



Memory & Storage: Crash Course Computer Science #19

Crash Course: Computer Science

<https://youtube.com/watch?v=TQCr9RV7twk>

<https://nerdfighteria.info/v/TQCr9RV7twk>

Hi, I'm Carrie Anne, and welcome to Crash Course Computer Science!

We've talked about computer memory several times in this series, and we even designed some in Episode 6. In general, computer memory is non-permanent.

If your Xbox accidentally gets unplugged and turns off, any data saved in memory is lost. For this reason, it's called volatile memory. What we haven't talked so much about this series is storage, which is a tad different.

Any data written to storage, like your hard drive, will stay there until it's over-written or deleted, even if the power goes out. It's non-volatile. It used to be that volatile memory was fast and non-volatile storage was slow, but as computing technologies have improved, this distinction is becoming less true, and the terms have started to blend together.

Nowadays, we take for granted technologies like this little USB stick, which offers gigabytes of memory, reliable over long periods of time, all at low cost, but this wasn't always true. INTRO The earliest computer storage was paper punch cards, and its close cousin, punched paper tape. By the 1940s, punch cards had largely standardized into a grid of 80 columns and 12 rows, allowing for a maximum of 960 bits of data to be stored on a single card.

The largest program ever punched onto cards, that we know of, was the US Military's Semi-Automatic Ground Environment, or SAGE, an Air Defense System that became operational in 1958. The main program was stored on 62,500 punchcards, roughly equivalent to 5 megabytes of data, that's the size of an average smartphone photo today. Punch cards were a useful and popular form of storage for decades, they didn't need power, plus paper was cheap and reasonably durable.

However, punchcards were slow and write-once, you can't easily un-punch a hole. So they were a less useful form of memory, where a value might only be needed for a fraction of a second during a program's execution, and then discarded. A faster, larger and more flexible form of computer memory was needed.

An early and practical approach was developed by J. Presper Eckert, as he was finishing work on ENIAC in 1944. His invention was called Delay Line Memory, and it worked like this.

You take a tube and fill it with a liquid, like mercury. Then, you put a speaker at one end and microphone at the other. When you pulse the speaker, it creates a pressure wave.

This takes time to propagate to the other end of the tube, where it hits the microphone, converting it back into an electrical signal. And we can use this propagation delay to store data! Imagine that the presence of a pressure wave is a 1 and the absence of a pressure wave is a 0.

Our speaker can output a binary sequence like 1010 0111. The corresponding waves will travel down the tube, in order, and a little while later, hit the microphone, which converts the signal back into 1's and 0's. If we create a circuit that connects the microphone to the speaker, plus a little amplifier to compensate for any loss, we can create a loop that stores data.

The signal traveling along the wire is near instantaneous, so there's only ever one bit of data showing at any moment in time. But in the tube, you can store many bits! After working on ENIAC, Eckert and his colleague John Mauchly, set out to build a bigger and better computer called EDVAC, incorporating Delay Line Memory.

In total, the computer had 128 Delay Lines, each capable of storing 352 bits. That's a grand total of 45 thousands bits of memory, not too shabby for 1949! This allowed EDVAC to be one of the very earliest Stored-Program Computers, which we talked about in Episode 10.

However, a big drawback with delay line memory is that you could only read one bit of data from a tube at any given instant. If you wanted to access a specific bit, like bit 112, you'd have to wait for it to come around in the loop, what's called sequential or cyclic-access memory, whereas we really want random access memory, where we can access any bit at any time. It also proved challenging to increase the density of the memory, packing waves closer together meant they were more easily mixed up.

In response, new forms of delay line memory were invented, such as magnetostrictive delay lines. These delay lines use a metal wire that could be twisted, creating little torsional waves that represented data. By forming the wire into a coil, you could store around 1000 bits in a 1 foot by 1 foot square.

However, delay line memory was largely obsolete by the mid 1950s, surpassed in performance, reliability and cost by a new kid on the block: magnetic core memory which was constructed out of little magnetic donuts, called cores. If you loop a wire around this core.... and run an electrical current through the wire, we can magnetize the core in a certain direction. If we turn the current off, the core will stay magnetized.

If we pass current through the wire in the opposite direction, the magnetization direction, called polarity, flips the other way. In this way, we can store 1's and 0's! 1 bit of memory isn't very useful, so these little donuts were arranged into grids. There were wires for selecting the right row and column, and a wire that ran through every core, which could be used to read or write a bit.

Here is an actual piece of core memory! In each of these little yellow squares, there are 32 rows and 32 columns of tiny cores, each one holding 1 bit of data. So, each of these yellow squares could hold 1024 bits.

In total, there are 9 of these, so this memory board could hold a maximum of 9216 bits, which is around 9 kilobytes. The first big use of core memory was MIT's Whirlwind 1 computer, in 1953, which used a 32 by 32 core arrangement. And, instead of just a single plane of cores, like this, it was 16 boards deep, providing roughly 16 thousand bits of storage.

Importantly, unlike delay line memory, any bit could be accessed at any time. This was a killer feature, and magnetic core memory became the predominant Random Access Memory technology for two decades, beginning in the mid 1950s even though it was typically woven by hand! Although starting at roughly 1 dollar per bit, the cost fell to around 1 cent per bit by the 1970s.

Unfortunately, even 1 cent per bit isn't cheap enough for storage. As previously mentioned, an average smartphone photo is around 5 megabytes in size, that's roughly 40 million bits. Would you pay 4 hundred thousand dollars to store a photo on core memory?

If you have that kind of money to drop, did you know that Crash Course is on Patreon? Right? Wink wink.

