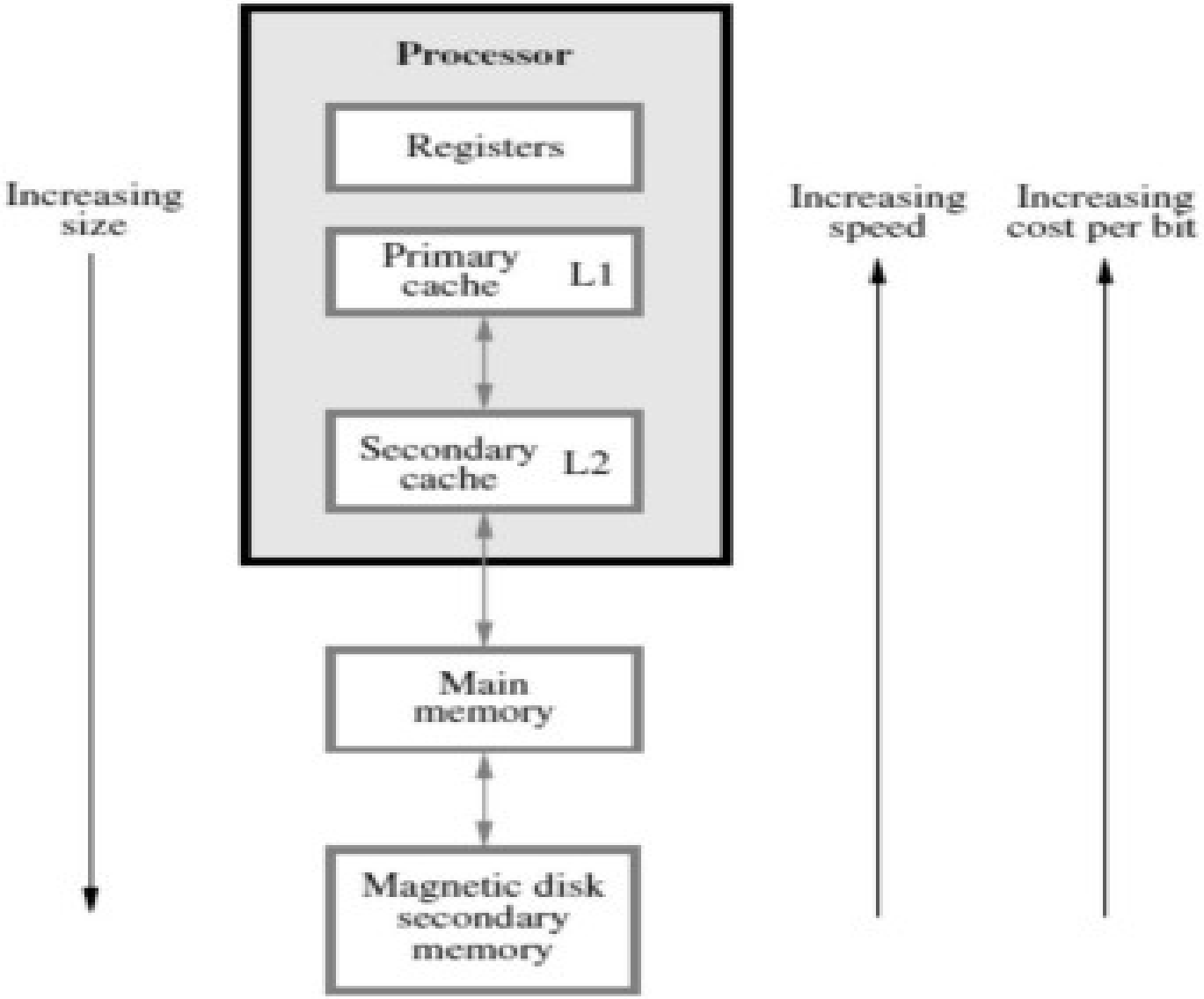

Figure 10: Memory Hierarchy

[Hamacher, Vranesic & Zaky, "Computer Organization" (5th Ed), McGraw Hill]

