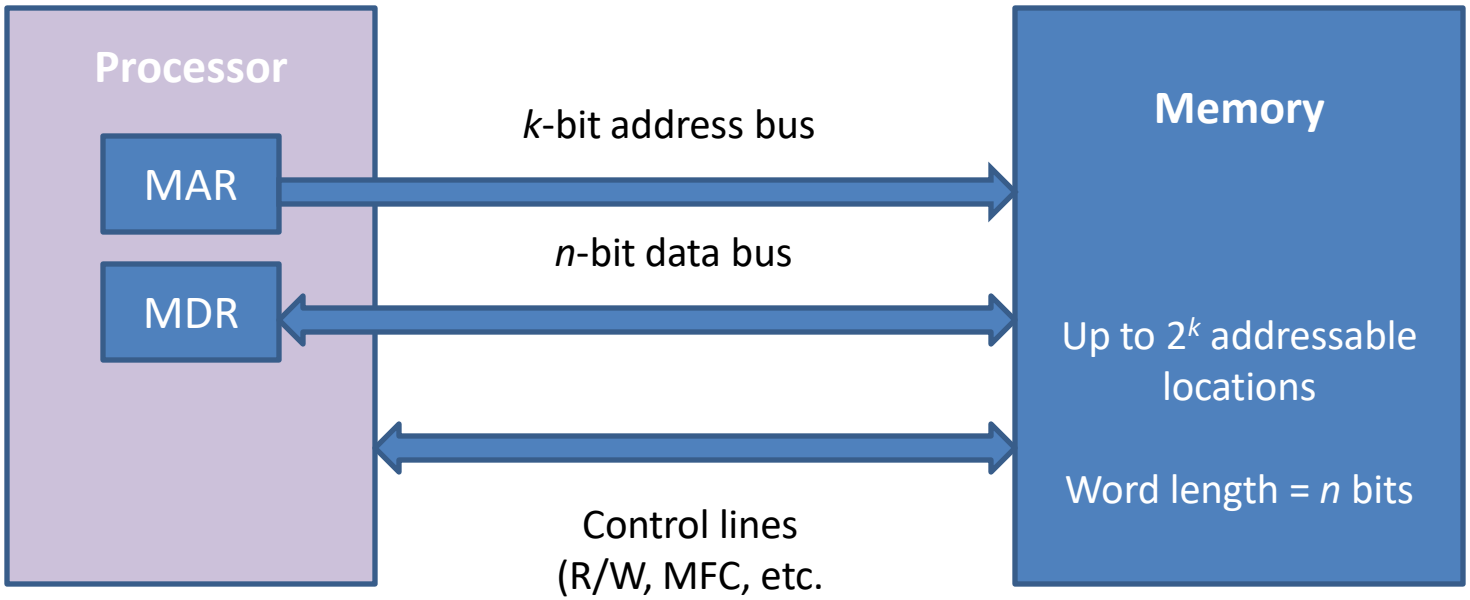
Computer Organization & Architecture

Introduction

- Maximum size of memory that can be used in any computer is determined by addressing scheme.
  - Ex., a computer that generates 16-bit addresses is capable of addressing upto 2^{16} =64 K memory locations.
  - Machines whose instructions generate 32-bit addresses can utilize a memory that contains upto 2^{32} = 4 G locations
  - Machines with 64-bit addresses can access up to 2^{64} =16 E ≈16×10^{18} locations.
- Number of locations represents size of address space of the computer

Introduction

- Memory is usually designed to store and retrieve data in word-length quantities
- Consider, a byte-addressable computer whose instructions generate 32-bit addresses.
  - When a 32-bit address is sent from the processor to the memory unit, high order 30 bits determine which word will be accessed
  - If a byte quantity is specified, low-order 2 bits of address specify which byte location is involved


Processor
MAR
MDR
k-bit address bus
n-bit data bus
Control lines
(R/W, MFC, etc.
Memory
Up to 2^k addressable locations
Word length = n bits

Ex. Byte addressable memory, 32-bit address, 32-bit word length

3	2	1	0	Word 0
7	6	5	4	Word 1
11	10	9	8	Word 2
15	14	13	12	Word 3
	…	…		
	…	…		
	…	…		
	…	…		
	…	…		
				Word $2^{30} - 1$

00000000 00000000 00000000 00001110

Measure of memory speed

- **Memory access time:** time that elapses between initiation of an operation to transfer a word of data and completion of operation
- **Memory cycle time:** minimum time delay required between the initiation of two successive memory operations
  - For example, time between two successive Read operations.
  - Cycle time is usually slightly longer than access time, depending on implementation details of memory unit.

Measure of memory speed

- A memory unit is called a random-access memory (RAM) if access time to any location is same, independent of location's address
  - Distinguishes such memory units from serial, or partly serial, access storage devices such as magnetic and optical disks.
  - Access time of latter devices depends on address or position of data
  - Semiconductor random-access memories (RAMs) are available in a wide range of speeds; cycle times range from 100 ns to less than 10 ns.

