

ITFreeTraining



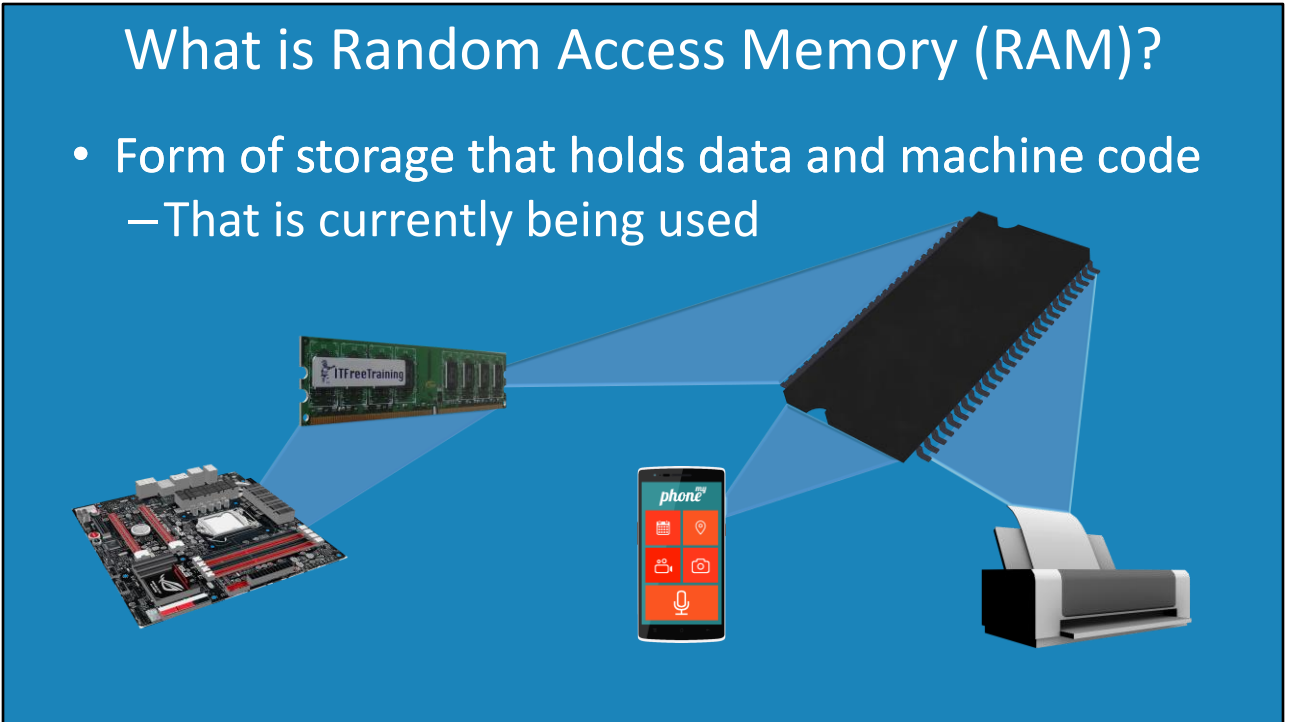
Basics of How RAM Works

For the free video please see
<http://itfreetraining.com/ap/3a04>

Welcome to the ITFreeTraining video on the basics of how Random Access Memory or RAM works. Since RAM was first created, there have been major changes in how it operates. Having an understanding of how RAM works will give you an idea of what to expect from different RAM types.

What is Random Access Memory (RAM)?

- Form of storage that holds data and machine code
– That is currently being used



0:18 So, what is RAM? RAM is essentially a form of storage that holds data and machine code. The difference from other types of storage is that RAM is currently in use by the computer and thus needs to be fast. By fast I mean, the speed of RAM is measured in milliseconds.

