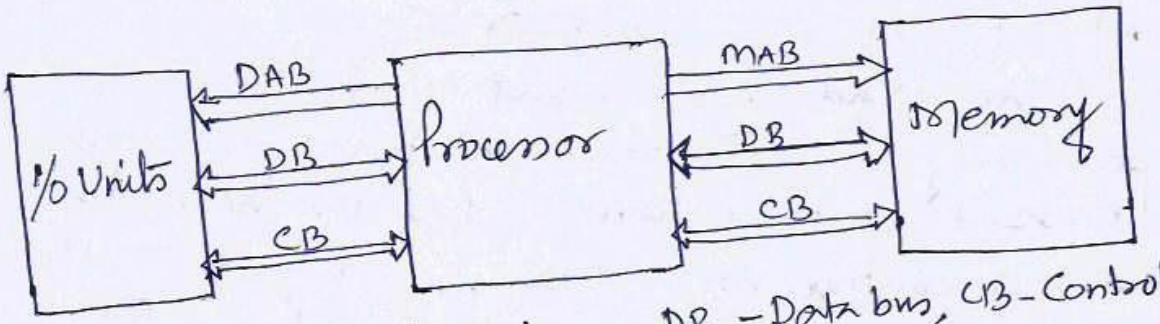
COMPUTER ARCHITECTURE

T Dey
Dept of Comp Sc
KRC

MEMORY ORGANIZATION

Illustrating interconnection of computer units via two system buses.



DAB - Device address bus, DB - Data bus, CB - Control bus
MAB - Memory address bus

A set of wires which carries a group of bits in parallel and has an associated control scheme is known as a bus.

A bus which carries a word to or from memory is known as a data bus.

In order to retrieve a word from memory it is necessary to specify its address. The address is carried by a Memory address bus whose width equals the number of bits in the MAR of the memory. Thus if a computer's memory has 64K, 32bit words, then the data bus will be 32 bits wide and address bus 16 bits wide. Besides buses to carry address and data, we also need control signals between the units of a computer.

For instance, if the processor has to send READ and WRITE commands to memory, START command to I/O units etc. Such signals are carried by a control bus. A system bus will thus consist of a data bus, a memory address bus and a control bus

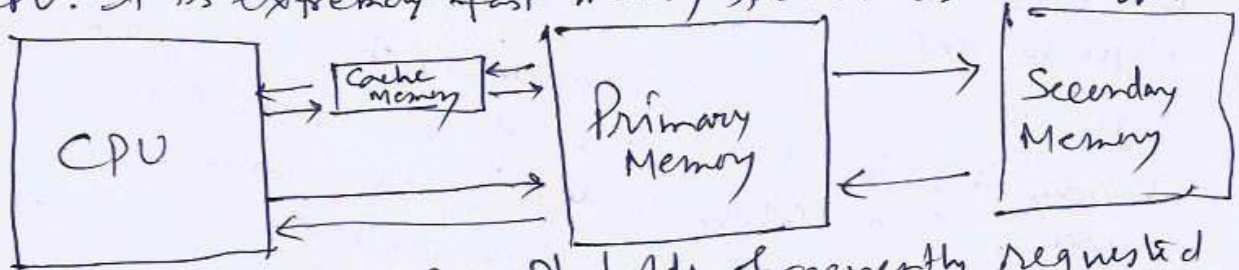
2/8

One method of connecting I/O units to the computer is to connect them to the processor via a bus. This bus will consist of a bus carrying a word from the addressed input unit to the processor or carrying a word from the addressed input unit to the processor or carrying a word from the processor to the addressed output unit. Besides these two buses, a control bus carries commands such as READ, WRITE, START & STOP etc. from the processor to I/O units. It also carries the I/O units' status information to the processor.

The interconnection of I/O units, processor and memory using two independent system buses is known as a two bus interconnection structure.

— — — — — — —

Cache memory is a special very high speed memory. It is used to speed up and synchronizing with high speed CPU. It is extremely fast memory type that acts as a buffer



between RAM & CPU. It holds frequently requested data and instruction so that they are immediately available to the CPU when needed. It is used to reduce the average time to access data from the Main-Memory. It is smaller & faster memory which stores copies of the data frequently used in main memory locations. There are different types of Caches in a CPU which store instructions and data.

23/08

PROCESSOR TO MEMORY COMMUNICATION

The following sequence of events takes place when information is to be transferred from the memory to the processor.

