APPENDIX E
INTERLEAVED MEMORY

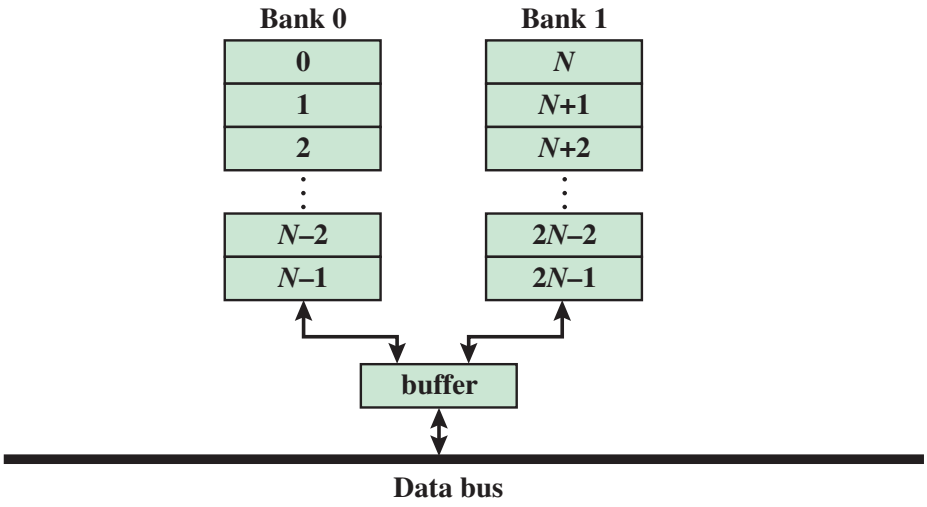
William Stallings

Supplement to
Computer Organization and Architecture, Ninth Edition
Prentice Hall 2012
ISBN: 013293633X
http://williamstallings.com/ComputerOrganization

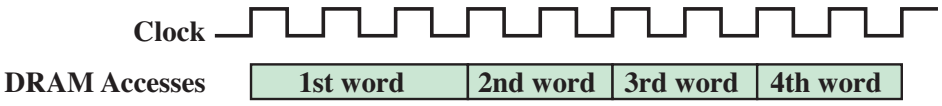
Main memory is typically of a series of DRAM chips. A number of these chips form a bank, with a port for transfer of data to and from the processor or an intermediate cache. Multiple memory banks can be connected together to form an interleaved memory system. Because each bank can service a request, an interleaved memory system with K banks can service K requests simultaneously, increasing the peak data transfer rate by a factor of K over the data transfer rate of a single bank. In most memory systems, the number of banks is a power of 2; that is, $K = 2^k$ for some integer k.

To get a feel for the use of interleaved memory, let us consider a simple system consisting of two DRAM memory banks. If the memory controller does not support interleaving, then the memory addresses are assigned sequentially in the first bank, followed by addresses in the second bank. Figure E.1a shows this organization for two banks of N words (assuming addressing is at the word level). Typically, the memory controller will perform a burst access (a single bus transaction that reads or writes multiple words) to move data between cache and memory. For example, the cache may have a line size of four 32-bit words and so data is transferred between memory and cache in blocks of four words. All the words in the block come from one bank of DRAM in a non-interleaved memory organization, so the time required to complete the transfer is a linear function of the number of words transferred. Figure E.1b shows the timing of the transfer. Note that the time needed to transfer each of the second, third, and fourth words is shorter than the time for the first word. This is because of a feature of contemporary DRAMs known as page-mode access. This is, in effect, a form of on-chip caching on the DRAM chip [JACO08].

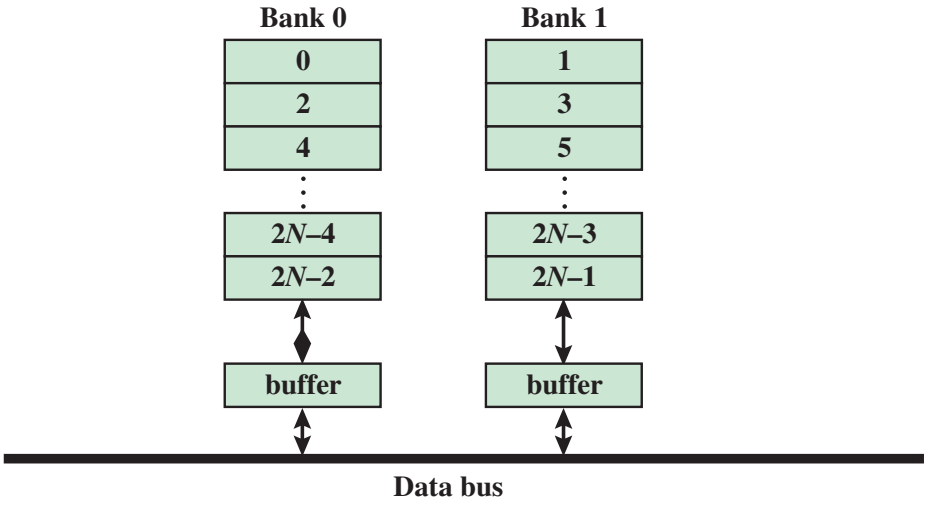
If the memory controller supports interleaved memory, then memory addresses are organized as shown in Figure E.1c. Memory location addresses alternate between the two banks. This configuration speeds up the burst transfer of four words, as shown in the timing diagram of Figure E.1d.