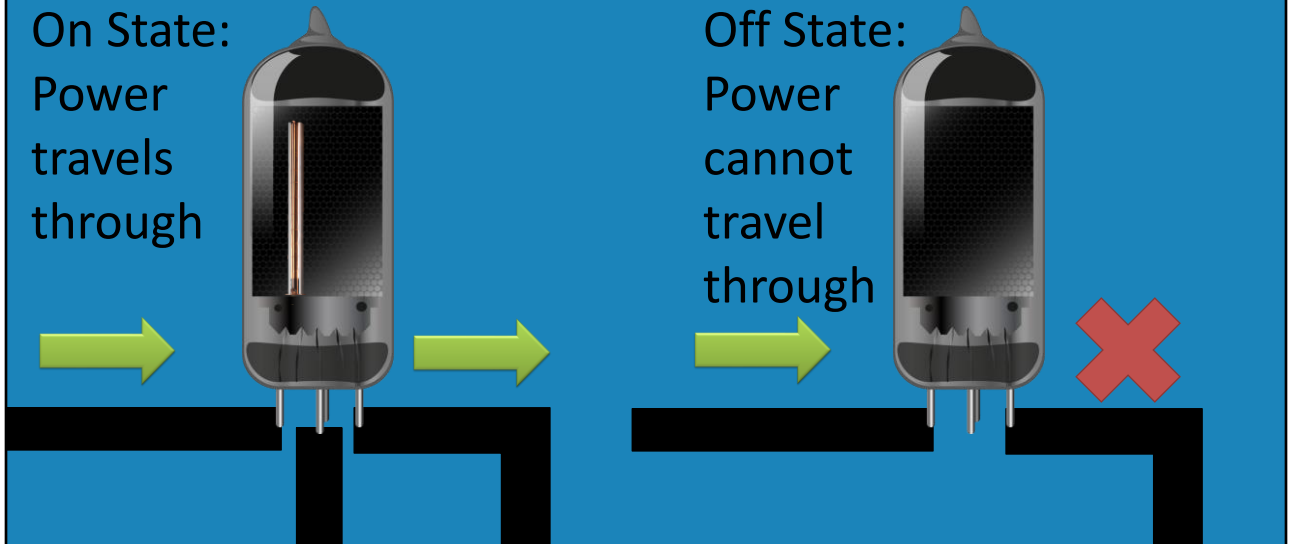
Inside your computer is a motherboard. This motherboard will contain one or more memory modules. Memory modules are also commonly referred to as memory sticks or sticks of RAM. These memory modules hold the data and software the computer needs to operate.

On the memory module will be a number of chips called DRAM chips or Dynamic RAM chips. The DRAM chips are what holds the data. Although there are many DRAM chips and they have improved over the years, these DRAM chips are the same kind that you would find in your hardware devices or mobile devices.

Essentially any device that runs code and needs to store data will have some kind of memory in it. In small hardware devices, the memory is often soldered onto the printed circuit board or PCB. In most computers, the memory modules can be replaced. However, in some computers the memory will be soldered onto the PCB. To understand how modern memory works, I will first look at one of the first ways that memory was implemented.