Step 1: The processor places the address in MAR via the memory address bus.

Step 2: The processor issues a READ command via the control bus.

Step 3: The memory places the retrieved data in MDR and transfers it via the data bus to the processor. Based on the read time of the memory, a specific number of processor clock intervals are allotted for completion of this operation. During this interval, the processor is forced to wait

For writing in memory, the steps are similar

Step 1: The processor places the address in MAR via the memory address bus.

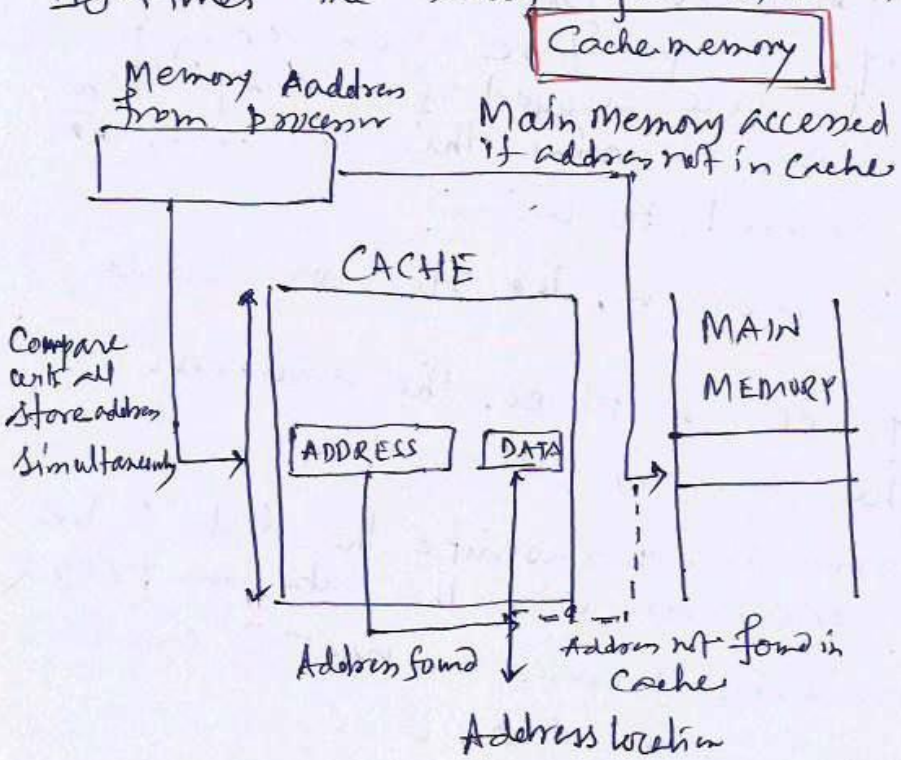
Step 2: The processor transmits the data to be written in memory via the data bus to MDR

Step 3: The processor issues a WRITE command to memory via the control bus.

Step 4: The data in MDR is written in memory in the address specified in MAR.

4/8

The main problem is the 1 to 10 speed mismatch between the processor and the memory. Thus the processor is forced to wait for data or instructions from memory. This speed mismatch is alleviated by using a small fast memory as an intermediate buffer between the main memory and the processor as shown in the figure C. This memory is known as a cache memory. The effective cycle time of a cache would be about a tenth of the main memory cycle time and its cost about 10 times the cost per byte of main memory.



