Shared Memory Organizations

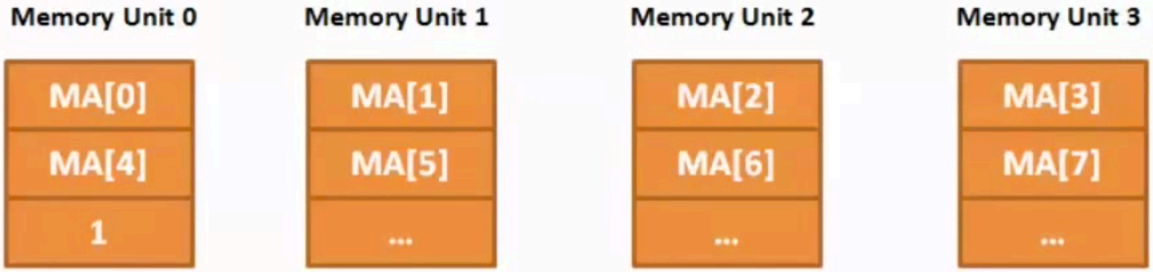
Main issue – Matching the speed of the CPU processing with data access from memory!

This is a bottleneck when working at higher speeds as memory system is unable to cope with CPU processing speeds.

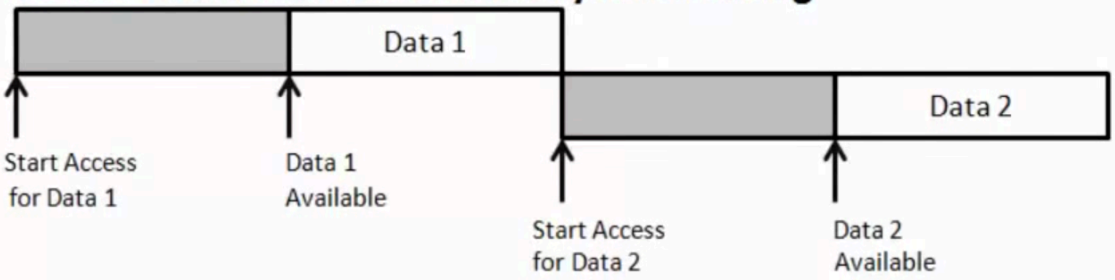
- *Goal: Maximize the effective memory bandwidth so that more words can be accessed per unit time; Matching the Memory bandwidth + bus bandwidth + processor bandwidth*

Memory Interleaving

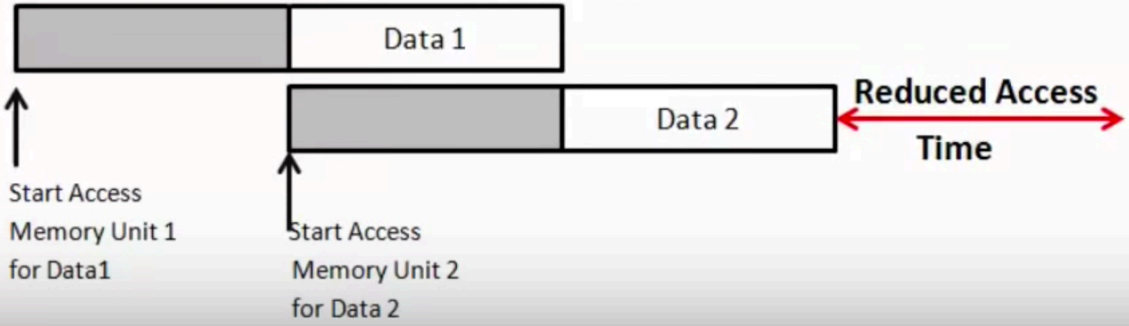
Technique to interleave successive memory addresses across multiple memory units