Cache memory

- Processor of a computer can usually process instructions and data faster than they can be fetched from the main memory
  - Memory access time is usually bottleneck in system
- One way to reduce memory access time is to use a cache memory
  - Small, fast memory inserted between larger, slower main memory and processor
  - Holds currently active portions of a program and their data

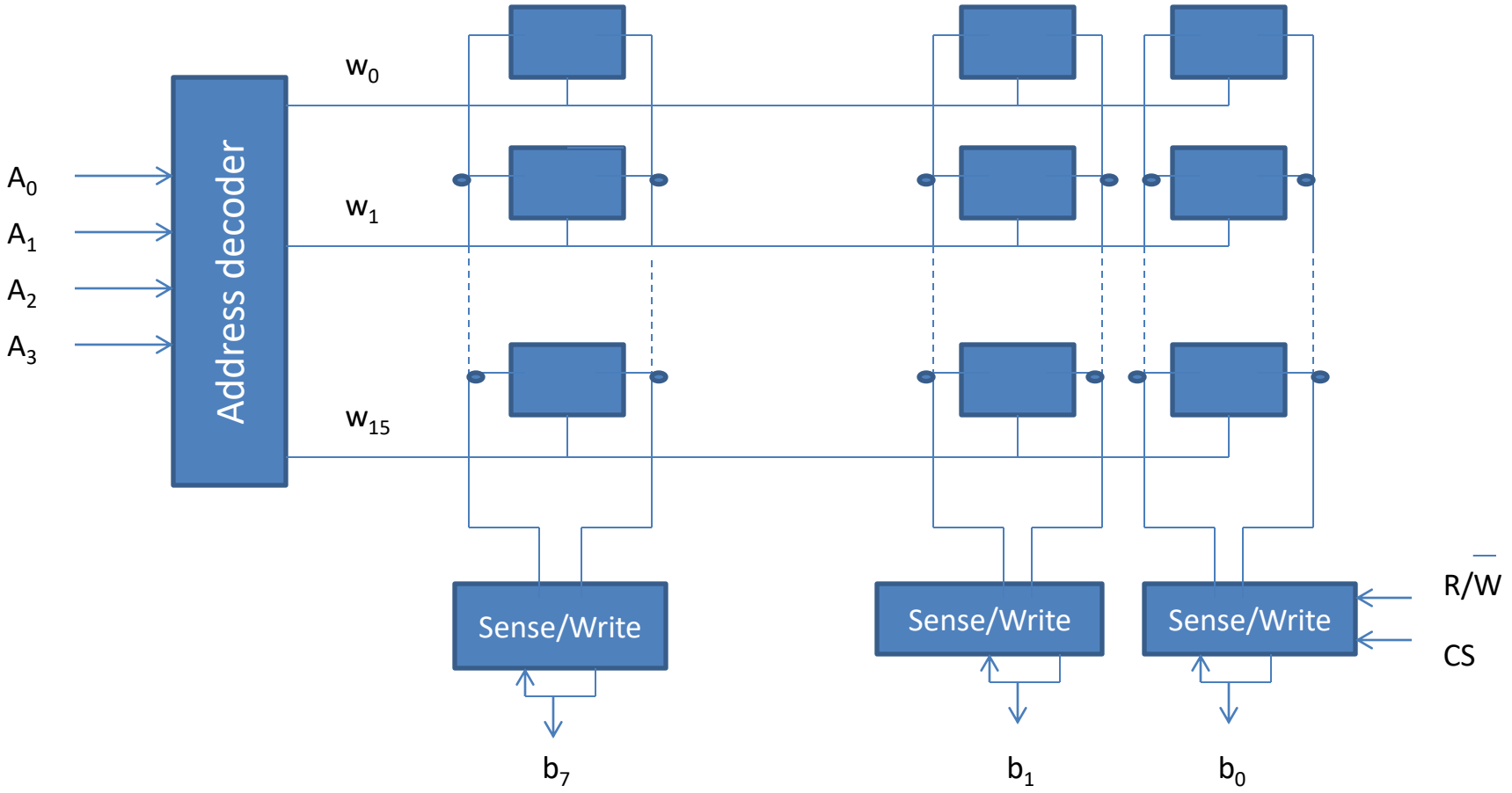
Virtual memory

- Virtual memory is another important concept related to memory organization
  - Only active portions of a program are stored in main memory, and remainder is stored on much larger secondary storage device.
  - Sections of program are transferred back and forth between main memory and secondary storage device in a manner that is transparent to application program
  - Application program sees a memory that is much larger than computer's physical main memory

Block transfer

- Data move frequently between main memory and cache and between main memory and disk
  - These transfers do not occur one word at a time
  - Always transferred in contiguous blocks involving tens, hundreds, or thousands of words
- Data transfers between main memory and high-speed devices such as a graphic display or an Ethernet interface also involve large blocks of data
  - A critical parameter for performance of main memory is its ability to read or write blocks of data at high speed.

Internal organization of memory chips



16*8 organization: Memory circuit stores 128 bits and requires 14 external connections for address, data, and control lines. Also needs two lines for power supply and ground connections

Internal organization of memory chips

- Ex., 1K (1024) memory cells
- Can be organized as a 128×8 memory
  - Requiring a total of 19 external connections
- Alternatively, can be organized into a 1K×1 format
  - 10-bit address is needed, but there is only one data line, resulting in 15 external connections
  - Required 10-bit address is divided into two groups of 5 bits each to form the row and column addresses for cell array
  - Row address selects a row of 32 cells, all of which are accessed in parallel
  - But, only one of these cells is connected to the external data line, based on column address.

Internal organization of memory chips