Chapter 4 - Cache Memory

Reading: Section 4.1 Computer Memory System Overview (pp. 101-108)

We have already had brief discussions of memory, in particular, we have discussed memory in terms of location. This is one of the most important aspects of memory as the location has a direct impact on speed and expense.

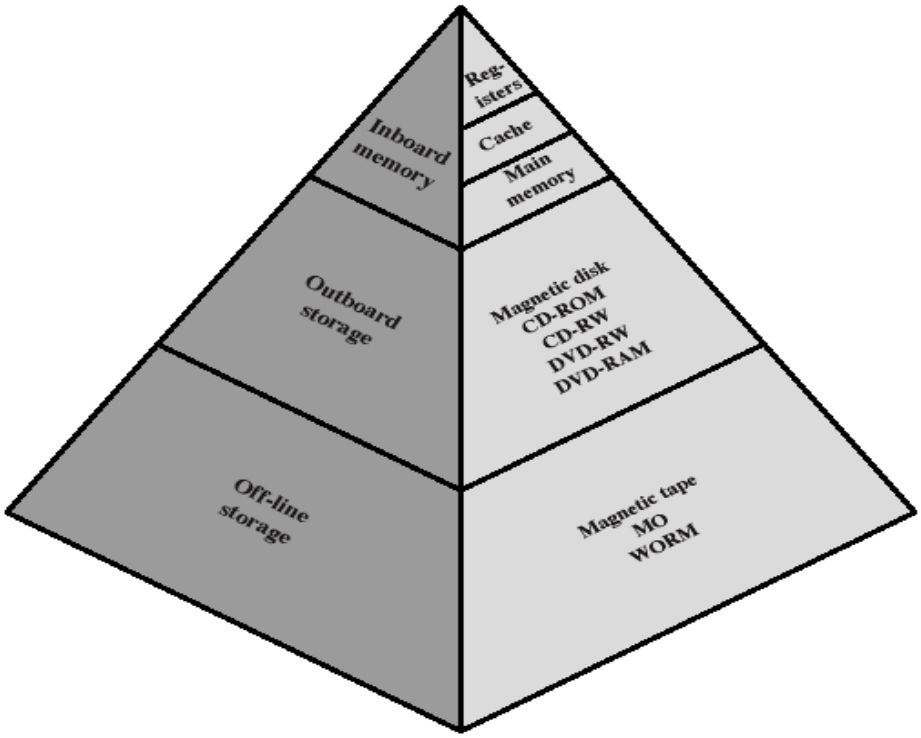

Figure 4.1 The Memory Hierarchy

Before discussing the memory hierarchy in more detail, let's discuss some terminology.

With memory, one of our biggest distinctions is "how is memory accessed"

1. Sequential Access - access through memory proceeds linearly. Typically, data is stored in records and some R/W mechanism is used to move linearly through the data until the desired record is found. (e.g. Tape Drive)

2. Direct Access - Using the same shared R/W mechanism individual records have a unique address and access is accomplished by going to a specific vicinity and then sequentially searching or counting until the final location is reached (e.g. Disk Drive)

3. Random - Each location has a unique address with a physically wired addressing scheme. Access time is constant. (e.g. MEM)

4. Associative - Is random access in nature where certain bits within a word are check and the appropriate location is computed. The word is retrieved from the location based on the bits and several words might be stored at the same location although only one at a time. Retrieval time is constant because accessing memory is still based on the random accessing scheme.

From a users point of view in regard to memory, we are concerned about cost, capacity, and performance. We will examine this in more detail at a later time.

We've already defined the memory hierarchy in this class from the processor out as follows:

```
closest   Registers      Fastest  Low Capacity
   |      Cache             |          |
   |      MEM               |          |
   |      Disk Cache        |          |
   |      Magnetic Disk     |          |
   V      Optical Disk      V          V
farthest  Magnetic Tape  Slowest  High Capacity
```

Q#1: What are we talking for capacity of each of the following:

a) Cache
b) RAM
c) Hard Disk
d) CD-ROM
e) Floppy Disk

From a design point we can take advantage of what is called "locality of reference" to take advantage of our small high speed memory. During program execution, memory references by the processor tend to cluster for data and instructions due to the fact that programs contain a number of functions and loops. This means there is reference quite often to a small set of instructions repeatedly.