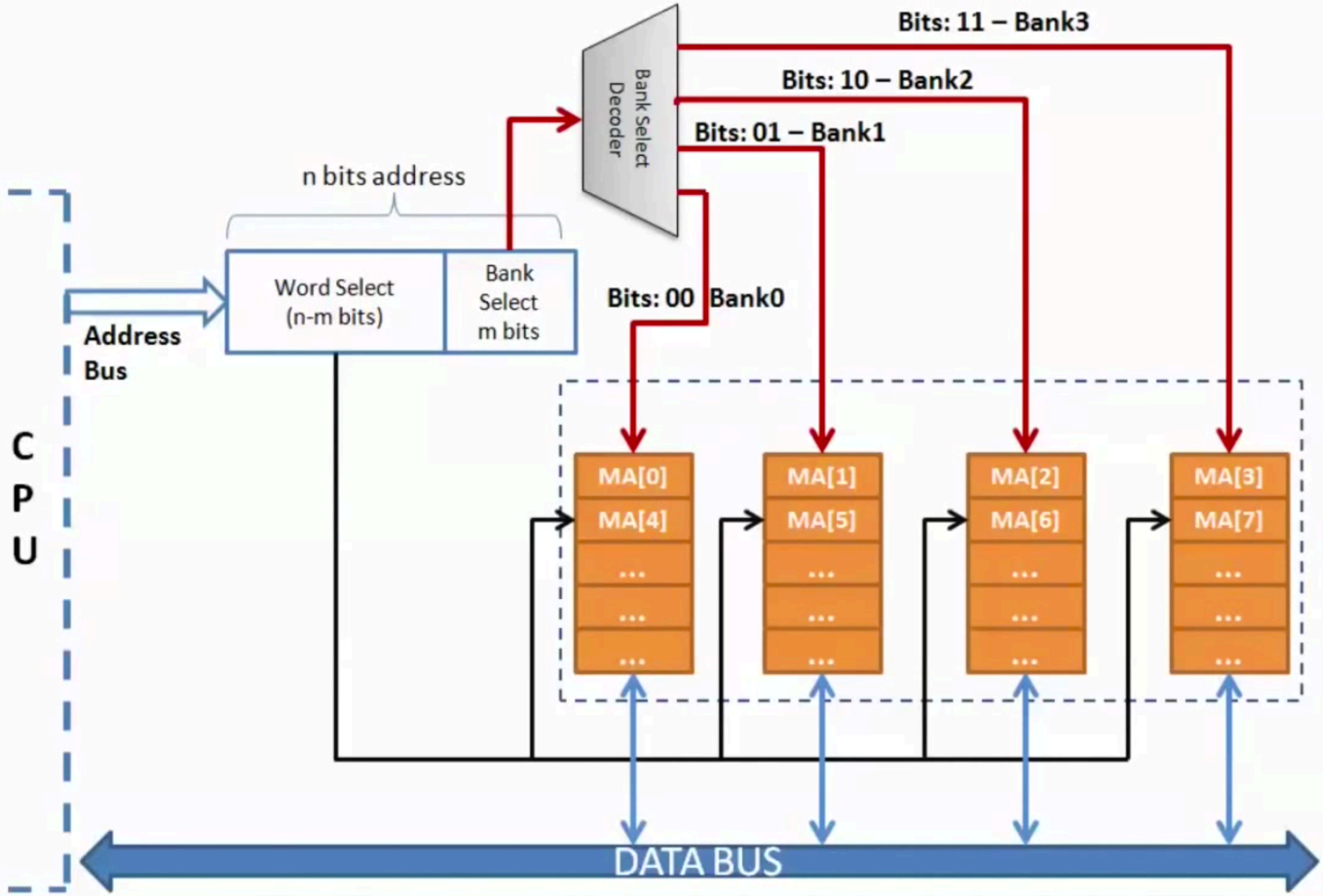
Access Pattern without Memory Interleaving