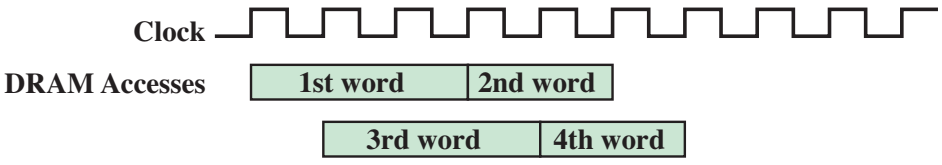
(a) Non-interleaved memory organization



(b) Non-interleaved memory timing



(c) Interleaved memory organization



(d) Interleaved memory timing

Figure E.1 Example of 2-Way Interleaved Memory

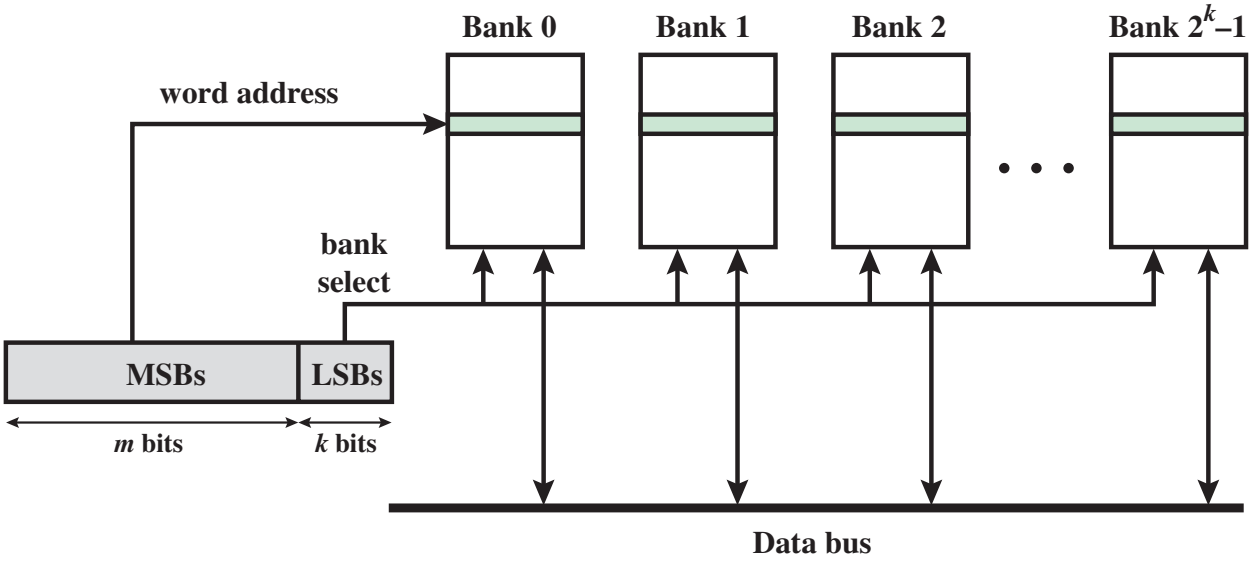


Figure E.2 Interleaved Memory

Because the four words of a burst access are spread across two physical banks of DRAM, the individual accesses can be overlapped to hide part, or all, of the DRAM access time delay.

Figure E.2 shows the organization of an interleaved memory with 2^k DRAM banks. Multiple memory banks are connected to a single bus (channel) and differentiated by the lower (least significant) k bits of the address bus. They share the bus by time division and overlapping the operations. If the address length is $m + k$, bits, then the upper (most significant) m bits of the address select a word within a memory bank, while the lower k bits select the given memory bank.

The interleaved memory system is most effective when the number of memory banks is equal to or an integer multiple of the number of words in a cache line.