Memory Devices

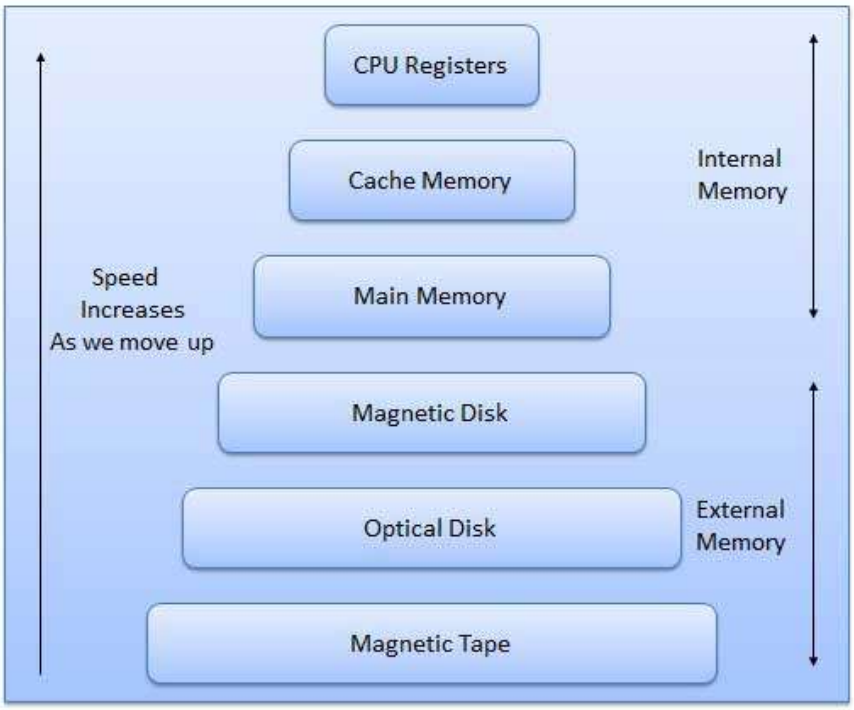
Memory is just like a human brain. It is used to store data and instruction. Computer memory is the storage space in computer where data is to be processed and instructions required for processing are stored.

The memory is divided into large number of small parts. Each part is called cell. Each location or cell has a unique address which varies from zero to memory size minus one.

For example if computer has 64k words, then this memory unit has 64 * 1024=65536 memory location. The address of these locations varies from 0 to 65535.

Memory is primarily of two types

- **Internal Memory** - cache memory and primary/main memory
- **External Memory** - magnetic disk / optical disk etc.



Characteristics of Memory Hierarchy are following when we go from top to bottom.

- Capacity in terms of storage increases.
- Cost per bit of storage decreases.
- Frequency of access of the memory by the CPU decreases.
- Access time by the CPU increases

Prepared by:- Anwar Bari

RAM

A RAM constitutes the internal memory of the CPU for storing data, program and program result. It is read/write memory. It is called random access memory (RAM).

Since access time in RAM is independent of the address to the word that is, each storage location inside the memory is as easy to reach as other location & takes the same amount of time. We can reach into the memory at random & extremely fast but can also be quite expensive.

RAM is volatile, i.e. data stored in it is lost when we switch off the computer or if there is a power failure. Hence a backup uninterruptible power system(UPS) is often used with computers. RAM is small , both in terms of its physical size and in the amount of data it can hold.

RAM is of two types

- Static RAM (SRAM)
- Dynamic RAM (DRAM)

Static RAM (SRAM)

The word **static** indicates that the memory retains its contents as long as power remains applied. However, data is lost when the power gets down due to volatile nature. SRAM chips use a matrix of 6-transistors and no capacitors. Transistors do not require power to prevent leakage, so SRAM need not have to be refreshed on a regular basis.

Because of the extra space in the matrix, SRAM uses more chips than DRAM for the same amount of storage space, thus making the manufacturing costs higher.

Static RAM is used as cache memory needs to be very fast and small.

Dynamic RAM (DRAM)

DRAM, unlike SRAM, must be continually **refreshed** in order for it to maintain the data. This is done by placing the memory on a refresh circuit that rewrites the data several hundred times per second. DRAM is used for most system memory because it is cheap and small. All DRAMs are made up of memory cells. These cells are composed of one capacitor and one transistor.

ROM

ROM stands for Read Only Memory. The memory from which we can only read but cannot write on it. This type of memory is non-volatile. The information is stored permanently in such memories during manufacture.

A ROM, stores such instruction as are required to start computer when electricity is first turned on, this operation is referred to as bootstrap. ROM chip are not only used in the computer but also in other electronic items like washing machine and microwave oven.

