

11 December 2006

By: Bogdan Solca, Hardware Editor



"Memories" of a PC

How PC RAM works and the RAM legacy

This week it's time to analyze some more crucial PC components. First, we are going to see why RAM is so important for the overall PC speed. You may remember that I mentioned some things about RAM when I presented the CPU and the motherboard. However, this article is going to supply the omitted details. So, here we go.

Ramming the definition RAM (random access memory) is the computer component where the operating system (OS), software programs, and data in current use are stored so that they can be quickly accessed by the computer's CPU. RAM is much faster to read from and write to than the other kinds of storage in a computer, the hard disk, floppy disk, and CD-ROM/DVD-ROM. But there is one important difference involved here: the data in RAM stays there only as long as your computer is running. When you turn the computer off, RAM is emptied and all that data is lost. When you turn your computer on again, your OS and other startup-related files are once again loaded from the HDD back into the system RAM. To better understand the concepts behind RAM, an analogy with the human brain will come in handy. RAM can be compared to a person's short-term memory and the hard disk to the long-term memory. The human short-term memory focuses on work at hand, but can only keep so many facts in view at one time. If the short-term memory fills up, your brain is sometimes able to refresh it from facts stored in long-term memory. Computer RAM works in a similar way. Depending on the amount of storage space, RAM may eventually fill up and in this case, the processor needs to continually go to the hard disk to overlay old data in RAM with new, slowing down the computer's speed. Unlike the hard disk which can become completely full of data so that it won't accept any more unless you delete something, RAM never runs out of storage space as it works hand in hand with the hard disk. However, when helped by the hard disk, RAM operates at slower speeds than the standard ones. **A lightning-fast dwarf** RAM is usually smaller than the hard disk, both in physical size (it's stored in microchips) and in the amount of data it can store. Nowadays, typical computers may come with around 1 billion bytes (1 GB) of RAM, while the hard disk may store more than 250 (250 GB) or even 1 trillion bytes (1 TB) and beyond. As I mentioned in the previous articles, RAM comes in the form of modules or sticks that house a series of memory chips. The stick plugs into specially designated connectors on the computer's motherboard. These connectors communicate through a bus or set of electronic paths to the CPU. Unlike RAM sticks, the hard disk stores data on a magnetized surface that looks like a phonograph record (platters). When buying a new motherboard, you should take into consideration the possibility of expanding the RAM capacity. Nowadays, PC motherboards usually feature 4 memory slots that allow you to add additional RAM modules. Having more RAM in your computer reduces the number of times the CPU has to read data in from your hard disk, an operation that takes much longer than reading data from RAM (RAM access time is in nanoseconds; hard disk access time is in milliseconds). The "random access" part usually means that any storage location in the memory can be accessed directly. A more appropriate term would have been "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. IBM always preferred to call it direct access storage or memory. Note that other forms of storage such as the hard disk and CD-ROM are also accessed directly, but the term random access is not mentioned with these forms of storage. In addition to hard disk, floppy disk, and CD-ROM 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 memory stick 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 (apart from the BIOS functions). **Core structure** Similar to a microprocessor, a memory chip is an integrated circuit (IC), comprising millions of transistors and capacitors. In the most common form of computer memory, dynamic random access memory (DRAM), a transistor and a capacitor are paired to create a memory cell, which represents a single bit of data. The capacitor holds the bit of information - a 0 or a 1. The transistor acts as a switch that lets the control circuitry on the memory chip read the capacitor or change its state. To better understand how these micro-structures work, a comparison between electronics and real-world common examples is always welcome. In this sense, a capacitor is like a small bucket that is able to store electrons. To store a 1 in the memory cell, the bucket is filled with electrons. To store a 0, the bucket is emptied. The problem with the capacitor's "bucket" is that it has a leak. In a matter of a few milliseconds, a full bucket becomes empty. Therefore, for dynamic memory to work, either the CPU or the memory controller has to come along and recharge all of the capacitors holding a 1 before they discharge. To do this, the memory controller reads the memory and then writes it right back. This refresh operation happens automatically thousands of times per second (remember that the access time is measured in nanoseconds for memory components). This refresh operation is where dynamic RAM gets its name from. Dynamic RAM has to be dynamically refreshed all the while the computer operates or it forgets what it is holding. Unfortunately, all the refreshing procedures take time and slow down the memory, introducing relative latencies. Memory cells are etched onto a silicon wafer in an array of columns (bitlines) and rows (wordlines). The intersection of a bitline and wordline constitutes the address of the memory cell. Thus, there appears to be an intricate structure of bits arranged in a two-dimensional grid. DRAM works by sending a charge through the appropriate column to activate the transistor at each bit in the column. We know that the basic actions that happen inside memory chips are reading and writing. When writing, the row lines contain the state the capacitor should take on. When reading, the sense-amplifier determines the level of charge in the capacitor. If it is more than 50 percent, it reads it as a 1; otherwise it reads it as a 0. There is a counter that tracks the refresh sequence based on which rows have been accessed and in what order. The length of time necessary to do all this is so short that it is expressed in nanoseconds (billionths of a second). A memory chip's latency rating of 70ns means that it takes 70 nanoseconds to completely read and recharge each cell. Memory cells alone would be worthless without some way to get information in and out of them. So the memory cells have a whole support infrastructure of other specialized circuits. These circuits perform functions such as:

- * Identifying each row and column: row address strobe (RAS) and column address strobe (CAS); these are two important functions that are measured in clock cycles.
- * Read/write activator: Chips Select (CS).
- * Keeping track of the refresh sequence (counter).
- * Reading and restoring the signal from a cell (sense amplifier).
- * Telling a cell whether it should take a charge or not (write enable).

