

THE MEMORY GUIDE

Mark E. Donaldson

INTRODUCTION AND OVERVIEW

Once upon a time, magazine ads for computers were actually readable. Not only were there fewer processors to choose from, there were also fewer choices in everything else, from hard disks to sound cards. Today, the choices boggle the mind. Acronyms run rampant. And as if it weren't bad enough having to deal with all the other jargon, we now have to confront a seemingly limitless variety of memory types. It used to be enough to have RAM, but now RAM always seems to have one or more extra letters in front of it. DRAM, VRAM, SRAM, SDRAM, WRAM--it's enough to confuse even the staunchest technology watcher.

The CPU, yours is probably an Intel 486 or Pentium type, is the heart of the computer, where data is processed and program instructions are interpreted. Integrated with the CPU is the system's main memory, called random-access memory or RAM. Together, these two components make up the core of your machine; components such as hard disks, controllers, and video cards are peripheral to this central activity, and are therefore known as peripherals.

The CPU uses its RAM as a storage area for data, calculation results, and program instructions, drawing on this storage as necessary to perform the tasks required by programs. In order to store data and draw from the data store, the CPU specifies the memory address of the required information. The address bus allows the CPU to send the address to RAM, and the data bus allows the actual data transfer to the CPU. The term bus itself refers to the connection between the two devices that allows them to communicate. An important measurement of RAM performance is access time, the amount of time that passes between the instant the CPU issues an instruction to RAM to read a particular piece of data from a particular address and the moment the CPU actually receives the data. Today's RAM chips typically have a 60-ns access time, which means it takes 60 nanoseconds (a nanosecond is a billionth of a second) to perform this round-trip function. This access time is much faster than that of the 100- to 120-ns chips of a few years ago, but it's still much slower than the ideal access time of zero, which would be realizable if the CPU itself stored all the data. To speed things further, the CPU has access to cache memory (usually referred to as "the cache"). At 20 ns or better, cache memory is faster than main memory, but systems contain less of it than main memory (cache memory is expensive), and therefore only select data, the data the CPU will probably need next, is placed inside it. The selection is handled by the cache controller.

Memory chips function by storing electronic charges. The chips are made up of a capacitor and a transistor, with the capacitor storing the charge and the transistor turning the charge on or off. With RAM chips, the system can alter the on or off state of the charges, but with ROM (read-only memory) chips the charges are either permanently on or permanently off. This article deals only with RAM.

All RAM technologies emphasize speed and attempt to offer more of it without an increase in cost. But CPU technology keeps getting faster, and memory technology must keep pace, hence the need for different types of RAM. It may be confusing, but rest assured, it's in our best interests.

THE MEMORY GUIDE

Mark E. Donaldson

RAM

This is the umbrella term for all memory that can be read from or written to in a nonlinear fashion. However, it has come to refer specifically to chip-based memory, since all chip-based memory is random-access. It is not the opposite of ROM. The computer can read ROM; it can read and write to RAM.

SIMM (Single in-line Memory Module) DIMM (Dual in-line Memory Module)

SIMM and DIMM refer not to memory types, but to modules (circuit boards plus chips) in which RAM is packaged. SIMMs, the older of the two, offer a data path of 32 bits. Because Pentiums are designed to handle a much wider data path than that, SIMMs must be used in pairs on Pentium motherboards (they can be used singly on boards based on 486 or slower processors). DIMMs, which are of more recent origin, offer a 64-bit path, which makes them more suitable for use with the Pentium and other more recent processors. From a buyer's standpoint, the good news is that one DIMM will handle the work of two SIMMs and thus can be used singly on a Pentium motherboard. DIMMs are more economical in the long run, because you can add one at a time to your system.

DRAM (Dynamic RAM)

Dynamic RAM is the standard main memory type in computers today and is what you're referring to when you tell someone your PC has 32MB of RAM. In DRAM, information is stored as a series of charges in a capacitor. Within a millisecond of being electronically charged, the capacitor discharges and needs to be refreshed to retain its values. This constant refreshing is the reason for the use of the term dynamic.