A diagram of the architecture and data flow of a typical cache memory unit

5/8

CACHE MEMORY

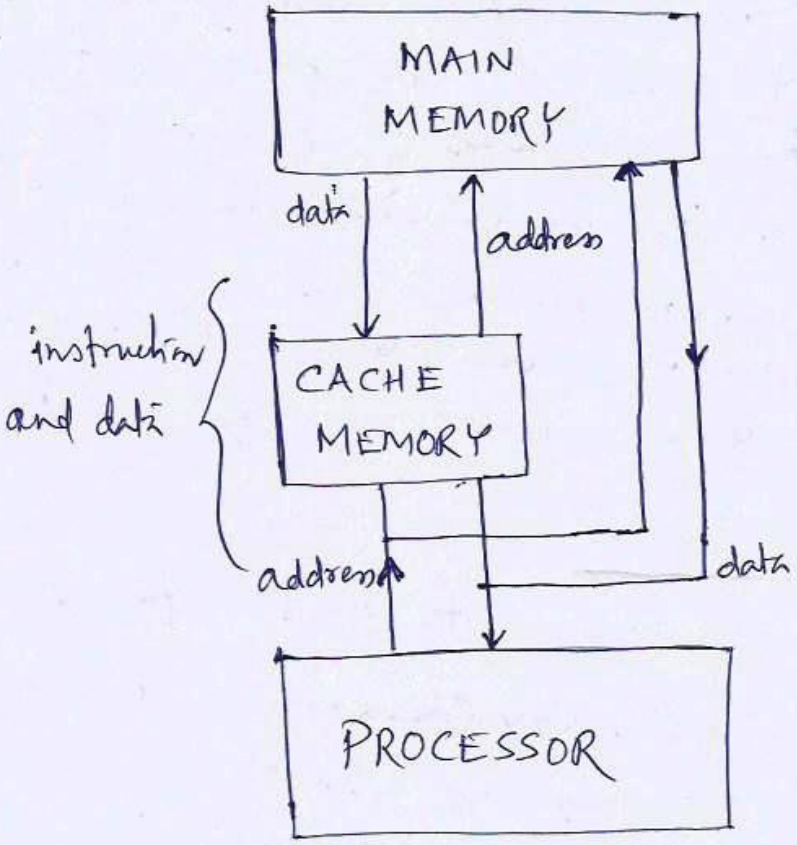


Figure - C

A CPU cache is a hardware cache used by the central processing unit (CPU) of a computer to reduce the average cost (time or energy) to access data from the main memory. It is fabricated using high speed semiconductor devices. A cache is a smaller, fast memory located closer to a processor core which stores copies of the data from frequently used main memory locations. It is faster than main memory. It consumes less access time as compared to main memory. It stores the program that can be executed within a short period of time.

Contd..

6/8

Cache memory is a chip-based computer ~~Computent~~ Component that makes retrieving data from the computer's memory more-efficient. It acts as a temporary storage area that the computer's processor can retrieve data from easily. This ~~temper-~~ temporary storage known as a cache. is more readily available to the processor than the computer's main memory source.

Types of Cache memory

Cache memory is fast and expensive.

L1 cache, or primary cache, is extremely fast but relatively small, and is usually embedded in the processor chip as CPU cache

L2 cache, or secondary cache, is often more capacious than L1, L2 cache may be embedded on the CPU, or it can be on a separate chip or coprocessor and have a high speed alternative system bus connecting the Cache and CPU.

Level 3 (L3) cache is specialized memory developed to improve the performance of L1 and L2. L1 or L2 can be significantly faster than L3, though L3 is usually double the speed of DRAM.

*disclaimer :- Information are collected through net and ref. book etc.

28/

Cache memory mapping

The basic characteristics of Cache memory is its fast access time. Therefore, very little or no time must be wasted when searching for words in the Cache. The transformation of data from main memory to a cache memory is referred to as a mapping process.

Cache memory traditionally works under three different configurations.

1. Direct mapping
2. Associative mapping
3. Set-associative mapping.

1. Direct mapping cache has each block mapped to exactly one cache memory location. Conceptually a direct mapped Cache is like rows in a table with three Columns. the cache block that contains the actual data fetched and stored, a tag with all or part of the address of the data that was fetched, and a flag bit that shows the presence in the row entry of a valid bit of data. * Direct mapping maps each block of main memory into only one possible Cache line or assign each memory block to a specific line in the cache.

8/8 Direct mapping's performance is directly proportional to Hit ratio.

2. Fully associative cache mapping is similar to direct mapping in structure but allows a ~~memory~~ memory block to be mapped to any cache location rather than to a prespecified cache memory location as is the case with direct mapping. It is considered to be the fastest and the most flexible mapping form.

3. Set associative cache mapping can be viewed as a compromise between direct mapping and fully associative mapping in which each block is mapped to a subset of cache locations. It is sometimes called N-way set associative mapping, which provides for a location in main memory to be cached to any of "N" locations in the L1 Cache. This form of mapping is an enhanced form of direct mapping where the drawbacks of direct mapping are removed. In this case, the cache consists of a number of sets, each of which consists of a number of lines.

Hit ratio:

The performance of cache memory is frequently measured in terms of quantity called Hit ratio.

Hit ratio = hit/(hit+miss) = no. of hits/total accesses

accesses.