

## Memory-Table of Contents

<b>Topic</b>	<b>Page</b>
Table of Contents	1
Memory Introduction	2
A Simple Example	2
Memory Types	3
A Real Life Example	4-5
Type of Data Stored in the Cache	5
RAM	5
Memory Cells	5
Common Types of RAM in a Computer	6
RAM Performance	6
Clock Speed	7
Latency	7-8
SD-RAM	8
Added Performance	9
Enhanced SD-RAM	9
DDR-SDRAM	10
Important Improvements	10
Different Naming Scheme	11
RD-RAM	11-12
MHz vs. Performance	12
Hot, Hot, Hot	12

## Memory Introduction

Memory is a very important part of the computer. Although memory is technically any form of storage on a computer, the word is mostly used to refer to fast, but temporary type of electronic storage. If there was no memory, your CPU would have to take all of its information from the hard drive and other forms of permanent storage, which are very slow compared to today's memory speed standards.



A variety of computer memory modules.  
© www.kingston.com, 2001

### **A Simple Example**

To demonstrate why memory is used to make a computer, I will use a very simple example of two students who are currently going to high school. We will use John and Jessica as the names for the two students. Since both of the students have many textbooks and notebooks, they are assigned a locker in which they can store their books and equipment. It would just be too hard to haul all of their books to every one of their classes, as well as to their homes.

John is very lazy, and never plans ahead. He never comes to any of his classes prepared. Whenever his teacher switches the subject within the period, John is forced to go to his locker and bring the required equipment. When he is at his locker, instead of bringing everything he will need for the rest of the class, he only brings what he will need the current part of the class. This wastes much of his time throughout the day, because he constantly has to make trips to his locker, which by the way, is on the other side of the school. Sometimes, he even has to go home to get some of his things, which wastes even more time. If only John could be more like Jessica, he would get things done faster, and save the energy of making hundreds of trips a week to his locker.

Jessica, on the other hand, plans ahead. When she goes to her classes, she takes everything she needs for her class. This way, when the teacher switches the subject, all Jessica has to do is pull the required equipment out of her desk, and she is all ready to go. When the period ends, she makes a short trip to her locker. She unloads all of her things from the last period, takes everything she needs for the next period, and she is all set.

Although John and Jessica are not computers, this example shows the advantages of having temporary memory. The computer stores data it accesses the most in the memory, so that the computer runs faster. If it had to take things out of the hard drive every time it performed an operation, operations would take a very long time to complete, and much of the time would be wasted. John would represent a computer with no memory, while Jessica would represent a computer with a sufficient amount of memory.

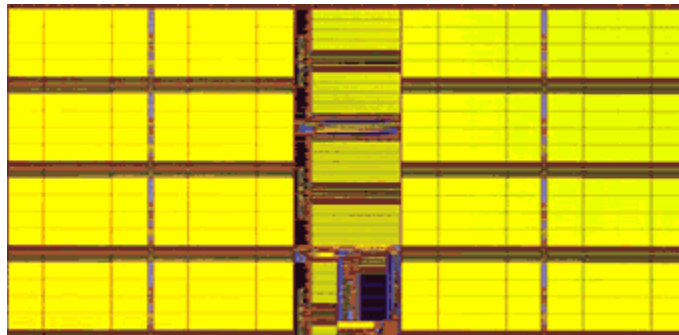
## Memory Types

There are many different types of memory in the computer. Each type of memory has its own role to fulfill. Although most memory types are built roughly the same way, there are some differences. This is why we will be discussing the different memory types, including:

- Cache
- RAM
- SD-RAM/SD-RAM based technology
- RD-RAM

As well, we will be talking about the different performance factors for memory and which memory is best suited for certain purposes.

If you have gone through the CPU section, you may already be familiar with the computer's Cache. A very important part of both the CPU and the whole computer, it is the fastest type of memory available to our computer. On the downside, it is very expensive to make, so not many CPUs have a large amount of it. You will soon learn how this small amount available to our computer greatly improves the performance.



This is 1/4 MB of Cache from the Intel Pentium III Processor  
© Intel Corp, 2001

### A Real Life Example

To demonstrate how Cache is used, I will once again use Jessica, the very efficient high school student.