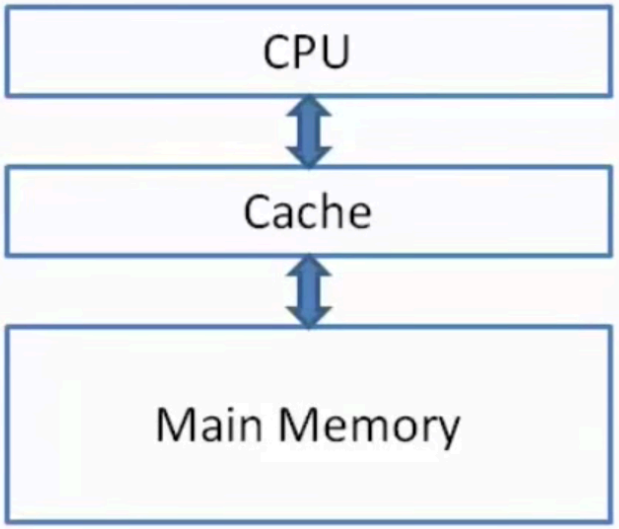
Access Pattern with Memory Interleaving



Abstract Architecture



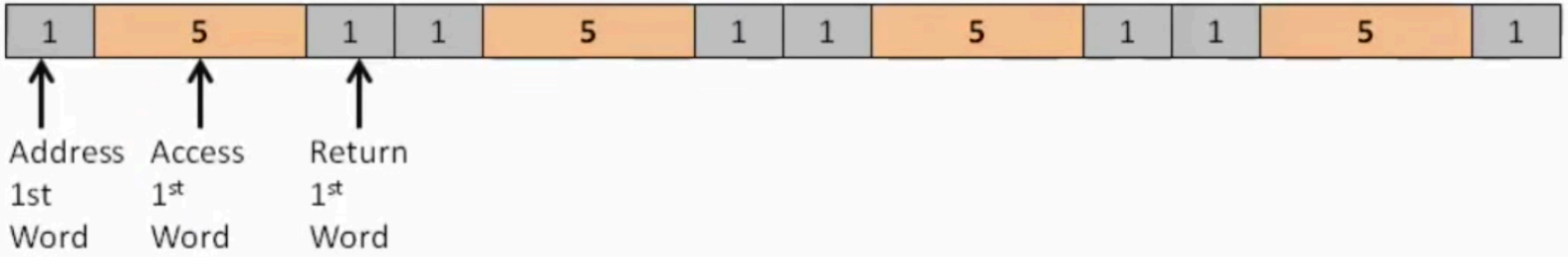
Non-Interleaved Memory Access with No Bank Division



4 word wide Cache

1 cycle to send address to DRAM
5 cycles for one word DRAM access
1 cycle to return data

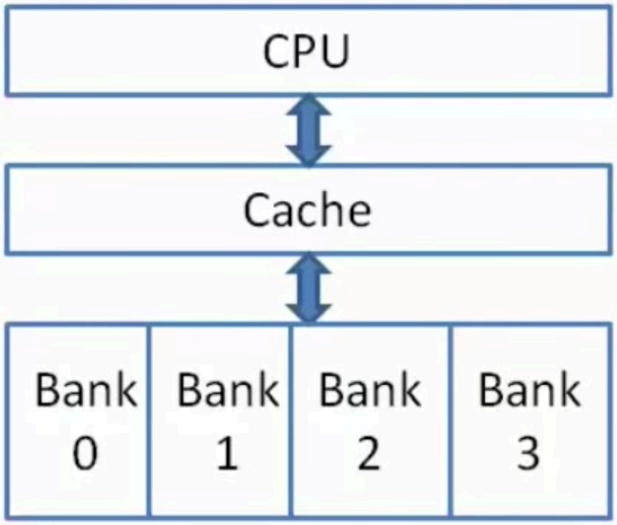
To Access 4 words of Data from Memory:



