Memory and Storage

From the programmer's point of view, the main memory holds running programs as well as the data the programs use. Although computer memory has much broader meaning that includes a variety of special-purpose memory devices such as the memory in a cell phone that holds the address book or the memory that holds the program for an embedded processor.

An architect views a memory as a solid-state digital device that provides storage for data values. When an architect begins to design a memory, two key choices arise:

(a) Technology and (b) Organization

Technology refers to the properties of the underlying hardware mechanism is used to construct the memory system and **Organization** refers to the way the underlying technology is used to form a working system – the different choices about how to combine individual bits into larger units and how to address the units.

Characteristics of Memory Technology

A wide range of technologies have been invented with several purpose and characteristics such as:

- Volatile or non-volatile
- Random or Sequential access
- Read-write or Read-only
- Primary or Secondary

Memory Volatility

Memory is classified as volatile if the contents of the memory disappear when power is removed. The main memory used in most computers is volatile.

Memory is known as non-volatile, when the contents survive even after power is removed. The secondary storage media are all non-volatile in nature, such as hard-disk, Compact disk or the memory of the camera, mobile etc.

Memory Access Paradigm

The most common forms of memory are classified as random access, which means that any value in the memory can be accessed at any time. The alternative to random access is sequential access in which values must be read from memory in the same order they were inserted.

Permanence of Values

Memory is characterized by whether values can be extracted, updated or both. The primary form of memory in conventional computer system allows a value to be accessed (read) and updated (written) at any time.

However some memories are characterized as:

ROM (Read Only Memory) – The memory contains data values that can be accessed but cannot be changed.

PROM (Programmable Read-Only Memory) – This allows data values to be entered once, and then accessed many times. Typically values are initially placed in PROM by using high voltage to alter the physical circuits on the chip, e.g. to destroy the electrical path that corresponds to a zero bit. Informally we say that values are burned into the memory.

EEPROM (Electrically Erasable and Programmable Read Only Memory) – is a form of non-volatile memory that permits values to change. However, storing a value in EEPROM memory requires activation of special circuits and takes much longer than reading a value. EEPROMs are used in situations where non-volatility is desired, but values change infrequently.

Flash Memory or Flash ROM – This is a popular variant of EEPROM Technology, which is commonly used in digital cameras. Although it takes much longer time to store an image in Flash Memory, but it is not critical because it happens in less time than human to aim and focus the camera.

Primary and Secondary Memory

Primary Memory is the fast, volatile, internal main memory of a computer. Secondary memory is the slower, non-volatile storage provided by an external electro-mechanical device such as a hard disk.

Physical Memory and Physical Addressing

Engineers used the term Random Access Memory (RAM) to denote the type of memory used as the primary memory system in most computers. The name implies that RAM provides random access. In addition RAM offers read-write capability and is volatile in nature, i.e. values do not persist after the computer is powered down.

Static and Dynamic RAM Technologies

The technologies used to implement Random Access Memory can be divided into two broad categories.

(a) Static RAM or SRAM (b) Dynamic RAM or DRAM

Conceptually SRAM stores each data bit in a miniature digital circuit composed of multiple transistors similar to the flip-flop.

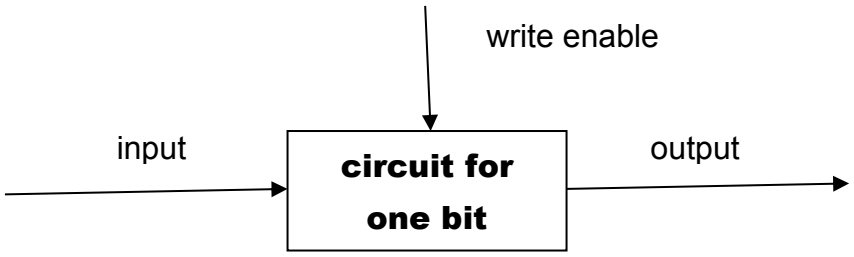


Illustration of a miniature Static RAM circuit that stores one data bit.

The circuit has two inputs and one output. Whenever the write enable input is on, the circuit sets the output voltage equals to the input (zero or one). Whenever the write enable input is off, the circuit ignores the input and keeps the output at the last setting. Thus to write a value, the hardware places the desired value in the input, turns on the write enable line and turns it off again.

Although it performs at high speed, SRAM has a significant disadvantage: power consumption and heat. The miniature circuit contains many transistors that operate continuously, where each transistor consumes a small amount of power and generates heat.