When Jessica is at a particular class, we already know that she always brings the required equipment and books. We also know that Jessica may not always be using all of her textbooks and notebooks at the same time. To make things more organized, she would have the things she does not need right away inside her desk. In the same way, the computer stores the things it does not need, but will need soon, in the Ram. If Jessica is writing a note that her teacher is dictating to the class, she would have a notebook as well as her writing equipment out. She would have a pencil in her hand because this is what she would use the most. On her desk, she would have other writing utensils such as an eraser, highlighter, ruler, etc. There is no point in constantly keeping these in her hand because she only requires it from

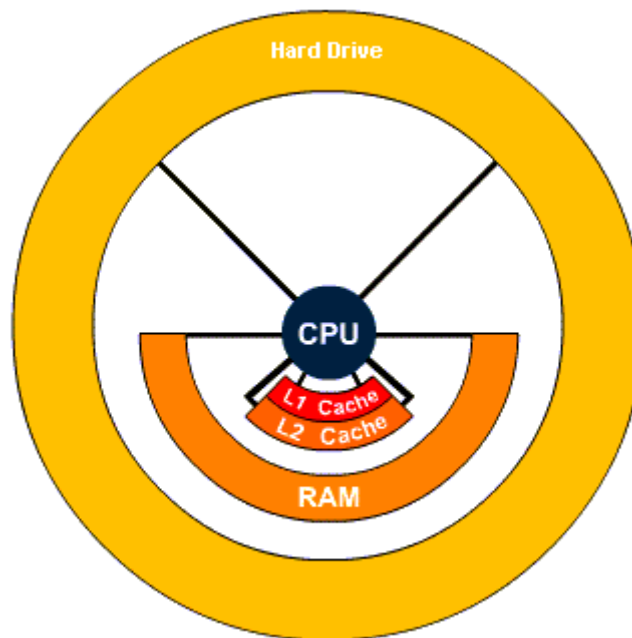
time to time. She keeps them *on* the desk, because she does require it more than the equipment *in* her desk.

Much like Jessica's desktop and hand, the Cache is used to store the things that are needed the most by the CPU. Level 1 Cache, like her hand, is used to store the data that will be needed the most by the CPU. The L1 Cache is very expensive to make, so only a small amount of it is available.

The desktop, which has all of the equipment Jessica will need from time to time was used to represent Level 2 Cache. It stores more data that the CPU will need frequently, but not as frequently as the data stored in the L1 Cache. L2 Cache is similar to L1 Cache, except that it is slower and less expensive to make. There is usually much more L2 Cache than L1 cache for these reasons.

If the information needed is not in the Cache, the computer looks in the RAM, and if it not available there, it takes it from permanent storage (usually the hard drive).

To demonstrate the above in a visual manner, here is a Pyramid representing the different levels of memory, and when they are accessed.



The CPU first checks if the required data is in the L1 Cache.
If the required data is not found in the L1 Cache, the CPU goes to the L2 Cache.
The CPU now goes to the RAM, because the required data has not been stored in the L1 or the L2 Cache.
Finally, if the required data cannot be found in the RAM, the CPU has no options but to access the comparatively slow Hard Drive.

As you can see, the CPU tries to see if the data it requires is in the L1 or L2 Cache. If it is not, it has no choice but to go to the slower RAM. In the case that the data it needs is not in either of these types of memory, it has to access them from permanent storage types, such as a CD-ROM Drive, or a Hard Drive.

## Type of Data Stored in the Cache

What type of data would need to be accessed over and over again? Well, this can vary depending on the program used. For example, in a Word Processor, the font that is currently used may be stored in the cache because it has to be accessed every time the user inputs any letter or number. In programming, loops are used very commonly. Loops let the computer know that it has to execute a block of code a certain number of times, which can be anywhere from two to thousands of times. You can see how much time the computer would save by not having to go to the much slower RAM thousands of times just to complete one block of code.

## RAM

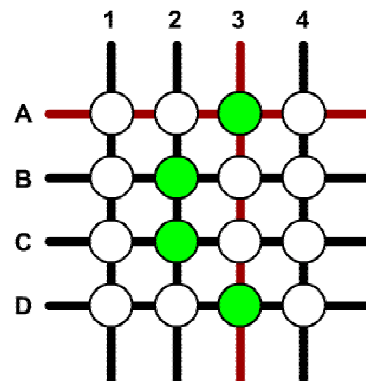
One of the most common memory types in your computer is Random Access Memory, or RAM. This memory is considered random access because the computer can access any part of the memory whenever it wants, which means that it is not restricted to go in any certain order.

Just like any other part of a computer, memory is a circuit which is made of millions of transistors and capacitors. In computer memory, every capacitor is paired up with a transistor. The capacitor holds the information, while the transistor is used to access or change the information that the capacitor is holding. To store a 1 in the capacitor, it is filled up with electrons. To store a 0 in the capacitor, it is emptied so that there are no electrons in it. This is all done at a very rapid pace, so that the computer can process the information as quickly as possible.

