INTERNATIONAL BURCH UNIVERSITY

Faculty of Engineering and Natural Sciences

Department of Information Technology

CEN 261: Computer Organization

Research Project

INTERNAL MEMORY

Sarajevo, 7th of January, 2022

Amira Abdo- 20002450

TABLE OF CONTENTS

1. Abstract

The memory, being a crucial part of every computer system, as well as of numerous electrical items in use every day, represents a technological concept which is bound to advance and improve majorly in the years to come, just as it has improved ever since the beginning of the computer-era. As such, it has gone through multiple stages, taking several forms which implement different mechanisms and use various materials and media for their task of data storing. All the evolution in computer memory, as well as computers themselves, brought us to the modern computer architectures which are based on implementing several different types of memory to optimize a device's performance, cost and storage size. This research report intends to present the reader with some basic information about computer memory: how it works, how exactly it impacts a computer's performance, what types of memory are used in modern computers, as well as how the information-seeking happens in one such system.

2. Introduction

Computer memory is a term used to describe the storage space of a computer, used for processing data as well as storing the instructions for data processing. It can be classified as either external or internal memory. The internal memory, which is the main focus of this research project, is the memory that is accessed directly by the processor without using the computer's input-output channels. It is also referred to as the main or primary memory of a computer, and it stores smaller amounts of data that is accessed quickly during runtime. The memory space is composed of a very large number of individual memory-storing parts which we call memory cells. Each of these cells is a unique memory location, and, as such, possess a unique address, which ranges from zero to the computer's memory size minus one. There are three main types of internal memory: RAM, ROM and cache memory, as well as registers inside of the processor, each of which will be described further in the paper.

3. Main memory through the years

The earliest type of data storage in computer were ultrasonic waves stored in tubes of mercury, referred to as mercury delay lines, as well as cathode-ray tubes which used electrical charge on the

tubes to store data. Around 1948 the magnetic drum was invented, using a coating of iron oxide on a rotating drum to store information as magnetic patterns.

Any device that can be in one of two possible states, corresponding to the bit values 0 and 1 can serve as computer memory. According to this, magnetic-core memory, the first relatively affordable Random Access Memory device, was invented in 1952. “It was composed of tiny, doughnut-shaped ferrite magnets threaded on the intersection points of a two-dimensional wire grid. These wires carried currents to change the direction of each core’s magnetization, while a third wire threaded through the doughnut detected its magnetic orientation.” (https://www.britannica.com/technology/computer/Main-memory)

Since accessing a memory address in a chip requires the address to first be specified, and since memory is slower than a CPU, a memory design in which a series of words can be transferred sequentially after the first address has been determined represents a notable advance in technology. One such design saw a wide-range usage around 2001, and it was known as synchronous DRAM (SDRAM).

However, transfer of data through the data bus (the connection between the CPU and memory through which data is fetched) acted as a bottleneck. To solve this issue, modern processor chips contain cache memory- a very small memory space made up of SRAM which holds copies of certain data from main memory.

Nonetheless, data transfer through the “bus”—the set of wires that connect the CPU to memory and peripheral devices—is a bottleneck. For that reason, CPU chips now contain cache memory—a small amount of fast SRAM. The cache holds copies of data from blocks of main memory. A well-designed cache allows up to 85–90 percent of memory references to be done from it in typical programs, giving a several-fold speedup in data access.

“The time between two memory reads or writes (cycle time) was about 17 microseconds (millionths of a second) for early core memory and about 1 microsecond for core in the early 1970s. The first DRAM had a cycle time of about half a microsecond, or 500 nanoseconds (billionths of a second), and today it is 20 nanoseconds or less. An equally important measure is the cost per bit of memory. The first DRAM stored 128 bytes (1 byte = 8 bits) and cost about $10,

or $80,000 per megabyte (millions of bytes). In 2001 DRAM could be purchased for less than $0.25 per megabyte."[1]

All of this led us to the modern organization of computer memory inside of a machine, which will be discussed in the following sections.

4. How memory works

Data in a computer system is processed and stored according to presence or lack of electronic or magnetic signals in its circuit or the media used. This is a characteristic referred to as two-state or binary representation of data, as the computer and the media can be in only one of two possible conditions, "on" or "off"- as a regular light switch. For example, semiconductor circuitries are either in a conducting or a nonconducting state. Media that uses magnetic waves as a guard exhibit these two states using magnetized spots whose magnetic fields have one of two possible polarities.

This binary characteristic is exactly the reason for the binary number system being the basis of representing data in computers. Therefore, in electronic circuitries used today, the conducting state represents the number 1 and the nonconducting state represents 0.

A single 1 or 0 'digit', which is, in fact, the conducting or the opposite state, represents one binary digit, or bit. Bits are commonly used to express a capacity of a memory chip. A more general way of storing data is using a byte- a combination of typically 8 bits. A byte represents one character of data in most computer architectures. It is usually used for expressing a computer's memory capacity.

4.1. Memory cell in internal memory systems

A computer's internal memory is made up of a semiconductor material, commonly silicon. This material is more expensive compared to, for example, magnetic tapes used for external memory, so this memory is usually small in size.