P#1: The following are various types of memory. Classify each by category (R/W, Read only, Read mostly), Erasure (Erasable, Non-erasable), Volatility (Volatile, Nonvolatile).

a) Random-access memory (RAM)
b) Read-only memory (ROM)
c) Programmable ROM
d) Erasable PROM

e) Flash Memory
f) Electrically Erasable PROM

Before taking a look at a 16-Mbit DRAM let's briefly discuss the two different RAM technologies.

A "dynamic" RAM or DRAM is made using circuitry that has capacitors which store data as a charge. Presence of a charge is a 1 and absence is a 0. Periodic refreshing is necessary.

A "static" RAM or SRAM is made from flip-flop logic-gate technology. No refreshing is necessary as the feedback mechanism of the FF keeps the value.

Finally, let's look at an 8-bit parallel register.

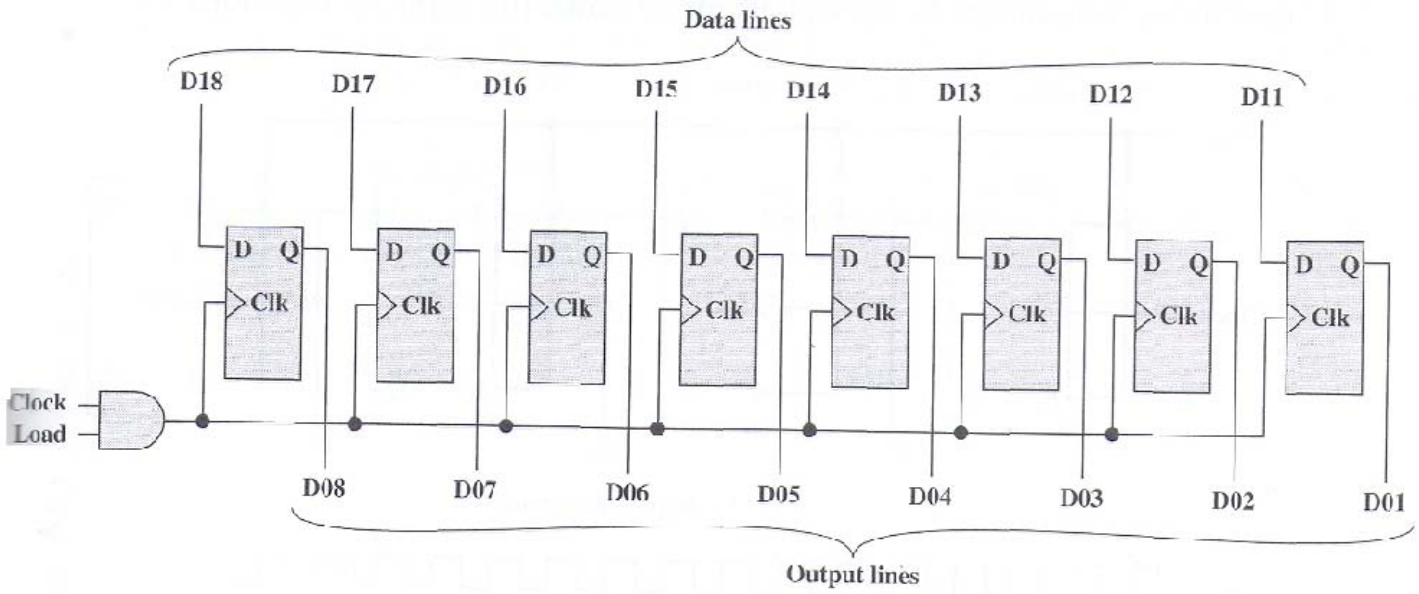

Figure B.30 8-Bit Parallel Register