Many memory modules have special specifications that determine the overall speed, so keep an eye on them when buying for upgrades. These specifications are measured in clock cycles. For example, specifications like 2-3-3-7-1T present the clock cycles attributed as following: 2 - the CAS latency 3 - the RAS to CAS delay (tRCD) 3 - the RAS Precharge (tRP) 7 - the Active to Precharge delay (tRAS) 1T - Command Rate (can be 1T or 2T for today's memories). The memory controller is specialized in identifying the type, speed and amount of memory and error checking. **Static RAM is for the privileged** There is another form of RAM - static RAM - which uses a completely different technology. In static RAM, a form of flip-flop holds each bit of memory. A flip-flop of a memory cell takes four or six transistors along with some wiring, but never has to be refreshed. This makes static RAM significantly faster than dynamic RAM. However, because it has more parts, a static memory cell takes up a lot more space on a chip than a dynamic memory cell. Therefore, you get less memory per chip, and that makes static RAM a lot more expensive. However, static RAM is to be found inside the CPU. You certainly remember the Level 1 and Level 2 cache memory

inside a CPU. Because static RAM is quite expensive and takes too much space, the L1 and L2 caches feature little storage space (nowadays, values between 64 KB and 2 MB are common). **All your "memories" are belong to us!** I don't want to bore you with present-day memory bank configurations, numbers of pins, error-checking bits. So we'll go directly to a list of the most common types of memory that exist nowadays. As we have already discussed about DRAM and SRAM, I'm going to skip these two. Keep in mind that all memories have to communicate with the L2 cache, before reaching the CPU core.

- * FPM DRAM: Fast page mode dynamic random access memory was the original form of DRAM. It waits through the entire process of locating a bit of data by column and row and then reading the bit before it starts on the next bit. Keep in mind that all memories have to make contact with the Maximum transfer rate to L2 cache which is approximately 176 MBps.
- * EDO DRAM: Extended data-out dynamic random access memory does not wait for all of the processing of the first bit before continuing to the next one. As soon as the address of the first bit is located, EDO DRAM begins looking for the next bit. It is about five percent faster than FPM. Maximum transfer rate to L2 cache is approximately 264 MBps.
- * SDRAM: Synchronous dynamic random access memory takes advantage of the burst mode concept to greatly improve performance. It does this by staying on the row containing the requested bit and moving rapidly through the columns, reading each bit as it goes. The idea is that most of the time, the data needed by the CPU will be in sequence. SDRAM is about five percent faster than EDO RAM and is the most common form in desktops today. Maximum transfer rate to L2 cache is approximately 528 MBps.
- * DDR SDRAM: Double data rate synchronous dynamic RAM is just like SDRAM, but supports a higher bandwidth (double the SDRAM one) which translates into greater speeds. Maximum transfer rate to L2 cache is approximately 1,064 MBps (this is the nominal speed for DDR SDRAM clocked at 133 MHz).
- * RDRAM: Rambus dynamic random access memory is a radical departure from the previous DRAM architecture. Designed by Rambus, RDRAM uses a Rambus in-line memory module (RIMM), which is similar in size and pin configuration to a standard DIMM. What makes RDRAM so different is its use of a special high-speed data bus called the Rambus channel. RDRAM memory chips work in parallel to achieve a data rate of 800 MHz, or 1,600 MBps. Since they operate at such high speeds, they generate much more heat than other types of chips. To help dissipate the excess heat, Rambus chips are fitted with a heat spreader, which looks like a long thin wafer. These types of memory have been presented in chronological order of their appearance as standards and each of them has specially designed counterparts for laptop and notebook PCs. But there also are other types that fit a series of peripherals inside a PC or a notebook/laptop.
- * Credit Card Memory: Credit card memory is a proprietary self-contained DRAM memory module that plugs into a special slot for use in notebook computers.
- * PCMCIA Memory Card: Another self-contained DRAM module for notebooks, cards of this type are not proprietary and should work with any notebook computer whose system bus matches the memory card's configuration.
- * CMOS RAM: CMOS RAM is a term for the small amount of memory used by your computer and some other devices to remember things like each and every PC component settings. This memory uses a small battery to provide it with the power it needs to maintain the memory contents. It is more of a ROM type from this point of view.
- * VRAM: VideoRAM, also known as multiport dynamic random access memory (MPDRAM), is a type of RAM used specifically for video adapters or 3-D accelerators. The "multiport" part comes from the fact that VRAM normally has two independent access ports instead of one, allowing the CPU and graphics processor to access the RAM simultaneously. VRAM is located on the graphics card and comes in a variety of formats, many of which are proprietary. The amount of VRAM is a determining factor in the resolution and color depth of the display. VRAM is also used to hold graphics-specific information such as 3-D geometry data and texture maps. True multiport VRAM tends to be expensive, so today, Nvidia and ATI, the two most important graphic cards makers, embraced the GDDR SDRAM (Graphics DDR SDRAM) standard. Current Nvidia GeForce 8800 series and ATI's

upcoming Radeon R600 Graphics Processing Unit feature GDDR 4 support. This list would conclude our short analysis of today's RAM. Remember that the easiest way of speeding things up in your PC, without changing the CPU and the motherboard, would be to add more reduced-latency RAM. Tomorrow, we will take a look at the evolution of graphics cards and see how they work.