SPEED, SIZE AND COST

[Hamacher, Vranesic & Zaky, "Computer Organization" (5th Ed), McGraw Hill]

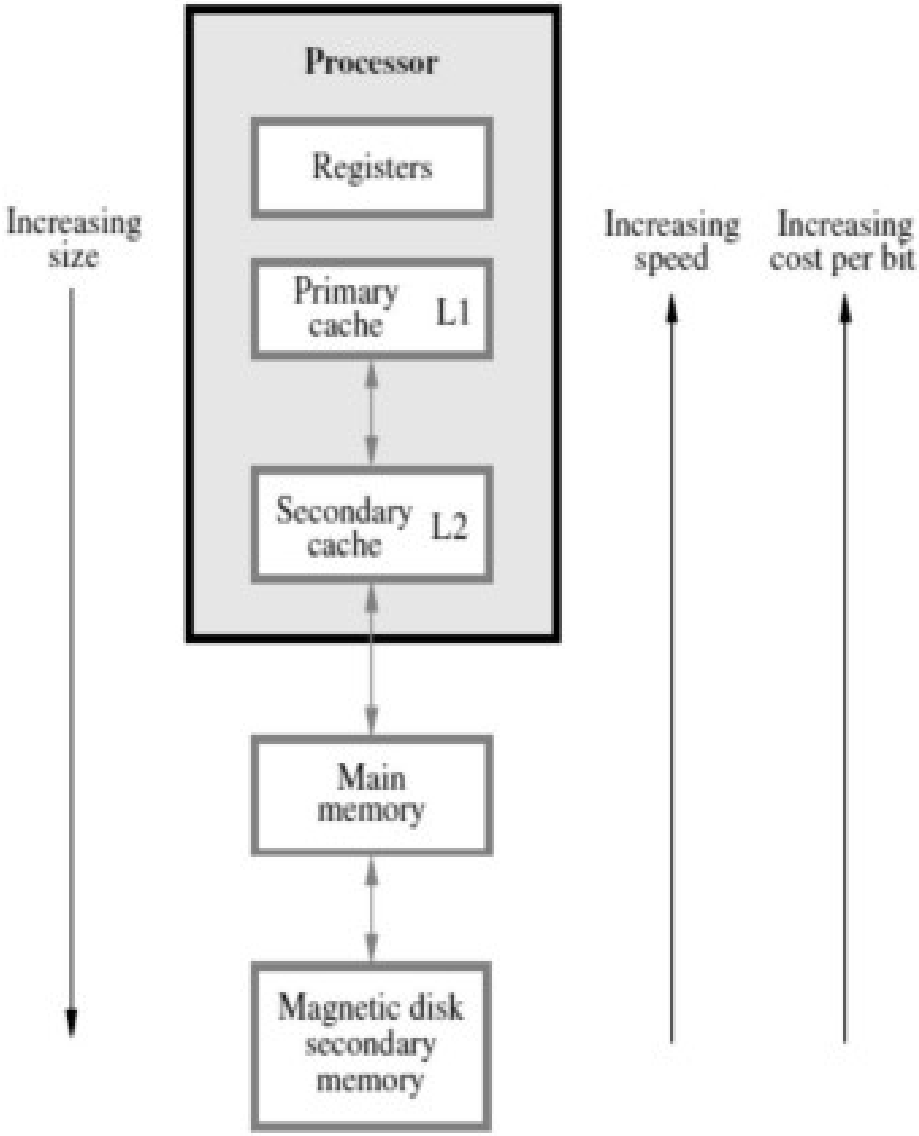

Figure 10: Memory Hierarchy

- Fastest access is to the data held in processor registers. Registers are at the top of the memory hierarchy.
- Relatively small amount of memory that can be implemented directly on the processor chip. This is processor cache.
- Two levels of cache. Level 1 (L1) cache is on the processor chip. Level 2 (L2) cache is in between main memory and processor.
- Next level is main memory, implemented using dynamic memory components. Much larger, but much slower than cache memory.
- Next level is magnetic disks. It provides huge amount of inexpensive storage.
- Speed of memory access is critical, the idea is to bring instructions and data that will be used in the near future as close to the processor as possible.