Time Taken to fetch 1 word from DRAM = 1 + 5 + 1 = 7 cycles

Time Taken to fetch 4 words from DRAM = 4 * (1 + 5 + 1) = **28 cycles**

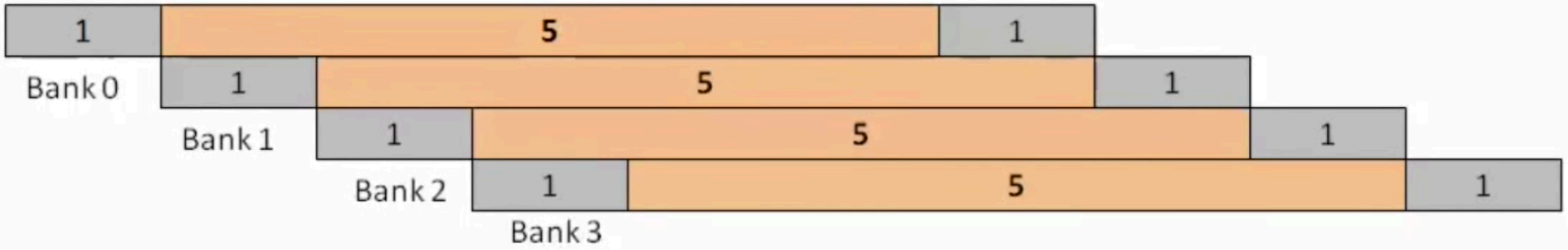
Interleaved Memory Access with 4 Banks



4 word wide Cache

1 cycle to send address to DRAM
5 cycles for one word DRAM access
1 cycle to return data

To Access 4 words of Data from Memory:



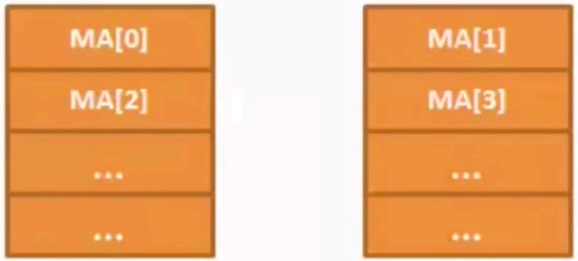
Time Taken to fetch 1 word from DRAM = 1 + 5 + 1 = 7 cycles

Time Taken to fetch 4 words from DRAM = (1 + 5 + 1) + (3*1) = **10 cycles**

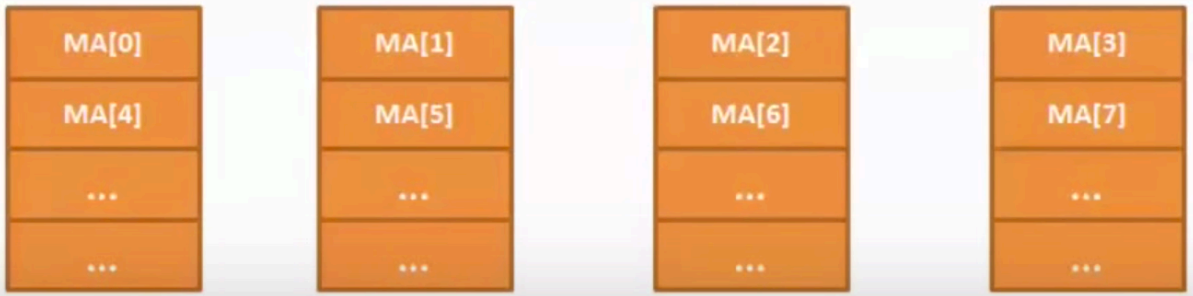
n – Way Memory Interleaving

The memory can be divided into multiple memory units. n – way interleaved memory means that the whole memory is divided into n number of memory units