Vacuum Tubes

- Originally used in computers



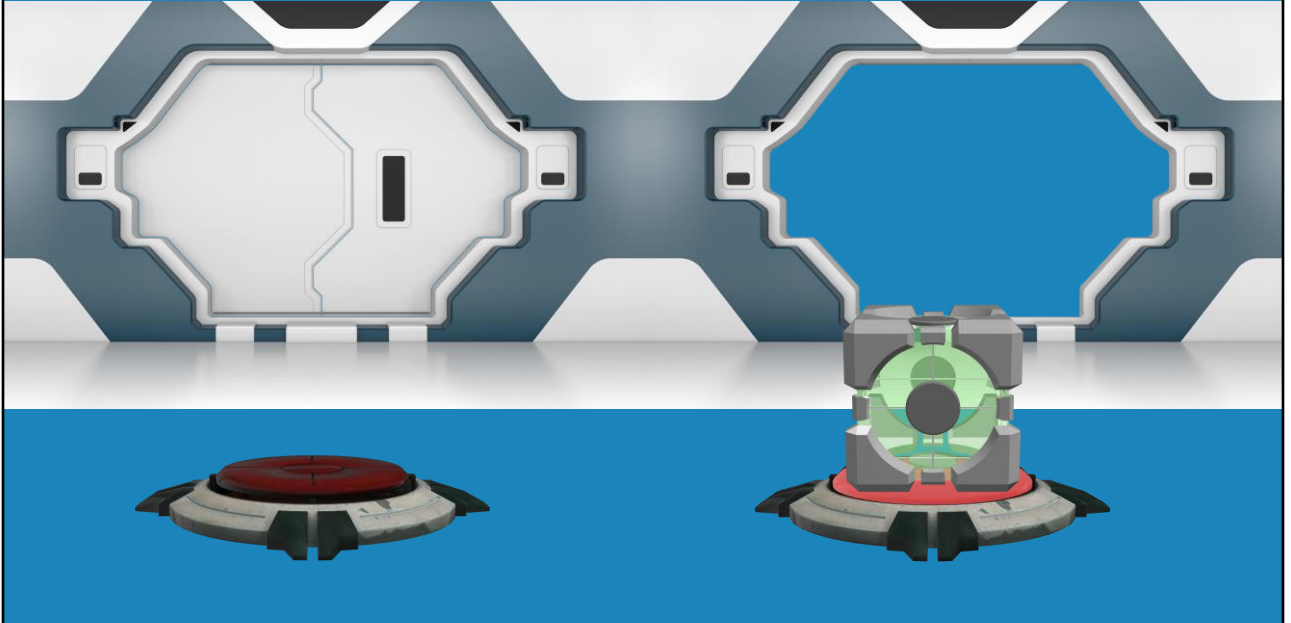
1:42 One of the first methods of storing data in a computer was with the use of vacuum tubes, back in the 1940's. A vacuum tube is essentially a glass tube with no air inside, just like a light globe. The glass tube contains a filament.

If the filament is connected to a power source, the filament heats up. Now consider that you have two wires connected to the vacuum tube. If power is applied to the left and the filament is active; power will be allowed to flow through to the right.

If power is applied, this essentially mean that the vacuum tube is in the 'on' state. Next consider what happens if I have a second vacuum tube in which the filament is not on. When power is applied, this time the power will not be allowed through. This is called the 'off' state.

What the vacuum tube allowed was for a basic switch to be created, which could be in the on or off state. Or to think of it another way, it could be one or zero. This is the basic concept of how a transistor works. Transistors are a fundamental part of electronics. To understand this better, let's look at a different example.

Example



2:56 The basics of a transistor work like this; consider that you have a door that is connected to a button, a bit like what you may find in a puzzle style computer game. If the button is not being held down the door remains closed. However, if something were to press the button and hold the button down, the door would open and remain open until the button is released. If the button is no longer being held down, the door will close.

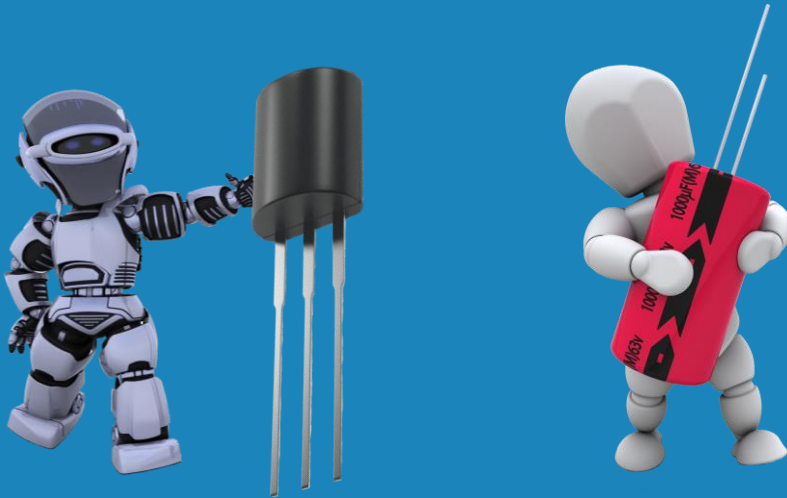
This is the basic fundamentals of how a transistor works. When multiple transistors are combined together, more complex results can be achieved. In order to store one bit of information you may think you only need one transistor. You also need to consider that you need a way of accessing the bit and changing the bit. To store a single bit and control functions can take four or more transistors depending on what method is used.

