Jahong Liu

System Design Basics 1: Computer Architecture

Computer Architecture

The building blocks of a computer and their usages.

Components

Disk

A disk is the primary storage device in a computer. It is **persistent**, meaning the data will be persistent regardless of the state of the machine (turned on or off). Most modern computers store information in disk on the order of TBs (terabytes).

Recall that a byte is made up of 8-bits, and bit is the smallest unit of measure in computers – a binary measure represented by 0 or 1. A teratype is 10^12 bytes, or a trillion bytes. A disk storage device like a USB drive might have storage on the order of **GB** (gigabytes), which is 10^9, or one billion bytes.

Two terms: **HDD** and **SDD**. Hard disk drives (HDD) and Solid-state drive (SDD) are both persistent storage devices, with the latter one being more popular and faster. However, it does generally cost a bit more.

HDDs are mechanical, and have read/write head. The older they get, the more wear and tear they collect which slows them down overtime. SSDs are significantly faster because they don't have moving parts, and reply on reading and writing data electronically (similar to RAM)

RAM

Random Access Memory is also used for storing information but is typically a lot smaller in size compared to disk drives. RAM sizes generally vary from 1GB - hundreds of GB because RAM is *much* more expensive than disk space. The benefit is that reading and writing to

RAM is *significantly* faster than disk. For instance, writing 1 MB of data to RAM might take on the order of microseconds (millionths of a second, or $1/10^6$ seconds), while writing the same amount to a disk might take on the order of milliseconds (thousandths of a second, or $1/10^3$ seconds)." Note: these numbers are rough estimates and can change over time.

RAM keeps the applications you have open in memory; include any variables your program has allocated. It is often described as volatile memory, meaning that the data gets erased once the computer is shut down. This is why it is important to save your work to disk before shutting down your computer.

It is important to note that the RAM and disk do not directly communicate with each other. They rely on the CPU to facilitate data transfer between them.

CPU

The central processing unit (CPU) is the intermediary between the RAM and disk. Also referred to as the "brain" of the computer; it reads/writes from the RAM and disk(s).

For example, when you write code and run it, your code is translated into a set of binary instructions stored in RAM. This sentence could be made clearer: "The CPU reads and executes these instructions, which may involve manipulating data stored elsewhere in RAM or on disk. An example of reading from disk, would be opening a file in your file system and reading it line-by-line.

All computations are done within the CPU, such as addition/subtraction/multiplication etc. This occurs in a matter of milliseconds. It fetches instructions from the RAM, decodes those instructions and executes the decoded instructions. On the lowest level, all these instructions are represented as bytes.

The CPU also consists of a cache. A cache is extremely fast memory that lies on the same die as the CPU.

Caches

Most CPUs have an L1, L2, and L3 cache, which are physical components that are much faster than RAM, but they only stores data on the order of KBs or tens of MBs

Whenever a read operation is requested, the cache is checked before the RAM and the disk. If the data requested is in the cache, and is unchanged since the last time it was accessed, it will be fetched from the cache, and not the RAM. Reading and writing to the cache is a lot faster than RAM and disk. It is up to the operating system to decide what gets stored in the cache.

Caching is an important concept applied in *many* areas beyond computer architecture. For example, web browsers use cache to keep track of frequently accessed web pages to load them faster. This stored data might include HTML, CSS, JavaScript, and images, among other things. If the data is still valid from the time the page was cached, it will load faster. But in this case, the browser is using the disk as cache, because making internet requests is a lot slower than reading from disk.

The cache being part of the CPU is only part of the reason why it is faster than RAM. Cache is what's known as SRAM. If you are interested, you might view this resource from **MIT,** which gives a gentle introduction.

Closing Notes

All three memory types have their own use. If you are writing an essay, and the computer were to shut down, you would want it stored on a disk drive. While multi-tasking, you would want the programs opened to be stored in the RAM.