2 – Way Memory Interleaving

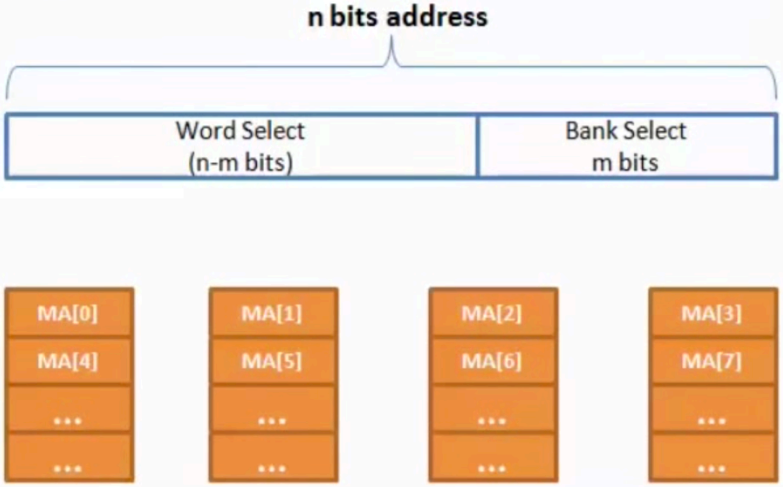


4 – Way Memory Interleaving

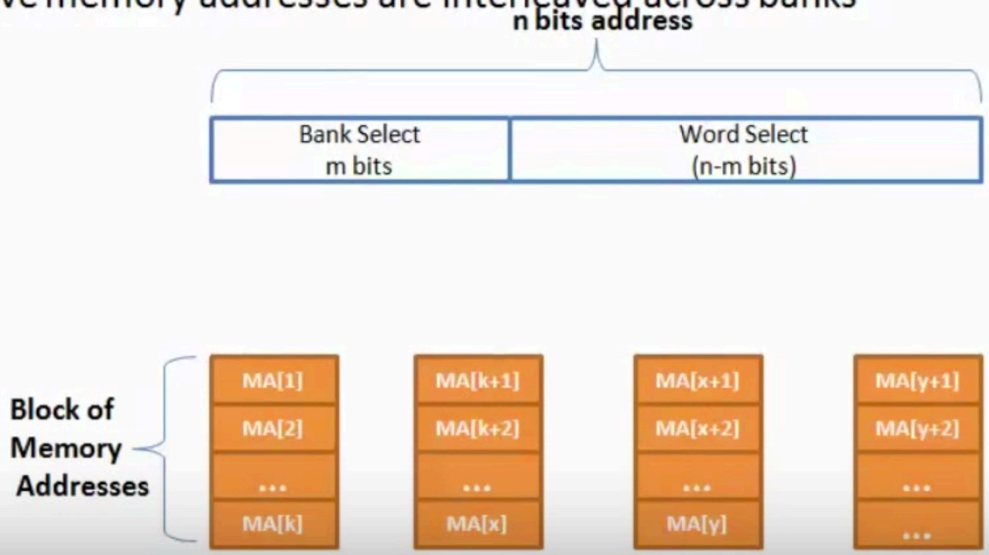


Types of Memory Interleaving