Nowadays, a computer CPU can contain over a billion transistors in a single chip, so using minimum of four transistors does not sound like a lot. However, if you consider a gigabyte of memory would be eight billion bits, now times this by four, this would give 32 billion transistors. You can start to understand why there is only a small amount of RAM in a CPU and the majority of the RAM is found on memory modules outside the CPU.

The advantage of transistor-based RAM is that it is very fast and thus why it is used in CPU's. The disadvantage is that it requires a large number of transistors in order to operate. More transistors mean that it is harder to make, is bigger in size or the size of each transistor needs to be reduced which leads to manufacturing problems.

Capacitor Based Memory

- Uses one transistor and capacitor per bit



4:47 In order to reduce the number of transistors required for computer memory, a new type of memory was developed. This utilizes a transistor and a capacitor. A capacitor is an electronic component which can hold a charge. This is similar to a battery; however, a battery holds a charge for a long time while a capacitor holds a charge for only a short period.

In electronics, a capacitor is used for many different things, but in the case of memory it works like this. If the capacitor is full it represents 'one' bit. If the capacitor is empty it is a 'zero' bit. Thus, a one or a zero bit can be stored by either filling the capacitor with a charge or emptying it. Let's have a closer look at how this works.

Capacitor Empty

- Voltage is low or no voltage represents a 'zero' bit



5:36 Consider that you have a water tank. The tank is like a capacitor. A capacitor stores power like a tank stores water. In the case of a capacitor in a memory chip, if there is no voltage or the voltage is low this represents a zero bit.

Capacitor Full

- Voltage is high or full represents a 'one' bit

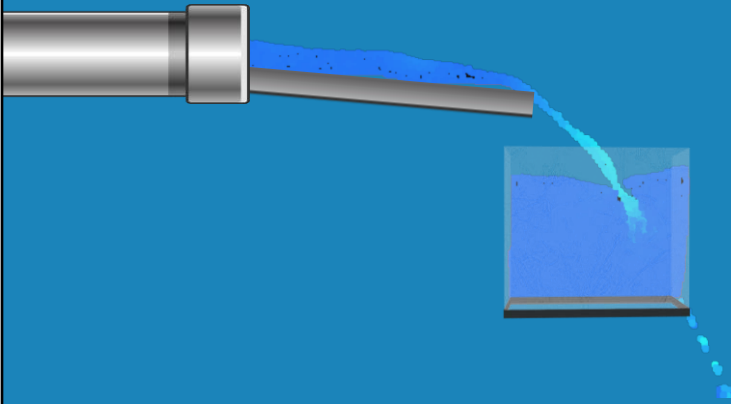


5:52 I will now fill the water tank. You probably guessed by now, when the water tank is full, this will represent a one bit. In a capacitor, if the voltage is high or the capacitor is full, this will represent a one bit.

This seems pretty simple, but unfortunately like most things in computing it is not that simple. Capacitors, as time goes by, lose their charge. In the case of memory capacitors this will be in milliseconds, so not that long at all.

Capacitor Refresh

- Periodically charge is added to the capacitor



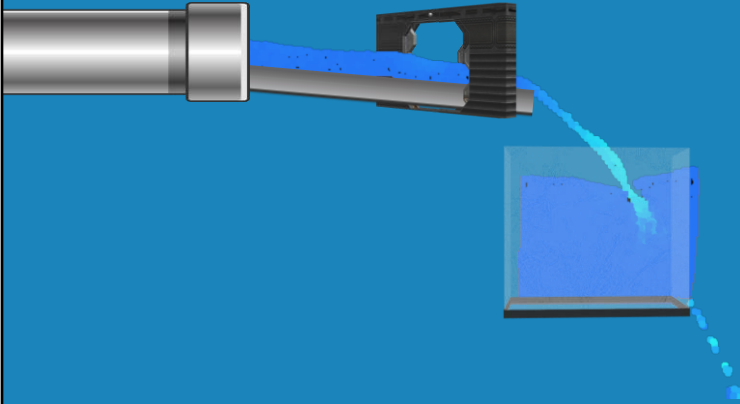
6:22 To think of this in simple terms, consider that our water tank now has a leak. The water will drain out of the tank. If the water level gets too low, the water tank will now represent a zero bit instead of a one bit.