## Memory Cells

Memory cells hold the information, but how is this achieved? Well, they are arranged on a 2 dimensional array. To change the state of a cell from 0 to 1, the column and row of that cell are charged, and the capacitor gets filled where the column and row meet, which is where the capacitor is located. Similarly, to change the state of a cell from 1 to 0, the column and row are charged, in order to open a circuit, allowing the electrons to escape from the capacitor. To help you understand this a little better, here is an illustration of how a capacitor holds a binary 1 and a binary 0.

I will use the Capacitor in row A, column 1 to explain. When you click on it the first time, row A is charged up, and column 1 is charged up. Only the point of intersection between the charged row and column becomes a 1. When the same row and column become charged up again, the capacitor releases the electrons, and now holds a binary 0. The other capacitors which are in that row or column remain unchanged.



### Common Types of Ram in a Computer

Although all computers have RAM, the type of RAM that they have can change from computer to computer. The type of RAM used can vary depending on one of these factors:

- the performance the computer needs
- type of RAM the computer can support
- speed of the system bus
- type of budget the Computer is built for

Keeping this in mind, here are some of the most common types of RAM in PCs.

<b>EDO-RAM</b>	A type of RAM which does not wait to process one bit before going to the next. This RAM is not used for newer computers, although it is in many Pentium II or lower class computers.
<b>SD-RAM</b>	RAM which replaced EDO-RAM in order to increase speed. At the same clock rate, it is almost the same speed as EDO-RAM. It was supposed to be faster because it uses burst mode to read data. This basically means that after it read the data the CPU required, it continued reading data after this in anticipation that the CPU will need the next piece of data in the memory right after the current data is processed. It is somewhat effective, but the faster clock frequency is what really makes it reach better performance.
<b>RD-RAM</b>	A type of RAM developed to operate much differently than other RAM types. Rambus created its own high-speed data bus, Rambus Channel, and made the RAM work in parallel (have more than one stream of information coming from the chip) in order to achieve phenomenal clock speeds compared to the competition. In reality, it is actually slower and more expensive than SD-RAM, which is a topic to be discussed later in the memory section.

Although RAM is used in computers, it is not limited only to them. For example, most health cards in the developed world have RAM to keep the cardholder's health information. TVs, Radios and even some Microwaves have RAM in order to keep custom information, such as favourite radio stations.

### RAM Performance

Over the years, computer speeds have been going up at a very rapid pace. For the past few years, the trend has been that computers double in speed after roughly 18 months. For RAM specifically, this is not true at all. Why? What is causing the RAM speeds to evolve at a much slower pace than the CPU, Cache or even CD-ROM drives? To answer this question, we have to take a look at the two main RAM performance factors: clock speed and latency.



This new RAM module has lower latency than some models from 3 years ago.

© www.mushkin.com, 2001

### **Clock Speed**

RAM clock speed has kept pace with the Computer's main bus. They do not need to be any higher because if they do go any faster, they will be outpacing the different components that need to use it. Also, in order to improve clock speed greatly, some other factors suffer. Since too much clock speed does not necessarily give more performance, the memory will be even slower because other parts had to be simplified in order to accommodate the clock speed, and performance is reduced. Improving the clock rate by a huge amount has been attempted successfully by Rambus (their memory will be talked about later in the memory segment), and it has proven that it does not make any difference.

### **Latency**

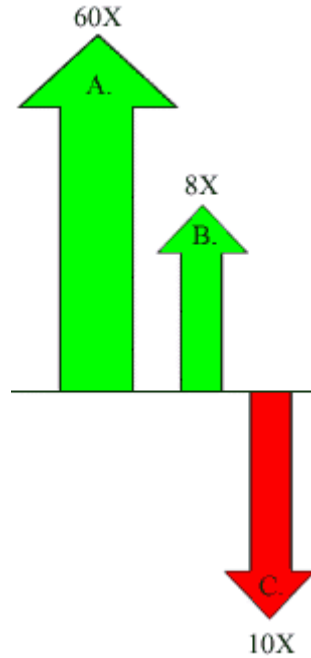
If it is not clock speed that slows something down, it has to be latency. Latency is the time that it takes for the memory to start sending data to other parts such as the CPU. The higher the latency, the lower the memory performance. Latency is very important, especially because information stored from the RAM is needed in many short bursts. This is because Cache data is checked first, and if the required data is not there, then the CPU checks in the RAM. Many times, the required data is in the Cache, because the Cache does store the most widely used information. Because there are so many short bursts, every time this short burst of data starts, the CPU has to wait for the data to be sent.