Anyway, there was tremendous research into storage technologies happening at this time. By 1951, Eckert and Mauchly had started their own company, and designed a new computer called UNIVAC, one of the earliest commercially sold computers. It debuted with a new form of computer storage: magnetic tape.



Memory & Storage: Crash Course Computer Science #19

Crash Course: Computer Science

<https://youtube.com/watch?v=TQCr9RV7twk>

<https://nerdfighteria.info/v/TQCr9RV7twk>

This was a long, thin and flexible strip of magnetic material, stored in reels. The tape could be moved forwards or backwards inside of a machine called a tape drive. Inside is a write head, which passes current through a wound wire to generate a magnetic field, causing a small section of the tape to become magnetized.

The direction of the current sets the polarity, again, perfect for storing 1's and 0's. There was also a separate read head could detect the polarity non-destructively. The UNIVAC used half-inch-wide tape with 8 parallel data tracks, each able to store 128 bits of data per inch.

With each reel containing 1200 feet of tape, it meant you could store roughly 15 million bits – that's almost 2 megabytes! Although tape drives were expensive, the magnetic tape itself was cheap and compact, and for this reason, they're still used today for archiving data. The main drawback is access speed.

Tape is inherently sequential, you have to rewind or fast-forward to get to data you want. This might mean traversing hundreds of feet of tape to retrieve a single byte, which is slow. A related popular technology in the 1950s and 60s was Magnetic Drum Memory.

This was a metal cylinder – called a drum – coated in a magnetic material for recording data. The drum was rotated continuously, and positioned along its length were dozens of read and write heads. These would wait for the right spot to rotate underneath them to read or write a bit of data.

To keep this delay as short as possible, drums were rotated thousand of revolutions per minute! By 1953, when the technology started to take off, you could buy units able to record 80,000 bits of data – that's 10 kilobytes, but the manufacture of drums ceased in the 1970s. However, Magnetic Drums did directly lead to the development of Hard Disk Drives, which are very similar, but use a different geometric configuration.

Instead of large cylinder, hard disks use, well... disks... that are hard. Hence the name! The storage principle is the same, the surface of a disk is magnetic, allowing write and read heads to store and retrieve 1's and 0's.

The great thing about disks is that they are thin, so you can stack many of them together, providing a lot of surface area for data storage. That's exactly what IBM did for the world's first computer with a disk drive: the RAMAC 305. Sweet name BTW.

It contained fifty, 24-inch diameter disks, offering a total storage capacity of roughly 5 megabytes. Yess!! We've finally gotten to a technology that can store a single smartphone photo!

The year was 1956. To access any bit of data, a read/write head would travel up or down the stack to the right disk, and then slide in between them. Like drum memory, the disks are spinning, so the head has to wait for the right section to come around.

The RAMAC 305 could access any block of data, on average, in around 6/10ths of a second, what's called the seek time. While great for storage, this was not nearly fast enough for memory, so the RAMAC 305 also had drum memory and magnetic core memory. This is an example of a memory hierarchy, where you have a little bit of fast memory, which is expensive, slightly more medium-speed memory, which is less expensive, and then a lot of slowish memory, which is cheap.

This mixed approach strikes a balance between cost and speed. Hard disk drives rapidly improved and became commonplace by the 1970s. A hard disk like this can easily hold 1 terabyte of data today – that's a trillion bytes – or roughly 200,000 five megabyte photos!

And these types of drives can be bought online for as little as 40 US dollars. That's 0.000000005 cents per bit. A huge improvement over core memory's 1 cent per bit!

Also, modern drives have an average seek time of under 1/100th of a second. I should also briefly mention a close cousin of hard disks, the floppy disk, which is basically the same thing, but uses a magnetic medium that's, floppy. You might recognize it as the save icon on some of your applications, but it was once a real physical object!

It was most commonly used for portable storage, and became near ubiquitous from the mid 1970s up to the mid 90s. And today it makes a pretty good coaster. Higher density floppy disks, like Zip Disks, became popular in the mid 1990s, but fell out of favor within a decade.

Optical storage came onto the scene in 1972, in the form of a 12-inch "laser disc." However, you are probably more familiar with its later, smaller, are more popular cousin, the Compact Disk, or CD, as well as the DVD which took off in the 90s. Functionally, these technologies are pretty similar to hard disks and floppy disks, but instead of storing data magnetically, optical disks have little physical divots in their surface that cause light to be reflected differently, which is captured by an optical sensor, and decoded into 1's and 0's. However, today, things are moving to solid state technologies, with no moving parts, like this hard drive and also this USB stick.

Inside are Integrated Circuits, which we talked about in Episode 15. The first RAM integrated circuits became available in 1972 at 1 cent per bit, quickly making magnetic core memory obsolete. Today, costs have fallen so far, that hard disk drives are being replaced with non-volatile, Solid State Drives, or SSDs, as the cool kids say.

Because they contain no moving parts, they don't really have to seek anywhere, so SSD access times are typically under 1/1000th of a second. That's fast! But it's still many times slower than your computer's RAM.

For this reason, computers today still use memory hierarchies. So, we've come along way since the 1940s. Much like transistor count and Moore's law, which we talked about in Episode 14, memory and storage technologies have followed a similar exponential trend.

From early core memory costing millions of dollars per megabyte, we're steadily fallen, to mere cents by 2000, and only fractions of a cent today. Plus, there's WAY less punch cards to keep track of. Seriously, can you imagine if there was a slight breeze in that room containing the SAGE program? 62,500 punch cards.

I don't even want to think about it. I'll see you next week.