

Chapter 08: The Memory System

Lesson 13: **Precharging and replication**

Objective

- Learn Precharge and Replication methods

Precharging

Time Between Successive Accesses

- Many memory technologies require that a certain amount of idle time elapse between memory accesses

Precharging Successive Accesses

- This time is used to prepare, or *precharge*, the circuitry for the next access
- Precharging the circuitry does some of the work of accessing the memory before the address arrives
- Reduces the delay from the time that an address is sent to the memory system until a memory operation completes

Precharge at End of Each Operation

- If the memory system is idle much of the time, doing the precharge at the end of each memory operation improves performance because there usually isn't another operation waiting to use the memory system

Case 1: Number of operations/s and Bandwidth

- If the memory system is idle much of the time, doing the precharge at the end of each memory operation the rate at which operations can be completed is determined by the memory latency

Case 2: Number of operations/s and Bandwidth

- If the memory system is being used most of the time, the rate at which operations can be completed is determined by the sum of the memory latency and the precharge time

Example of Case 2

- If a memory system has a bandwidth of 120,000,000 bytes/s, transfers 2 bytes of data per access, and has a precharge time of 5 ns between accesses
- Assume— no pipelining and that the memory system does not have multiple banks

Solution for memory system's latency

- To get a bandwidth of 120,000,000 bytes/s at 2 bytes/access, the memory system must have a throughput of 60,000,000 accesses/s, or one access every 16.7ns
- Since the time between accesses is the sum of the memory latency and the precharge time, the memory latency must be $16.7 \text{ ns} - 5 \text{ ns} = 11.7 \text{ ns}$

Replication

Replicated Memory System

- Provide multiple copies of the entire memory
- This means that each copy of the memory can handle any memory request
- But it increases the amount of memory required by a factor equal to the number of copies

Replicated Memory

- The replicated copy of the memory system does the precharge when original copy is undergoing read operation and vice versa
- Therefore, each memory read-operation the rate at which operations can be completed is determined only by the memory latency, and not memory latency plus precharge time

Replicated Memory System

- To keep the contents of each memory the same, all store operations must be sent each copy of the memory, making stores much more expensive than loads in terms of the amount of bandwidth they consume

Example

- Finding the memory bandwidth of a replicated memory system with 4 memories, where each memory provides a bandwidth of 10 MB/s
- Computing how much memory is required if the system needs to be able to store 32 MB of data

Example

- Computing the bandwidth and total memory required of a banked memory system with 4 banks and all other parameters the same as the first system

Solution

- In both cases, the memory bandwidth will be 40 MB/s, the sum of the bandwidths of the banks/copies
- For the replicated memory system, a total of 128 MB of memory will be required while the banked system will only require 32 MB of storage

Analysis of Result

- The example above illustrates a common hazard when analyzing memory systems
- We calculated the bandwidth of the banked memory system by multiplying the bandwidth of each bank by the number of banks

...Analysis of Result

- But since the requests will sometimes be sent to banks that are already busy, so they will have to wait before they can proceed, When this happens, the actual rate at which data is transferred to or from the memory system will be lower than the calculated bandwidth

...Analysis of Result

- In the extreme case where all memory requests go to the same bank, the system described above could have a bandwidth as low as 10 MB/s

Summary

We learnt

- Precharge method to improve bandwidth
- Replication to improve bandwidth during read

End of Lesson 13 on
Precharging and replication