Following are the varioys types of ROM

MROM (Masked ROM)

The very first ROMs were hard-wired devices that contained a pre-programmed set of data or instructions. These kind of ROMs are known as masked ROMs. It is inexpensive ROM.

PROM (Programmable Read only Memory)

PROM is read-only memory that can be modified only once by a user. The user buys a blank PROM and enters the desired contents using a PROM programmer.Inside the PROM chip there are small fuses which are burnt open during programming. It can be programmed only once and is not erasable.

EPROM(Erasable and Programmable Read Only Memory)

The EPROM can be erased by exposing it to ultra-violet light for a duration of upto 40 minutes. Usually, a EPROM eraser achieves this function. During programming an electrical charge is trapped in an insulated gate region. The charge is retained for more than ten years because the charge has no leakage path. For erasing this charge, ultra-violet light is passed through a quartz crystal window(lid). This exposure to ultra-violet light dissipates the charge. During normal use the quartz lid is sealed with a sticker.

EEPROM(Electrically Erasable and Programmable Read Only Memory)

The EEPROM is programmed and erased electrically. It can be erased and reprogrammed about ten thousand times. Both erasing and programming take about 4 to 10 ms (milli second). In EEPROM, any location can be selectively erased and programmed. EEPROMs can be erased one byte at a time, rather than erasing the entire chip. Hence, the process of re-programming is flexible but slow.

Serial Access Memory

Sequential access means the system must search the storage device from the beginning of the memory address until it finds the required piece of data. Memory device which supports such access is called a Sequential Access Memory or Serial Access Memory. Magnetic tape is an example of serial access memory.

Direct Access Memory

Direct access memory or Random Access Memory, refers to condition in which a system can go directly to the information that the user wants. Memory device which supports such access is called a Direct Access Memory. Magnetic disk, optical disks are an examples of direct access memory.

Cache Memory

Cache memory is a very high speed semiconductor memory which can speed up CPU. It acts as a buffer between the CPU and main memory. It is used to hold those parts of data and program which are most frequently used by CPU. The parts of data and programs are transferred from disk to cache memory by operating system, from where CPU can access them.

Advantages

- Cache memory 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.
- It stores data for temporary use.

Disadvantages

- Cache memory has limited capacity.
- It is very expensive.

Virtual memory is a technique that allows the execution of processes which are not completely available in memory. The main visible advantage of this scheme is that programs can be larger than physical memory. Virtual memory is the separation of user logical memory from physical memory.

This separation allows an extremely large virtual memory to be provided for programmers when only a smaller physical memory is available. Following are the situations, when entire program is not required to be loaded fully in main memory.

- User written error handling routines are used only when an error occured in the data or computation.
- Certain options and features of a program may be used rarely.
- Many tables are assigned a fixed amount of address space even though only a small amount of the table is actually used.
- The ability to execute a program that is only partially in memory would counter many benefits.
- Less number of I/O would be needed to load or swap each user program into memory.
- A program would no longer be constrained by the amount of physical memory that is available.
- Each user program could take less physical memory, more programs could be run the same time, with a corresponding increase in CPU utilization and throughput.

Auxiliary Memory

Auxiliary memory is much larger in size than main memory but is slower. It normally stores system programs, instruction and data files. It is also known as secondary memory. It can also be used as an overflow/virtual memory in case the main memory capacity has been exceeded. Secondary memories can not be accessed directly by a processor. First the data / information of auxillary memory is transferred to the main memory and then that information can be accessed by the CPU. Characteristics of Auxiliary Memory are following

- **Non-volatile memory** - Data is not lost when power is cut off.
- **Reusable** - The data stays in the secondary storage on permanent basis until it is not overwritten or deleted by the user.
- **Reliable** - Data in secondary storage is safe because of high physical stability of secondary storage device.
- **Convenience** - With the help of a computer software, authorised people can locate and access the data quickly.
- **Capacity** - Secondary storage can store large volumes of data in sets of multiple disks.
- **Cost** - It is much lesser expensive to store data on a tape or disk than primary memory.

Design of Direct Mapped Cache :

Cache memory is a small (in size) and very fast (zero wait state) memory which sits between the CPU and main memory. The notion of cache memory actually rely on the correlation properties observed in sequences of address references generated by CPU while executing a programm(principle of locality).When a memory request is generated, the request is first presented to the cache memory, and if the cache cannot respond, the request is then presented to main memory.

- **Hit:** a cache access finds data resident in the cache memory
- **Miss:** a cache access does not find data resident, so it forces to access the main memory.

