Chapter 08: The Memory System

Lesson 10:

Memory Access Cycle Time, Latency, Throughput, and Bandwidth

Objective

- Learn computation of access time, latency and bandwidths

Latency, Throughput and bandwidth

Latency and throughput term use in processor pipelines

- Latency— the time taken to complete an individual operation
- Throughput— the rate at which operations can be completed

Latency and throughput term use in memory

- Latency— the time taken to complete an individual memory read or write operation
- Throughput— the rate at which operations can be completed

Bandwidth term use in Memory

- Bandwidth– the total rate at which data can be moved between the processor and memory
- Bandwidth– can be thought of as the product of the throughput and the amount of data referenced by each memory operation

Example

- Assume─ a memory system latency of 10 ns per operation
- Assume─ A data width of 32 bits
- Assume─ only one operation can be performed at a time and there is no delay between operations

Solution for the throughput and bandwidth of the memory system

- Throughput = l/(latency) when operations execute sequentially
- = 1/ (10 ns) system = 100 million operations/s.
- Each operation references 32 bits of data, the bandwidth = 32 ×100 million operations/s = 3.2 billion bits/s, or (3.2 billion/8) = 400 million bytes/s

Actual bandwidth achievable in practice

- Generally much less than the peak bandwidth when running programs on a computer
- No requests going to the memory system at certain times
- Conflicts for memory banks, and other factors at certain times

Summary

We learnt

- Computation of access time, latency and bandwidths

End of Lesson 10 on

Memory Access Cycle Time, Latency, Throughput, and Bandwidth