# Chapter 08: The Memory System

Lesson 12:

**Memory Interleaving**

# Objective

- Learn memory interleaving method for the pipeline like access

# Interleaving (Pipeline)

# Improving Performance by Banking and Interleaving

- Interleaving– for facilitating the simultaneous accesses to the memory system

# Interleaving Like Pipeline

- Memory systems can be pipelined in the same way that processors are pipelined, allowing operations to overlap execution to improve throughput
- Interleaving permits accesses like instructions in a pipeline

# Interleaving Example 1

- Assume that first vertical column activates by chip-enable 0, second by chip-enable 1, and so on
- $A_j$ to $A_{n-1}$ are the address bus signals directly connected to the processor address bus
- A0 and A1 select a column out of the four columns
- Each byte sequentially one after one read or written in interleaving of the four banks

# Example 1 of interleaved memory

- Memory system with a latency of 40 ns that transfers 1 byte per operation
- Pipelined to allow 4 operations to overlap execution (assume no pipelining overhead)

# Solution for bandwidth

- Dividing the latency of 40 ns by the number of overlapped operations (4 gives a rate of 1 operation per 10 ns as the throughput of the memory system
- At 1 byte of data per operation, this gives a bandwidth of $10^8$ bytes/s

# Banking (parallel) plus Interleaving (Pipeline)

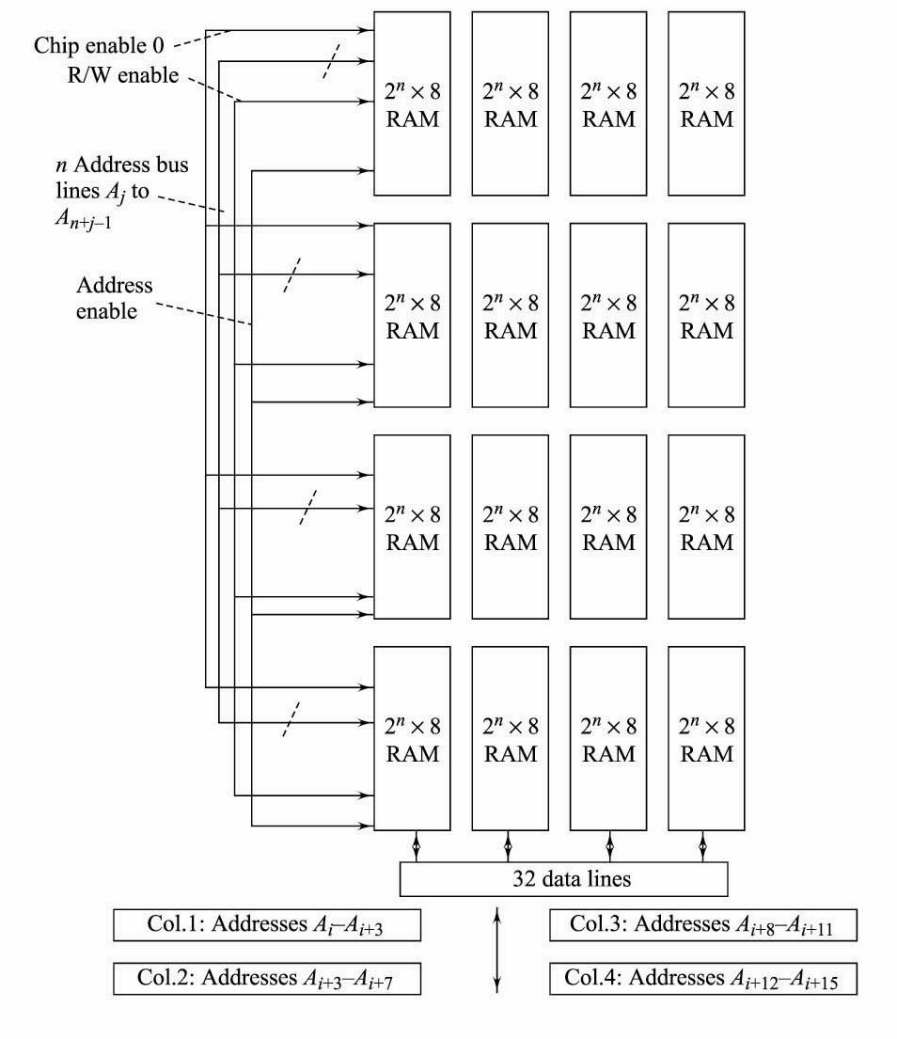
# Banking plus Interleaving

- Instead of two lower address bits two other address bits could be used to select a bank
- Generally, relatively low-order address bits are used to select the bank, so that references to sequential memory addresses go to different banks
- Lowest order bits used for interleaving (pipeline like access)

# Interleaving Example 2

- Interleaving of data of addresses $A_2$ to $A_{j-1}$, and $A_{n+j}$ and $A_{m-1}$ select a vertical column bank when four banks simultaneously parallel accessed
- Each word sequentially one after one read or written using horizontally placed set of 4 RAMs and there is simultaneous access of the four vertical banks

# Interleaving pipelined access from 4 RAMs and parallel access from 4 vertical banks and



# Example 2 banking plus interleaving

- Assume─ A memory system has four banks, each of which has a latency of 100 ns and is pipelined to allow 8 operations to overlap execution
- Each bank returns 4 bytes of data in response to a memory request

# Solution for the peak throughput and peak bandwidth of this system

- Each bank has a latency of 100 ns and can pipeline 8 operations
- Therefore, the throughput of each bank is 1 operation every 100 ns/ 8 = 12.5 ns, or 80,000,000 operations/s
- Since there are 4 banks, the peak throughput of the memory system is 4 × 80,000.000 = 320,000,000 operations/s

# Solution for the peak throughput and peak bandwidth of this system

- With each memory operation returning 4 bytes of data, this gives a peak bandwidth of $4 \times 320{,}000{,}000 = 1{,}280{,}000{,}000$ bytes/s
- Peak bandwidth computations assumes no idle time between one and other subsequent accesses after the first

# Summary

# We learnt

- Memory bank concept for the parallel access
- Memory interleaving method for the pipeline like access

End of Lesson 12 on

**Memory Interleaving**