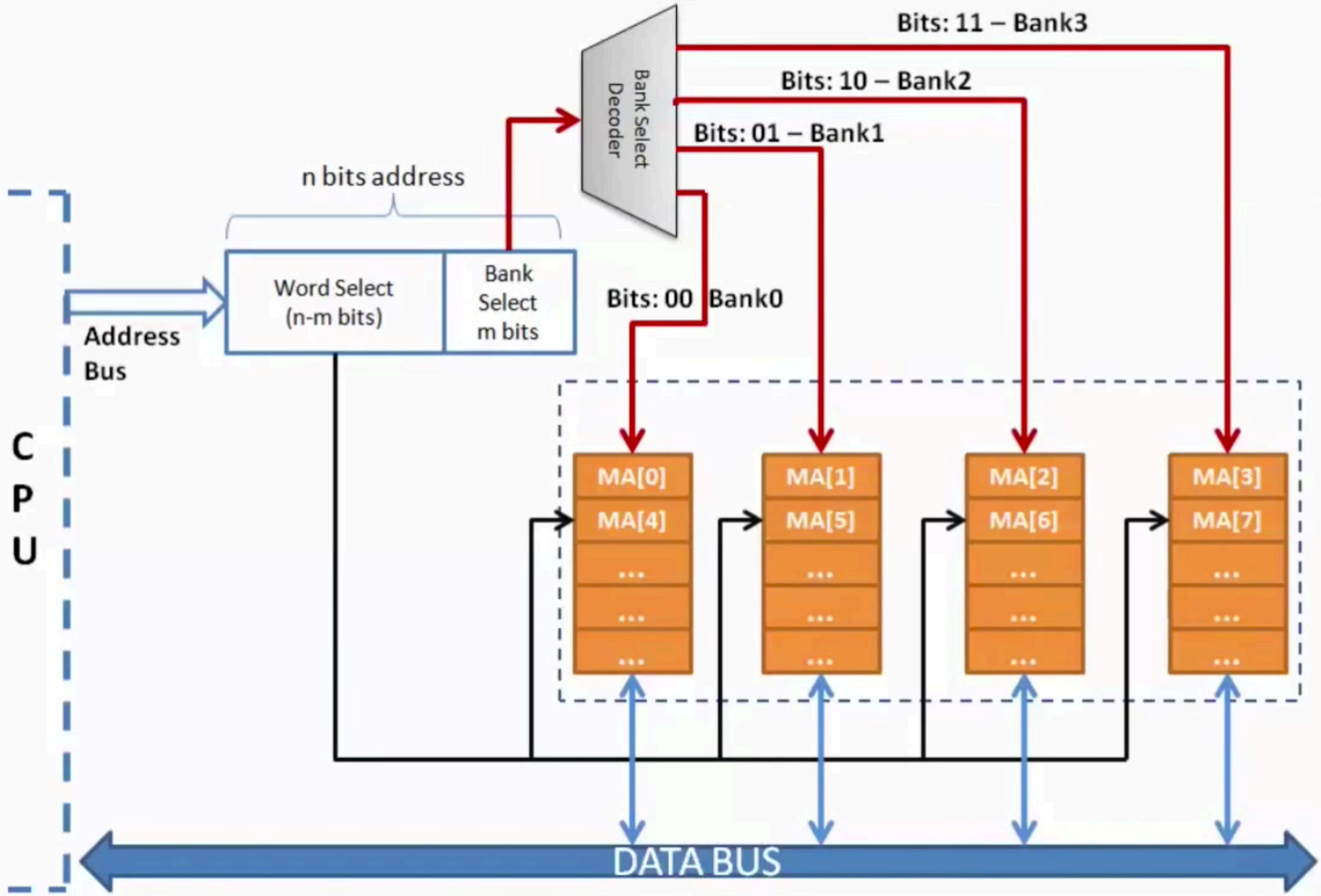
1. **Low Order Interleaving** : Least significant address bits are used for bank select. All successive memory addresses are interleaved across banks