Cache treats main memory as a set of blocks.As the cache size is much smaller than main memory so the number of cache lines are very less than the number of main memory blocks. So a procedure is needed for mapping main memory blocks into cache lines.cache mapping scheme affects cost and performance. There are three methods in block placement-

- **Direct Mapped Cache**
- **Fully Associative Mapped Cache**
- **Set Associative Mapped Cache**

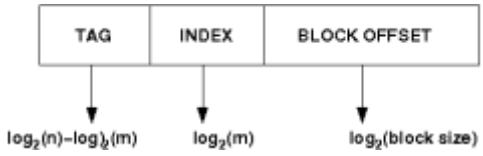
Direct Mapped Cache

A given memory block can be mapped into one and only cache line.

Block identification: let the main memory contains n blocks(which require $\log_2(n)$) and cache contains m blocks, so n/m different blocks of memory can be mapped (at different times) to a cache block. Each cache block has a tag saying which block of memory is currently present in it, each cache block also contain a valid bit to ensure whether a memory block is in the cache block currently.

- **Number of bits in the tag: $\log_2(n/m)$**
- **Number of sets in the Cache: m**
- **Number of bits to identify the correct set: $\log_2(m)$**

The memory address is divided into 3 parts- tag(most MSB), index, block offset(most LSB) in order to do the cache mapping.

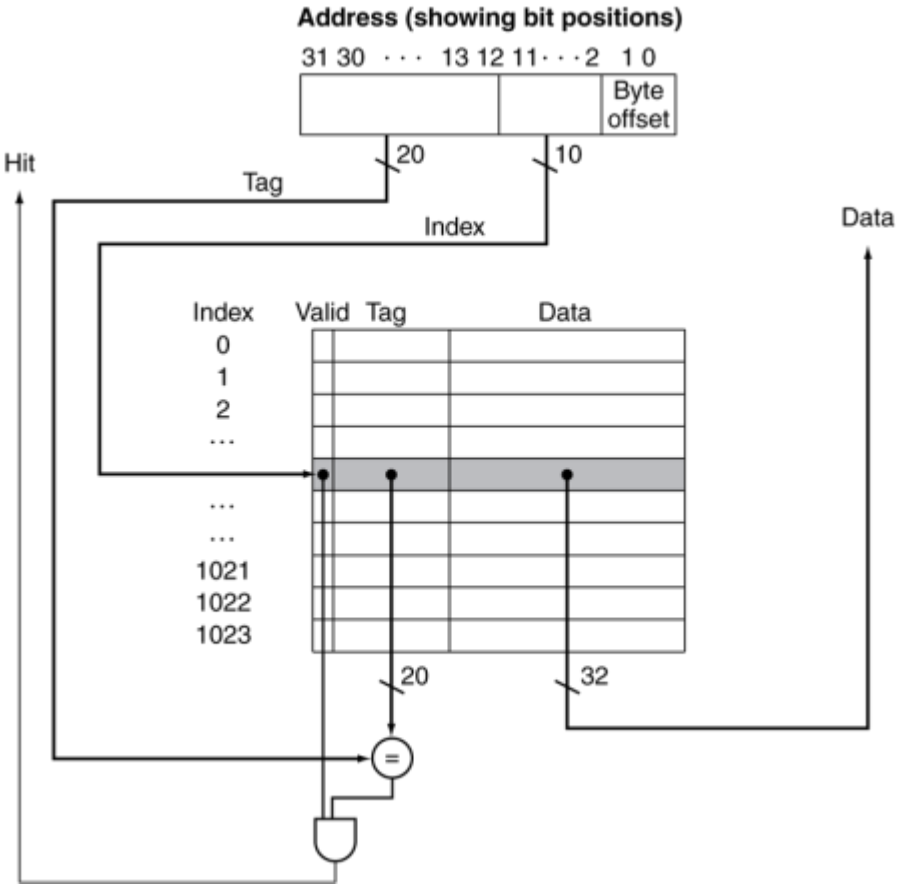


- **Select set using index, block from set using tag.**
- **Select location from block using block offset.**
- **tag + index = block address**

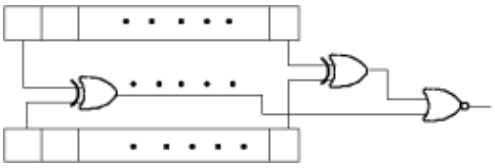
If a miss occur CPU bring the block from the main memory to the cache, if there is no free block in the corresponding set it replaces a block and put the new one. CPU uses different replacement policies to decide which block is to replace. The disadvantage of the direct mapped cache is that it is easy to build, but suffer the most from thrashing due to the 'conflict misses' giving more miss penalty.