Module V (8 Hours)

The Memory System – basic concepts, semiconductor RAM memories - organization – static and dynamic RAM, Structure of larger memories, semiconductor ROM memories, Speed, Size and cost, Cache memory – mapping functions – replacement algorithms , Virtual memory – paging and segmentation.

- Hamacher, Vranesic & Zaky, "Computer Organization" (5th Ed), McGraw Hill.

CACHE MEMORIES

- The cache is a small and very fast memory, interposed between the processor and the main memory.
- Its purpose is to make the main memory appear to the processor to be much faster than it actually is.
- The effectiveness of this approach is based on a property of computer programs called locality of reference.
- Analysis of programs shows that most of their execution time is spent in routines in which many instructions are executed repeatedly.
- These instructions may constitute a simple loop, nested loops, or a few procedures that repeatedly call each other.

CACHE MEMORIES

- The actual detailed pattern of instruction sequencing is not important—the point is that many instructions in localized areas of the program are executed repeatedly during some time period.

- This behaviour manifests itself in two ways: temporal and spatial.
- Temporal means that a recently executed instruction is likely to be executed again very soon.
- The spatial aspect means that instructions close to a recently executed instruction are also likely to be executed soon.

CACHE MEMORIES

- Conceptually, operation of a cache memory is very simple.
- The memory control circuitry is designed to take advantage of the property of locality of reference.
- Temporal locality suggests that whenever an information item, instruction or data, is first needed, this item should be brought into the cache, because it is likely to be needed again soon.
- Spatial locality suggests that instead of fetching just one item from the main memory to the cache, it is useful to fetch several items that are located at adjacent addresses as well.
- The term cache block refers to a set of contiguous address locations of some size.
- Another term that is often used to refer to a cache block is a cache line.

CACHE MEMORIES

- When the processor issues a Read request, the contents of a block of memory words containing the location specified are transferred into the cache one word at a time.
- Subsequently, when the program references any of the locations in this block, the desired contents are read directly from the cache.
- Usually, the cache memory can store a reasonable number of blocks at any given time, but this number is small compared to the total number of blocks in the main memory.

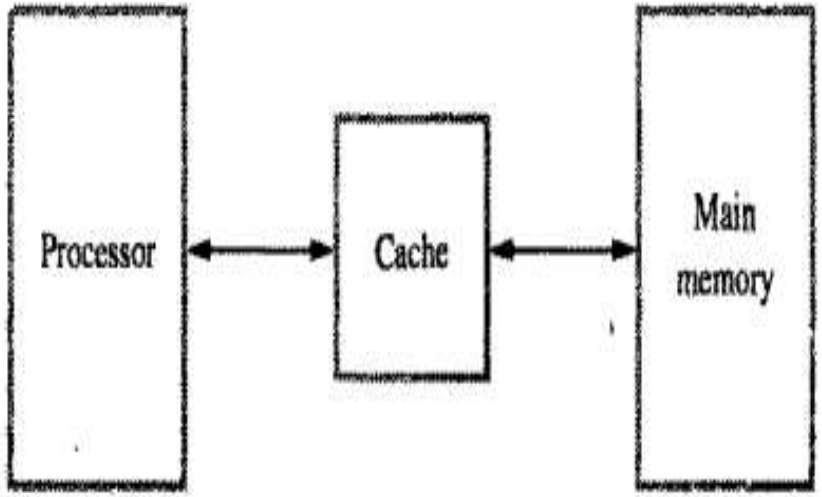

Figure 11 Use of a cache memory.