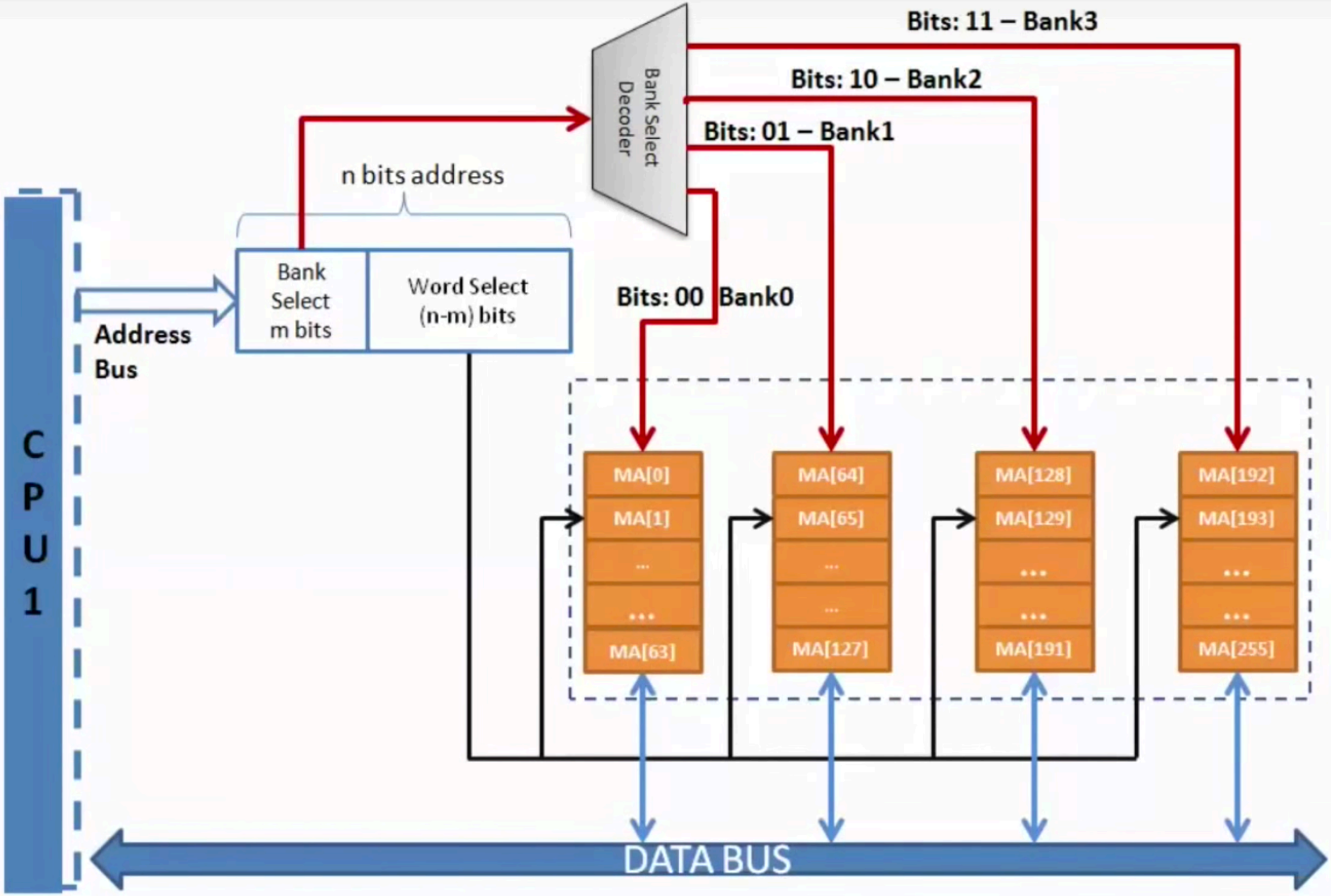
2. **High Order Interleaving** : Most significant address bits are used for bank select. Block of successive memory addresses are interleaved across banks



Low Order Interleaving



High Order Interleaving



- Low-Order interleaving supports pipelined block access of contiguous memory locations

- Under low-order interleaving technique graceful degradation of performance is achieved

Example of Low-order interleaving – Fault-tolerance:
8 memory modules viewed as banks

8-way implies 8 banks – 1 module per bank – failure of 1 module can provide 7 words;

4-way implies 4 banks – 2 modules per bank – failure of 1 module can provide 6 words;

2-way implies 2 banks – 4 modules per bank - failure of 1 module means 1 bank failure; but this can still provide 4 words;

1-way implies 8 modules per bank – failure of 1 module will be a total failure of the system (total of 1 bank)

- High-Order interleaving is preferred for shared memory systems. With supporting hardware circuitry, when one CPU is accessing a memory module other CPU can access other memory module, especially ndner read modes.

- High-order interleaving technique does not support pipelined block access of contiguous locations;
