1. RAM Defined

RAM (random access memory) is the place in a computer where the operating system, application programs, and data in current use are kept so that they can be quickly reached by the computer's processor. RAM is much faster to read from and write to than the other kinds of storage in a computer. However, the data in RAM stays there only as long as your computer is running. When you turn the computer off, RAM loses its data. When you turn your computer on again, your operating system and other files are once again loaded into RAM, usually from your hard disk.

RAM can be compared to a person's short-term memory and the hard disk to the long-term memory. The short-term memory focuses on work at hand, but can only keep so many facts in view at one time. If short-term memory fills up, your brain sometimes is able to refresh it from facts stored in long-term memory. A computer also works this way. If RAM fills up, the processor needs to continually go to the hard disk to overlay old data in RAM with new, slowing down the computer's operation. Unlike the hard disk which can become completely full of data so that it won't accept any more, RAM never runs out of memory. It keeps operating, but much more slowly than you may want it to.

2. How Big is RAM?

RAM is small, both in physical size (it's stored in microchip modules) and in the amount of data it can hold. It's much smaller than your hard disk. RAM comes in the form of "discrete" (meaning separate) microchip modules that plug into holes in the computer's motherboard. These holes connect through a bus or set of electrical paths to the processor. The hard drive, on the other hand, stores data on a magnetized surface that looks like a phonograph record.

Today's personal computers come with 8 or more gigabytes of RAM. Users of graphic applications usually need more gigabytes of memory. Most personal computers are designed to allow you to add additional RAM modules up to a certain limit. Having more RAM in your computer reduces the number of times that the computer processor has to read data in from your hard disk, an operation that takes much longer than reading data from RAM. (RAM access time is expressed in nanoseconds; hard disk access time is expressed in milliseconds.)

3. Why Random Access

RAM is called "random access" because any storage location can be accessed directly. Originally, the term distinguished regular core memory from offline memory, usually on magnetic tape in which an item of data could only be accessed by starting from the beginning of the tape and finding an address sequentially. Perhaps it should have been called "nonsequential memory" because RAM access is hardly random. RAM is organized and controlled in a way that enables data to be stored and retrieved directly to specific locations. A term IBM has preferred is *direct access* storage or memory. Note that other forms of storage such as the hard disk and DVD are also accessed directly (or "randomly") but the term *random access* is not applied to these forms of storage.

In addition to hard disk and DVD storage, another important form of storage is read-only memory (ROM), a more expensive kind of memory that retains data even when the computer is turned off. Every computer comes with a small amount of ROM that holds just enough programming so that the operating system can be loaded into RAM each time the computer is turned on.

4. What RAM Looks Like

In general, RAM is much like an arrangement of post-office boxes in which each box can hold a 0 or a 1. Each box has a unique address that can be found by counting across columns and then counting down by row. In RAM, this set of post-office boxes is known as an array and each box is a cell. To find the contents of a box (cell), the RAM controller sends the column/row address down a very thin electrical line etched into the chip. There is an *address line* for each row and each column in the set of boxes. If data is being read, the bits that are read flow back on a separate *data line*. In describing a RAM chip or module, a notation such as 256Kx16 means 256 thousand columns of cells standing 16 rows deep.

In the most common form of RAM, dynamic RAM, each cell has a charge or lack of charge held in something similar to an electrical capacitor. A transistor acts as a gate in determining whether the value in the capacitor can be read or written. In static RAM, instead of a capacitor-held charge, the transistor itself is a positional *flip/flop* switch, with one position meaning 1 and the other position meaning 0.

Externally, RAM is a chip that comes embedded in a personal computer motherboard with a variable amount of additional modules plugged into motherboard sockets. To add memory to your computer, you simply add more RAM modules in a prescribed configuration.

5. How Data Is Accessed

When the processor or CPU gets the next instruction it is to perform, the instruction may contain the address of some memory or RAM location from which data is to be read (brought to the processor for further processing). This address is sent to the RAM controller. The RAM controller organizes the request and sends it down the appropriate address lines so that transistors along the lines open up the cells so that each capacitor value can be read. A capacitor with a charge over a certain voltage level represents the binary value of 1 and a capacitor with less than that charge represents a 0. For dynamic RAM, before a capacitor is read, it must be power-refreshed to ensure that the value read is valid. Depending on the type of RAM, the entire line of data may be read that the specific address happens to be located at or, in some RAM types, a unit of data called a page is read. The data that is read is transmitted along the data lines to the processor's nearby data buffer known as level-1 cache and another copy may be held in level-2 cache or level-3cache.

6. How RAM Effectiveness is Measured

The amount of time that RAM takes to write data or to read it once the request has been received from the processor is called the *access time*. Typical access times vary from 9 nanoseconds to 70 nanoseconds, depending on the kind of RAM. Although fewer nanoseconds is better, user-perceived performance is based on coordinating access times with the computer's clock cycles. Access time consists of latency and *transfer time*. Latency is the time to coordinate signal timing and refresh data after reading it.