[Hamacher, Vranesic & Zaky, "Computer Organization" (5th Ed), McGraw Hill]

CACHE MEMORIES

- The correspondence between the main memory blocks and those in the cache is specified by a mapping function.
- When the cache is full and a memory word (instruction or data) that is not in the cache is referenced, the cache control hardware must decide which block should be removed to create space for the new block that contains the referenced word.
- The collection of rules for making this decision constitutes the cache's replacement algorithm.

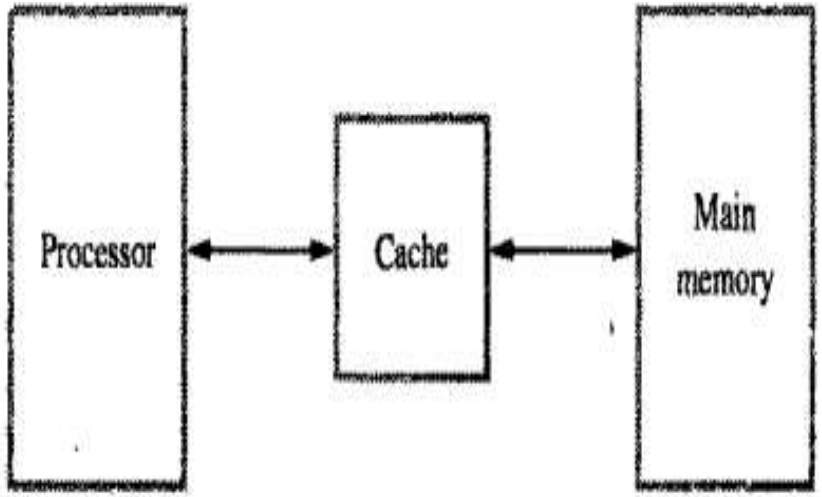

Figure 11 Use of a cache memory.

[Hamacher, Vranesic & Zaky, "Computer Organization" (5th Ed), McGraw Hill]

CACHE MEMORIES

- The processor does not need to know explicitly about the existence of the cache.
- It simply issues Read and Write requests using addresses that refer to locations in the memory.
- The cache control circuitry determines whether the requested word currently exists in the cache.
- If it does, the Read or Write operation is performed on the appropriate cache location. In this case, a read or write hit is said to have occurred.
- The main memory is not involved when there is a cache hit in a Read operation.

CACHE MEMORIES

- For a Write operation, the system can proceed in one of two ways, write-through or write-back protocol.
- In the first technique, called the write-through protocol, both the cache location and the main memory location are updated.
- The second technique is to update only the cache location and to mark the block containing it with an associated flag bit, often called the dirty or modified bit.
- The main memory location of the word is updated later, when the block containing this marked word is removed from the cache to make room for a new block. This technique is known as the write-back, or copy-back, protocol.

CACHE MEMORIES

- The write-through protocol is simpler than the write-back protocol, but it results in unnecessary Write operations in the main memory when a given cache word is updated several times during its cache residency.

- The write-back protocol also involves unnecessary Write operations, because all words of the block are eventually written back, even if only a single word has been changed while the block was in the cache.

CACHE MEMORIES

- A Read operation for a word that is not in the cache constitutes a Read miss.
- It causes the block of words containing the requested word to be copied from the main memory into the cache.
- After the entire block is loaded into the cache, the particular word requested is forwarded to the processor.
- Alternatively, this word may be sent to the processor as soon as it is read from the main memory.
- The latter approach, which is called load-through, or early restart, reduces the processor's waiting time somewhat, at the expense of more complex circuitry.

CACHE MEMORIES

- During a Write operation, if the addressed word is not in the cache, a write miss occurs.
- When a Write miss occurs in a computer that uses the write-through protocol, the information is written directly into the main memory.
- For the write-back protocol, the block containing the addressed word is first brought into the cache, and then the desired word in the cache is overwritten with the new information.