The alternative to Static Ram is known as Dynamic RAM or DRAM, which consumes less power. At the lowest level, to store information, DRAM uses a circuit that acts like a capacitor, a device that stores electrical charge. When a value is written to DRAM, the hardware charges or discharges the capacitor to reflect the digital value. Later, when the value is read from DRAM, the hardware examines the charge on the capacitor and generates the appropriate digital value.

But when the charge is kept for a long time, the capacitor gradually loses its charge. If a value is left long enough, the charge dissipates and the bit becomes zero. So before the entire charge dissipates, the bit needs to be read and written in it again, so that it causes the capacitor to start again with a full charge.

In practice, computers that use DRAM contains an extra hardware circuit, known as a refresh circuit that performs the task of reading and then writing a bit.

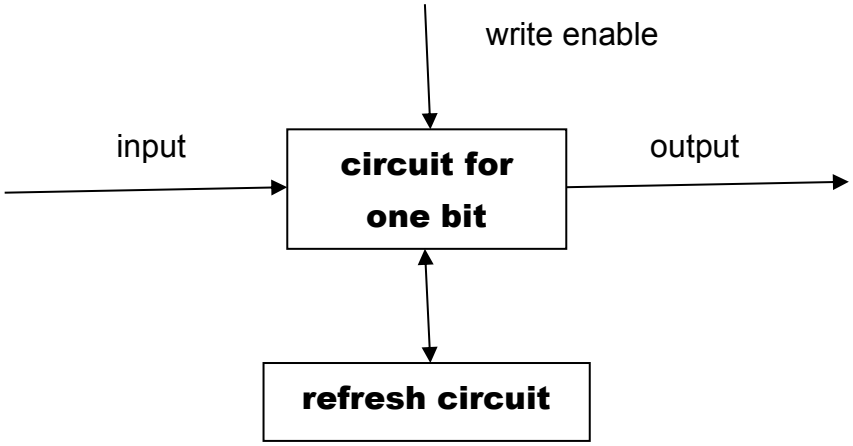


Illustration of a miniature DRAM circuit that stores one data bit with an external refresh circuit.

Complexity also arises because a refresh circuit must coordinate with normal memory operation. It should make sure that the value of ROM does not change during the refresh operation – i.e. read the value and write it back again. Despite the slowness, and the need for a refresh circuit, the cost and power consumption advantages of DRAM are so great that most computer memory is composed of DRAM rather than SRAM.

Latency and Memory Controllers

In addition to the memory chip themselves, additional hardware known as a **memory controller** provides an interface. To access memory, the processor presents a read or write request to the controller. the controller translates the memory address and request into signals appropriate for the underlying memory and passes the signal to the memory chip.

The controller returns an answer as quickly as possible, but after a response the controller needs additional clock cycles to reset hardware circuits and prepare for the next operation.

The read-cycle and write-cycle time are used as measures of memory system performance because they measure how quickly the memory system can handle successive requests.

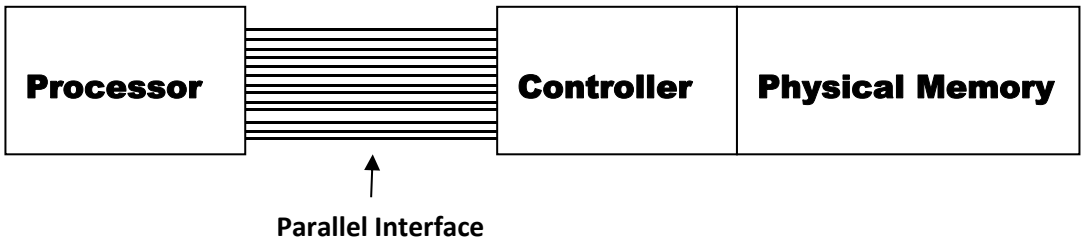
Like most other digital circuits in a computer, a memory system use a clock that controls exactly when a read or write operation begins. Now the processor clock differs from the clock used in the memory and this difference in clock rates affect the performance.

To eliminate this delay, some memory systems use a synchronized clock system. That is the clock pulses used with the memory system are aligned with the clock pulses used to run the processor. This synchronization can be used with DRAM or SRAM, which results in two technologies.

SDRAM – Synchronized Dynamic Random Access Memory

SSRAM - Synchronized Static Random Access Memory

Memory Address and Memory Bus



To achieve high performance, memory system use parallelism. If the parallel connection between the processor and memory contains N wires, it allows N bits to be transferred simultaneously. The technical name for this hardwire connection between a processor and memory is called memory bus.

The bits that comprise physical memory are divided into blocks of N bits per block, where N is the memory transfer size. A block of N bits is called a **word**, and the transfer size is called the **word size** or the **width of a word**.

Each word of the physical memory is assigned a unique number known as **physical memory address**.