1.Introduction

There are different definition of computer architecture is built on four basic viewpoints.
These are the structure, the organization, the implementation, and the performance.
The structure define is the interconnection of various hardware components.
The organization defines the dynamic interplay and management of the various components. The implementation defines the detailed design of hardware components, and the performance specifies the behavior of the computer system.
Figure : Computer System Component 2

2.Memory Management

2-1-Types of Memory

A- Main Memory

The program must be in main memory to be executed. Main memory is the only large storage area that the processor can be access directly . Its an array of words or bytes, ranging in size from hundreds of thousands to hundreds of millions. Memory is the central to the operation of modern computer system. The CPU fetches instructions from memory according to value of the program counter. These instruction may cause additional loading from and storing to specific memory addresses. For example first fetches an instruction from memory. The instruction is then decoded and may cause operands to be fetches from memory . after the instruction has been executed on the operands, results may be stored back in memory.
Ideally, we would want the programs and data to reside in main memory permanently . this arrangement is not possible for the following two reasons:
1- Main memory is usually too small to store all needed programs and data permanently.
2- Main memory is volatile storage device that loses its components when power is turned off or otherwise lost.
Thus, must computer systems provide secondary storage as extension of main memory. The main requirement for secondary storage is that it be able to hold large quantities of data permanently .
Note:- memory unit sees only as a stream of memory addresses; it does not know how they are generated.

B-Virtual Memory 3

Is a technique that allows the execution of processes that may not be completely in main memory. The main visible advantage of this scheme is that programs can be larger than physical memory. Further, its abstracts
main memory into an extremely large, uniform array of storage, separating logical memory as viewed by the user from physical memory. This technique frees programmers from concern over memory storage limitations .
The set of all logical addresses generated by a program is referred to as a
logical or virtual addresses, while the real addresses in the main memory that corresponding to the logical addresses called physical addresses. Each logical address will be converted to physical address using MMU
(Memory Management Unit), required hardware support by adding relocation register contain value loading to it from the operating system,
where each logical address will be added to the relocation register for generating corresponding physical address in memory.
Figure: Dynamic relocation using relocation register.

A typical memory hierarchy starts with a small, expensive, and relatively
fast unit, called the cache, followed by a larger, less expensive, and relatively slow main memory unit.
Figure: Typical memory hierarchy
Figure: Major differences between cache –main memory and mainsecondary memory hierarchies 5

Cache Memory

A cache is a small, fast memory located close to the CPU that holds the most recently accessed code or data. When the CPU finds a requested data item in the cache, it is called a cache hit. When the CPU does not find a data item it needs in the cache, a cache miss occurs. A fixed-size block of data, called a block, containing the requested word is retrieved from the main memory and placed into the cache. Temporal locality tells us that we are likely to need this word again in the near future, so placing it in the cache where it can be accessed quickly is useful. Because of spatial locality, there is high probability that the other data in the block will be needed soon.

The time required for the cache miss depends on both the latency of the memory and its bandwidth, which determines the time to retrieve the entire block. A cache miss, which is handled by hardware, usually causes the CPU to pause, or stall, until the data are available.

Likewise, not all objects referenced by a program need to reside in main memory.

The following figure shows the principal components of a cache memory.

Words are stored in a cache data memory and are grouped into small page called cachevblocks or block frames or lines. 6

The content of the cache‟s data memory are copies of a set of main memory blocks. Each cache block is marked with its block address, referred to as a tag, so the cache knows to what part of the memory space the block belongs . The collection of tag addresses currently assigned to the cache , is stored in cache tag memory or directory which implemented as associative memory. For example , if bock Bj containing data entries Dj is assign to M1 , then Bj is in the tag memory and Dj is in the cache‟s data memory.

The probability of finding the requested item in the first level is called the hit ratio, h1. The probability of not finding (missing) the requested item
in the first level of the memory hierarchy is called the miss ratio,
h2=(1-h1).