<u>Clock</u>	<u>Cycle</u>
13Mh	80 nsec
19Mh	60 nsec
25Mh	40 nsec
33Mh	30 nsec
50Mh	20 nsec
66Mh	15 nsec
100Mh	10 nsec
200Mh	5 nsec

FPM RAM (Fast Page-Mode RAM)

Until the advent of EDO RAM (see below), all main memory found in PCs was of the fast page-mode variety. That's why the name wasn't well known: There was no need to state the type, since there was only one. The access times of FPM RAM dropped as the technology matured, from 120 ns (nanoseconds) down to the now-common access time of 60 ns. The Pentium processor, however, allows for a bus speed of 66 MHz, which is faster than FPM RAM can keep up with. The speed of a 60-ns RAM module performing random page access (where page refers to a region of address space) is below 30 MHz, far slower than the bus speed. So DRAM makers came up with the concept of the RAM cache.

THE MEMORY GUIDE

Mark E. Donaldson

EDO RAM (Extended-Data-Out RAM)

Despite the hype surrounding it, EDO RAM is no more than another type of FPM RAM. Essentially, it recognizes that most of the time when the CPU requests memory for a particular address, it's going to want some more addresses nearby. Instead of forcing each memory access to start afresh, EDO RAM hangs onto the location of the previous access, thereby speeding access to nearby addresses. EDO RAM speeds up the memory cycle, with improvements in memory performance of as much as 40 percent. But EDO RAM is effective only up to a bus speed of 66 MHz, and that's quickly being bypassed by the most recent crop of AMD, Cyrix, and Intel processors.

BEDO RAM (Burst Extended-Data-Out RAM)

As the need for faster access to DRAM has increased, technologies have been developed to provide it. One such technology is known as bursting, in which large blocks of data are sent and processed in the form of an uninterrupted "burst" of smaller units. What this means to DRAM is that the burst carries details not only about the address of the first page, but also of the next few. BEDO RAM can handle four data elements in one burst, and this allows the final three elements to avoid experiencing the delays of the first, all the addresses are ready to be processed. The DRAM is given the first address, and then can process the rest at a rate of 10 ns each. BEDO RAM, however, despite its substantial speed increase, still has difficulty moving past the 66-MHz bus barrier. BEDO RAM exists because SDRAM manufacturers were uninterested in pricing SDRAM to be competitive with EDO RAM; as a result, more work was done with EDO to add bursting technologies for speed rivaling that of SDRAM. Hence BEDO RAM.

SDRAM (Synchronous Dynamic RAM)

Resources galore are being poured into SDRAM development, and it has begun making its appearance in the PC ads. The reason for its increasing popularity is twofold. First, SDRAM can handle bus speeds of up to 100 MHz, and these are fast approaching. Second, SDRAM is synchronized with the system clock itself, a technical feat that has eluded PC engineers until now. SDRAM technology allows two pages of memory to be opened simultaneously. A new standard for SDRAM is being developed by the SCiZZL Association at Santa Clara University (California) along with many industry leaders. Called SLDRAM, this technology improves on SDRAM by offering a higher bus speed and by using packets (small packs of data) to take care of address requests, timing, and commands to the DRAM. The result is less reliance on improvements in DRAM chip design, and ideally a lower-cost solution for high-performance memory.

SRAM (Static Random-Access Memory)

The difference between SRAM and DRAM is that where DRAM must be refreshed constantly, SRAM stores data without an automatic refresh. The only time a refresh occurs, in fact, is when a write command is performed. If the write command doesn't occur, nothing in the SRAM changes, which is why it's called static. The benefit of SRAM is that it's much faster than DRAM, reaching speeds of 12 ns as compared with BEDO's 50 ns. The disadvantage is that SRAM is

THE MEMORY GUIDE

Mark E. Donaldson

much more expensive than DRAM. SRAM's most common use in PCs is in the second-level cache, also called the L2 cache.

L2 Cache

Caching is the art of predicting what data will be requested next and having that data already in hand, thus speeding execution. When your CPU makes a data request, the data can be found in one of four places: the L1 cache, the L2 cache, main memory, or in a physical storage system (such as a hard disk). L1 cache exists on the CPU, and is much smaller than the other three. The L2 cache (second-level cache) is a separate memory area, and is configured with SRAM. Main memory is much larger and consists of DRAM, and the physical storage system is much larger again but is also much, much slower than the other storage areas. The data search begins in the L1 cache, then moves out to the L2 cache, then to DRAM, and then to physical storage. Each level consists of progressively slower components. The function of the L2 cache is to stand between DRAM and the CPU, offering faster access than DRAM but requiring sophisticated prediction technology to make it useful. The term cache hit refers to a successful location of data in L2, not L1. The purpose of a cache system is to bring the speed of accessing memory as close as possible to the speed of the CPU itself.