Design issues:

Bellow is a simple cache which holds 1024 words or 4KB, memory address is 32 bits. The tag from the cache is compared against the most significant bits of the address to determine whether the entry in the cache corresponds to the requested address as the cache has 2^{10} or 1024 words and a block size of one word, 10 bits are used to index the cache, leaving 32-10-2=20 bits to be compared against the tag. If the tag and the most significant 20 bits of the address are equal and the valid bit is on then the request hits in the cache otherwise miss occurs. No replacement policy has been implemented in the circuit.



The comparator Circuit through which tag is compared with specified bits of address:



Design of Associative Cache :

Cache memory is a small (in size) and very fast (zero wait state) memory which sits between the CPU and main memory. The notion of cache memory actually rely on the correlation properties observed in sequences of address references generated by CPU while executing a programm(principle of locality).When a memory request is generated, the request is first presented to the cache memory, and if the cache cannot respond, the request is then presented to main memory.

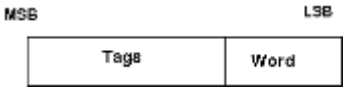
- **Hit:** a cache access finds data resident in the cache memory
- **Miss:** a cache access does not find data resident, so it forces to access the main memory.

Cache treats main memory as a set of blocks.As the cache size is much smaller than main memory so the number of cache lines are very less than the number of main memory blocks. So a procedure is needed for mapping main memory blocks into cache lines.cache mapping scheme affects cost and performance. There are three methods in block placement-

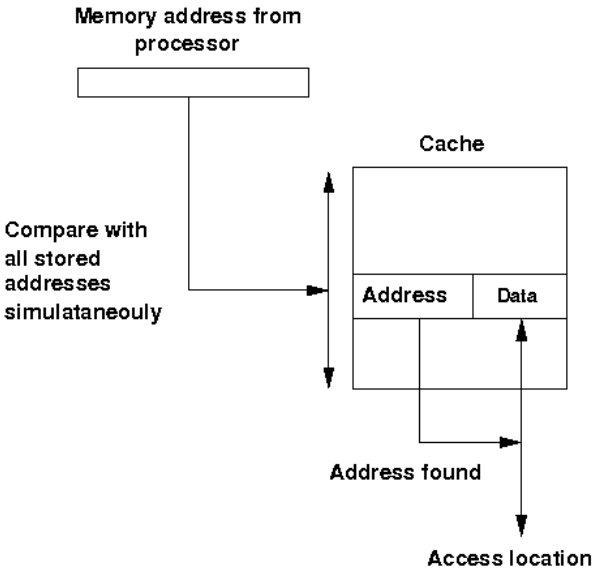
- **Direct Mapped Cache**
- **Fully Associative Mapped Cache**
- **Set Associative Mapped Cache**

Associative Cache

Any main memory block can mapped into any cache line. main memory address is divided into two groups which are tags and word bits. Words are low-order bits and identifies the location of a word within a block and tags are high-order bits which identifies the block.



Block diagram of a associated cache :

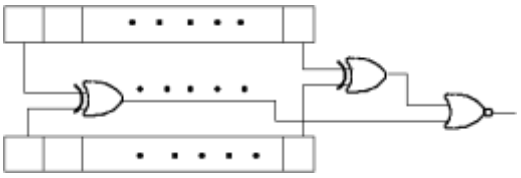


If a miss occur CPU bring the block from the main memory to the cache, if there is no free block in the corresponding set it replaces a block and put the new one. CPU uses different replacement policies to decide which block is to replace. The disadvantage of the associative cache is its high cost for implementing parallel tag comparison, but suffer the most from thrashing due to the 'conflict misses' giving more miss penalty.

Design issues:

No replacement policy has been implemented in the experiment.

The comparator Circuit through which tag is compared with specified bits of address:



virtual memory

In computing, **virtual memory** is a memory management technique that is implemented using both hardware and software. It maps memory addressesused by a program, called *virtual addresses*, into *physical addresses* in computer memory. Main storage as seen by a process or task appears as a contiguous address space or collection of contiguous segments. The operating system manages virtual address spaces and the assignment of real memory to virtual memory. Address translation hardware in the CPU, often referred to as amemory management unit or *MMU*, automatically translates virtual addresses to physical addresses. Software within the operating system may extend these capabilities to provide a virtual address space that can exceed the capacity of real memory and thus reference more memory than is physically present in the computer.

The primary benefits of virtual memory include freeing applications from having to manage a shared memory space, increased security due to memory isolation, and being able to conceptually use more memory than might be physically available