To prevent this from occurring, periodically charge is added to the capacitor. As long as more charge is added before too much charge is lost, the capacitor will not change its state from a one to a zero.

This works well, but what happens if the capacitor holds no charge and thus is a zero bit. In this case, we want to prevent charge going into the capacitor. How is this achieved?

Transistor On State

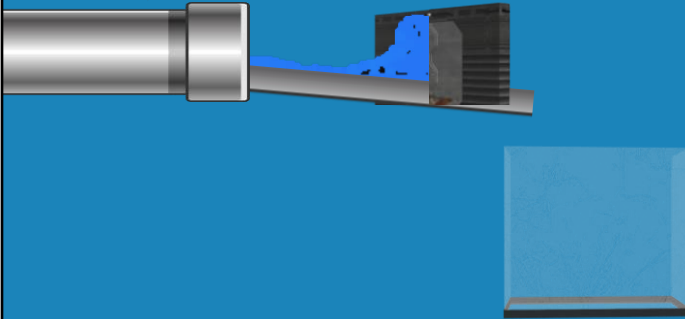
- Allows charge to go into the capacitor



7:00 To have some control over whether the capacitor is topped up with charge, this is where the transistor comes into play. In this example, I have used a door to represent the transistor. In the open state, the door will allow water to flow to the tank. In a memory module, the transistor is connected to the capacitor and thus if the capacitor is charged, like in this case, the transistor will allow power to travel to the capacitor and charge it.

Transistor Off State

- Does not allow charge to go into the capacitor



7:27 Now let's have a look at the case of when the capacitor is empty. To do this, it is a simple matter of draining the power from the capacitor or the tank of water in this case. Once this is done, the transistor is no longer powered and thus will not allow power to the capacitor. In this example, the doors will close and water will no longer be able to flow into the capacitor.

This is the basics of how capacitor-based RAM works. Let's compare transistor-based and capacitor-based RAM to have a better understanding of when they may be used.

SRAM vs DRAM

Static RAM (SRAM)


Large number of transistors
Costs more per bit
Faster than DRAM

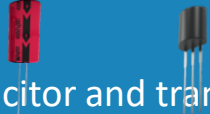


Used for CPU cache

Requires constant power



Dynamic RAM (DRAM)


One capacitor and transistor
Costs less per bit
Slower than SRAM



Used for memory modules

Requires power to be refreshed



8:03 Transistor-based RAM is commonly referred to as static ram or SRAM. Capacitor-based RAM is often referred to as Dynamic RAM or DRAM. SRAM requires a large number of transistors. Because of this, the chips need to be larger or each transistor on the chip needs to be smaller.

DRAM in contrast, requires one capacitor and one transistor per bit. This may not sound like a lot, but when you start looking at gigabytes of RAM, this adds up pretty fast. Essentially this means that SRAM costs more per bit to manufacture than DRAM which therefore costs less to manufacture per bit.

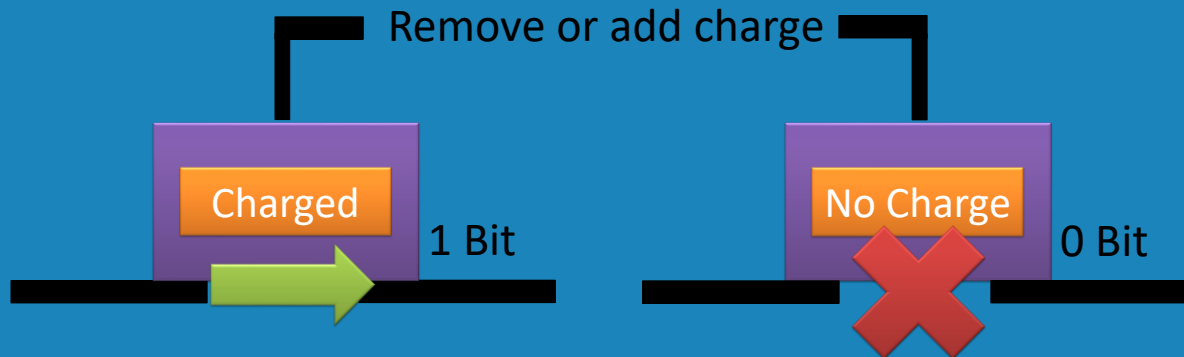
If cost was the only concern then you would always use DRAM. However, SRAM is faster than DRAM. This means that there is a trade-off between cost and performance. It is possible for engineers to use either SRAM or DRAM, but generally SRAM and DRAM are used for particular types of applications. For example, SRAM is generally used in the CPU for cache. This is because only a small amount of RAM is required and it needs to be as fast as possible. In contrast, DRAM is used for memory modules. Although SRAM can be used, due to its much higher cost, DRAM is commonly used for memory modules.