Async SRAM (Asynchronous SRAM)

Async SRAM has been with us since the days of the 386, and is still in place in the L2 cache of many PCs. It's called asynchronous because it's not in sync with the system clock, and therefore the CPU must wait for data requested from the L2 cache. The wait isn't as long as it is with DRAM, but it's still a wait.

Sync SRAM (Synchronous Burst SRAM)

Like SDRAM, Sync SRAM is synchronized with the system clock, so it's faster than the Async SRAM commonly used for L2 caches, with speeds of about 8.5 ns. Unfortunately, Sync SRAM isn't being produced in sufficient quantities to drive its cost down, so it seems destined for a relatively short life. That's especially true because it loses the ability to synchronize at bus speeds higher than 66 MHz. For the new breed of machines, therefore, let's welcome PB SRAM.

PB SRAM (Pipeline Burst SRAM)

Using burst technology, SRAM requests can be pipelined, or collected so that requests within the burst are executed on a nearly instantaneous basis. PB SRAM uses pipelining, and while it's slightly behind system synchronization speeds, it's a possible improvement over Sync SRAM because it's designed to work well with bus speeds of 75 MHz and higher. Look for PB SRAM to be a major player in Pentium II systems and beyond.

VRAM (Video RAM)

VRAM is aimed precisely at video performance, and you'll find it primarily on video accelerator cards or on motherboards that incorporate video technology. VRAM is used to store the pixel values of a graphical display, and the board's controller reads continuously from this memory to refresh the display. Its purpose is not only to give you faster video performance than you'd get

THE MEMORY GUIDE

Mark E. Donaldson

with a standard video board, but to reduce strain on the CPU. VRAM is dual-ported memory; there are two access ports to the memory cells, with one used to constantly refresh the display and the other used to change the data that will be displayed. Two ports means a doubling of bandwidth, and faster video performance as a result. By comparison, DRAM and SRAM have only one access port.

WRAM (Windows RAM)

Like VRAM, WRAM is a dual-ported type of RAM and it is used exclusively for graphics performance. WRAM is similar to VRAM in its operation, but it offers a higher overall bandwidth (roughly 25 percent higher), in addition to several graphics features that applications developers can exploit. These include a double-buffering data system several times faster than VRAM's buffer, resulting in considerably faster screen refresh rates.

SGRAM (Synchronous Graphics RAM)

Unlike VRAM and WRAM, and despite the fact that its primary use is on video accelerator cards, SGRAM is a single-ported RAM type. It speeds performance through a dual-bank feature, in which two memory pages can be opened simultaneously; it therefore approximates dual-ported. SGRAM is proving to be a significant player in 3-D video technology because of a block-write feature that speeds up screen fills and allows fast memory clearing. Three-dimensional video requires extremely fast clearing, in the range of 30 to 40 times per second.

L2-Cache Specifications

There are two components to the L2 system cache in a computer system: the cache RAM itself and the Tag SRAM. The rated speeds you'll need for these two different components will depend on how you have your L2 cache configured. To be absolutely certain, get the specification directly from your motherboard manufacturer.

Your 100-MHz Pentium uses a system clock speed of 60 MHz, so 17-ns SRAM modules should suffice for that processor. The Tag SRAM should be either 10 or 15 ns, depending on whether you configure the L2 cache as asynchronous or burst. If you're planning to upgrade to a 133-MHz Pentium (which uses a 66-MHz system clock speed, as do the 166- and 200-MHz Pentiums), you'll want to use 15-ns modules for the cache SRAM and 7 or 15 ns for the Tag SRAM.

Note that some motherboards may come with some L2-cache SRAM already soldered in place on the motherboard, and even though there is a COAST-compatible socket present, you cannot add memory to the cache. Again, check with the motherboard manufacturer for details on your system's specifications.