If data was sent in very long bursts, latency would not be much of a factor to memory performance. If you do not understand the concept of latency, think of this. Drag races are very short, sometimes as little as a quarter of a mile, or a third of a kilometre. It is therefore very important for the cars to accelerate very quickly, because chances are that they will already finish the race before they hit their top speed. This is why latency is important, because there are many short bursts of data, and long streams of data that are uncommon.

If a car was racing for 300 kilometres in a straight line, acceleration would not be important, top speed would be much more important. The same would be true if there were very long bursts of data, latency would not be a very big factor.

With the current way the computers are set up, latency is very important. Over the past few years, latency has become worse and worse, something which is very uncommon to computer evolution.

## 486 Era Compared to Today



A.	CPU Speed
B.	BUS Speed
C.	Latency

As you can see, BUS speed has gone up 8X from the 486 era. CPU speed is 60X larger than the 486 era. Latency is **10X worse** than in the 486 era. To make up for the bad latency, engineers are making much higher RAM speeds, and adding a larger amount of memory. Think how fast memory would be if all this was done, and latency was as good as it was 5-10 years ago.

**SD-RAM**

In the early 1990's, most PCs were equipped with EDO-RAM. Although EDO memory was a very good type of memory, it also had some very big inefficiencies. Suddenly, just making the EDO-RAM faster was not good enough. EDO-RAM is asynchronous, which means that it does not necessarily run at the same clock rate as the rest of the computer. This means that a clock would have to be used just for the memory, plus the CPU would spend much of its time doing nothing but waiting for the data to get sent.

To fix this problem, SD-RAM (Synchronous Dynamic RAM) was created. This type of RAM was able to run at the same frequency as the rest of the computer.

### Added Performance

The early type of SD-RAM used a 2-clock. This means that it was set up so that each clock cycle it could get access to two of the chips from each memory stick. To improve performance, the SD-RAM was altered so that it used a 4-clock.

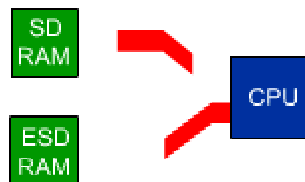
Another improvement that had to be made was the clock speed of the memory. When Intel CPUs started migrating to 100MHz buses, SD-RAM had to keep pace. Currently, speeds of SD-RAM are also available in 133Mhz, as well as 150MHz and 166MHz for high performance workstations. Today, SD-RAM is rated by the highest speed it can reliably run on. For example 133MHz SD-RAM would be referred to as PC133 SD-RAM.



Mushkin's 128MB high performance PC133 SDRAM  
© www.mushkin.com, 2001

### Enhanced SD-RAM

If there is a weakness to SD-RAM, it would have to be its latency. Although Memory modules seem to be getting faster and faster, latency, which is the real performance problem, is getting bigger and bigger. In order to overcome this problem, some manufacturers have actually added a small amount of high speed memory that acts as a Cache to the memory module. This effectively lowers the latency significantly, and improves the performance of the memory. Just like the CPUs cache, the goal of the memory Cache is to hold the most frequently used information. This performance gain does have a price, as ESD-RAM can be up to 4 times as expensive as regular SD-RAM modules.

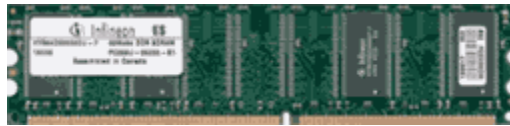


As you can see, ESD-RAM's added cache gives it an excellent performance gain over standard SD-RAM.

There have not been many higher performing types of SD-RAM at a relatively low price. One type is quickly becoming popular, theoretically being twice as fast as standard SD-RAM. It is called DDR SD-RAM, and the way it achieves the double data rate is actually quite simple and will be demonstrated in the next section.

## DDR SD-RAM

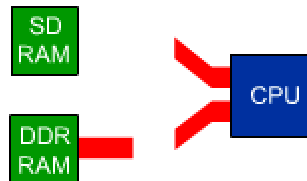
Double Data Rate SD-RAM is the next generation in SD-RAM memory. As the name implies, it can double the data rate that normal SD-RAM makes, which theoretically makes it twice as fast. Contrary to what some people think, although it is twice as fast as SD-RAM, it would not make a computer run twice as fast as the computer with normal SD-RAM. This is because the computer does not depend solely on memory for performance; there are a variety of other factors such as CPU performance, Video Card performance, etc. The computers that do need fast RAM for 3D modeling programs and other professional applications, DDR-SDRAM is a very good choice, because of its high speed, and the not so large price tag.