The last difference that I would like to highlight between SRAM and DRAM is that SRAM needs constant power to operate. In contrast, DRAM requires periodic power in the form of a refresh to be performed. While the refresh is occurring, the DRAM cannot be read or written to. This is one of the reasons why DRAM is typically slower than SRAM.

Both these RAM types are what is referred to as volatile memory. Volatile memory refers to memory that if power is lost the data is lost. However, some memory keeps its data even when it is not powered, this memory is referred to as non-volatile. That is, it keeps its data when power is switched off. How does this differ from SRAM and DRAM?

Non-volatile RAM

- Flash Based RAM (USB/Solid State Drives SSD)
- Holds charge in a floating gate



10:12 The most common non-volatile RAM used nowadays is flash based RAM, commonly used in solid state drives also referred to as SSD. This kind of RAM combines the capacitor and transistor into one. To achieve this, it holds charge in what is referred to as a floating gate.

I have simplified the process down, but it essentially works like this. A bit of data is held in what is referred to as a cell. Inside the cell a charge can be held. This is different from a capacitor in that the charge is held inside an insulation material. The insulation material makes sure the charge cannot escape. Unlike a capacitor, once the charge is in the cell, it does not require power. The cell is estimated to hold a charge, depending on what temperature the storage is kept at, for 20 to 100 years.

Like a transistor, there is a connection in and out of the cell. This acts like a switch; if there is charge in the cell, power is allowed to flow through which means the cell is holding a one bit. In contrast, if the cell is not holding a charge, the power will not be allowed through. This means the cell is holding a zero bit.

In order to change the charge in the cell, another connection is made to the cell. This connector can either add a charge or drain the cell so there is no charge. This is a simple version of the process, but you should get the idea. You can see that flash RAM is a simple design compared with DRAM and in particular SRAM. So why would we not use it in preference to those other technologies?

Flash vs SRAM and DRAM

- Performance (Slow reads and very slow writes)
- Problems reducing the size of flash RAM



Reads/Writes slow



11:54 Flash RAM compared to SRAM and DRAM generally has a much slower read and write performance. In particular, write performance is very slow compared to read. This is because it takes time to ensure the cell is either filled up or completely removed of charge. Read performance is generally good, however, not as good as SRAM or DRAM.

Next there are problems with reducing the size of flash RAM. Due to the demand of flash RAM, the reduction of its size has been faster than other technologies have been; however, there are problems with the manufacturing process as it gets smaller. For example, if the cell is made too small, there are difficulties charging it and keeping it charged. This leads to an increase in errors and failures.

For these reasons, flash RAM generally gets used for storage. Flash RAM is often used to store the contents of the BIOS so it is preserved during reboots. Now days, as Flash RAM has increase in speed and lowed in cost, it also has been replacing traditional hard disks for storage.

That concludes this video on the basics of how RAM works. I hope to see you in the other videos from us. Until those videos, I would like to thank you for watching.

References

“CompTIA A+ Certification Exam Guide Ninth Edition” pages 138-154

“Flash memory” https://en.wikipedia.org/wiki/Flash_memory

“Nano-RAM” <https://en.wikipedia.org/wiki/Nano-RAM>

“Memory cell (computing)” [https://en.wikipedia.org/wiki/Memory_cell_\(computing\)](https://en.wikipedia.org/wiki/Memory_cell_(computing))

“Flash memory” https://en.wikipedia.org/wiki/Flash_memory

Credits

Trainer: Austin Mason <http://ITFreeTraining.com>

Voice Talent: HP Lewis <http://hplewis.com>

Quality Assurance: Brett Batson <http://www.pbb-proofreading.uk>