The Memory Hierarchy

Access Times

1ns – 2ns

3ns – 10ns

25ns – 50ns

30ns – 90ns

5ms – 20ms

100ms - 5s

10s – 3m

Registers

Level 1 Cache

Level 2 Cache

Main Memory

Fixed Rigid Disk

Optical Disk

Magnetic Tape

Terminology with memory:
- *Hit* – The requested data resides in a given level of memory
- *Miss* – The requested data is not found in the given level of memory
- *Hit Rate* – The percentage of memory accesses found in a given level of memory
- *Miss Rate* – Percentage of memory accesses not found in a given level, = 1 – Hit Rate
- *Hit Time* – Time required to access the requested information in a given level

Data request:
Processor sends request to the fastest, smallest memory first (normally the cache). If data found in cache then loaded into CPU, else the request is forwarded to the next level of memory and so on until data is found. Once data found normally a whole block of data is transferred back up into the cache, because a good chance that next data CPU needs was in the same vicinity as the most recently requested.