High speed Infineon DDR RAM Module  
© www.infineon.com, 2001

### Important Improvements

So how does DDR SD-RAM achieve twice the data transfer rate of standard SD-RAM? The concept is actually quite simple. We know that SD-RAM transfers data every clock pulse. To be more exact, it transfers data on every rising clock pulse. You see, the clock pulse rises, and then it falls, then a new one does the same thing over and over. DDR takes the advantage of the rising and falling clock pulses and treats them as though they were two clock pulses instead of one, allowing more data to be transferred. Here is a simple example:



Another important improvement that DDR has over its predecessor is that it consumes less power, and only needs 2.5 Volts from the voltage supply. This makes it a very attractive option for notebook computers, which need to consume as little power as possible in order to extend battery life.

DDR is also built similarly to standard SD-RAM, which means that it will not be as expensive as some of the other modifications to SD-RAM.

### Different Naming Scheme

DDR RAM was first named the same way as SD-RAM. For example, 100MHz DDR would be called PC200 because of its double data rate. Since Rambus, a DDR competitor, decided to name its memory modules PC800 even though they are not running at 800MHz, DDR needed a new naming scheme. They used the amount of Megabytes that DDR can transfer every second. They took the width of the BUS and multiplied it by the number of clock cycles per second, and multiplied this by two (because of the DDR property) to figure out the transfer rate in MB/sec:

**64-bit bus \* 100MHz (1,048,576) \* 2 times the transfer rate = 134217728 bits/sec.**

This is now converted into bytes by dividing the number by 8:

**134217728 / 8 = 1677721600 bytes/sec.**

Now, to convert this to MB per second, the number is divided by the amount of bytes in 1 MB (1,048,576)

**1677721600 / 1,048,576 = 1600 MB/sec.**

Since they knew that 100MHz DDR RAM transfers at 1600 MB/sec, they decided to call it PC1600.

Here are some more of the popular speeds of DDR-RAM, and what they are named:

Speed Rating	Common Name
133 MHz	PC2100
150 MHz	PC2400
166 MHz	PC2700
200 MHz	PC3200

RAM has advanced quickly over the years, and DDR is one of the newer and more promising additions to the SD-RAM family.

### RD-RAM

RD-RAM stands for Rambus Direct RAM. Rambus has been in the news many times over the past year. In 1996, it decided to take a radically different approach to memory management for their RD-RAM. They have tried to create the fastest RAM. Some people feel that they were successful, because after all, they do have the fastest RAM when it comes to MHz. The only problem is that they do not have the fastest RAM when it comes to performance. This, combined with a very expensive price tag, has made Rambus very unsuccessful. Even Intel, Rambus' all time supporter, seems to be giving up on it. The early Pentium 4's all had to use Rambus memory. Now, Intel is slowly migrating to SD-RAM, which is expected to take 70% of the Pentium 4 Market. Now that we know some of the background, let's see how Rambus' new approach works.

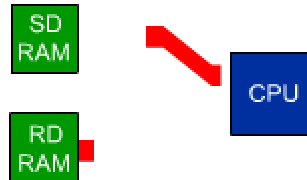


PC800 RD-RAM Module  
© www.mushkin.com, 2001

### MHz vs. Performance

Rambus tried to use an approach that has worked for Intel very well. They tried to make the clock speed much faster, so that people would buy their products. In order to change the MHz by an enormous amount, they had to make other parts more inefficient. To make their PC800 Memory run at 400MHz (they have a DDR approach too, and technically, their memory runs at 400MHz), they had to cut down the bus width to 16-bit. If you remember from the CPU history, 16-bit dates back to the 286 era, and even the 386 uses a 32-bit bus. Technically,  $1/4$  the bus width \* 8 times the clock rate would still make RD-RAM transfer twice as fast as PC100 SD-RAM. In order to reduce 32-bit information into 16-bit and then change it back to 32-bit, performance suffers greatly.

As well, Rambus RAM has one of the biggest latency times, making it very inefficient. Rambus RAM is not as fast as typical PC150 SD-RAM, but the real problem is the price. RD-RAM can be up to 3 times more expensive than standard SD-RAM modules. Here is the impact that RD-RAM's poor latency has on its performance compared to standard SD-RAM:



As you can see, latency is very important in today's memory.

### HOT, HOT, HOT

So what has the enormous clock speed done to the temperature of the RD-RAM module? It has brought it high enough that it has to use a heat sink (the blue metal covering up much of the memory you see above is used as a heat sink). Making the clock speed fast before companies do it has some disadvantages. Rambus is not all bad though. It is an excellent concept, and they are very brave to try something new. When you fail, you have to stand up and try again. This is precisely what Rambus is doing. They are working on even higher speeds for their memory modules, and RD-RAM will surely give DDR-RAM a run for its money in the near future.