1- Source: https://www.britannica.com/technology/computer/Main-memory

Although there are different methods for building the semiconductor internal memory, there are some universal characteristics shared by all semiconductor cells:

- "Every memory cell exhibits two states which represent binary 0 and 1.
- Every memory cell can be read to sense the state it is representing.
- Every memory cell can be written to set it to a particular state i.e., either 0 or 1."[2]

As seen in Figure 1, every memory cell is connected to three lines: select, control and read/write. The select line is used to indicate if the particular memory cell is currently in use for the read/write operation. The control line determines if the operation is a read or a write. For a write operation, an electric signal passes through the read/write line which sets the cell to either 0 or 1. In case of a read operation, the same read/write line is used to output the cell's state.

Figure 1 shows these lines the memory cell is connected to.

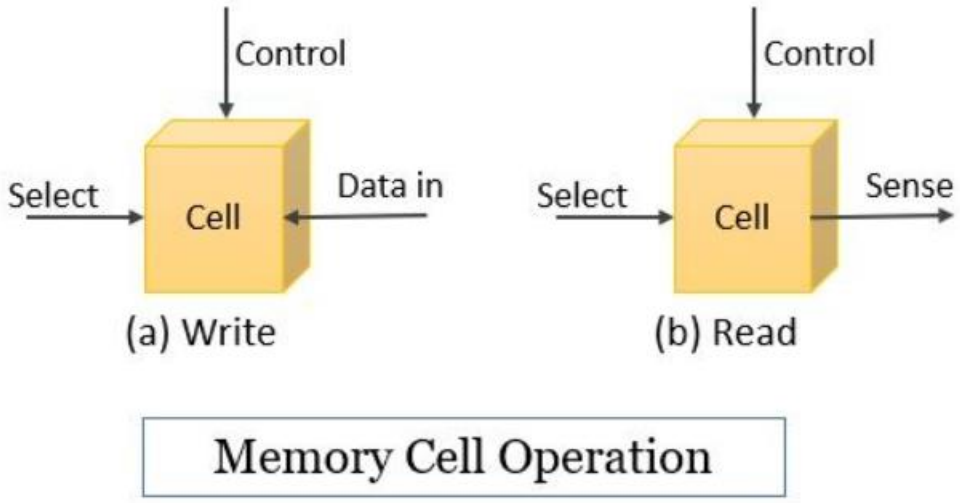

Figure 1: Memory Cell Operation

Source: https://binaryterms.com/internal-memory-in-computer-architecture.html

2- Source: https://binaryterms.com/internal-memory-in-computer-architecture.html

Since the memory cell is the fundamental building block of computer memory, a computer's memory is made up of billions of these cells. The way in which they are interconnected and organized depends on the memory type, which will be discussed in the following sections.

5. Types of internal memory

There are three main types of internal memory: Random Access Memory (RAM), Read Only Memory (ROM) and cache memory. These types are further divided into sub-types, as seen in Figure 2.

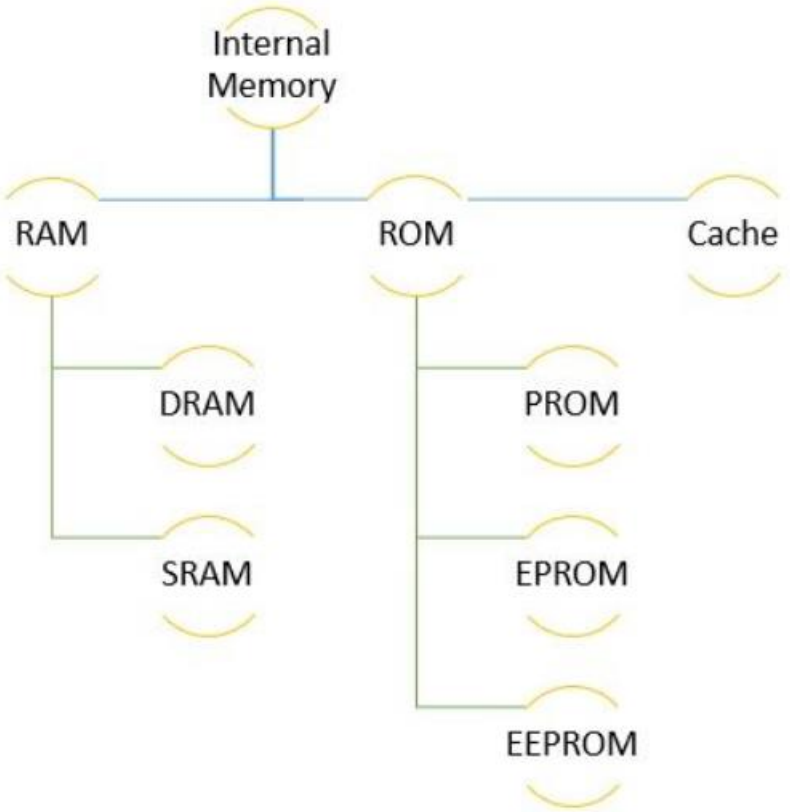

Figure 2: Types of internal memory

Source: https://binaryterms.com/internal-memory-in-computer-architecture.html

5.1. Random Access Memory

Random access memory, or RAM, is a type of computer memory which can be read and written into in any order. It is typically used for storing working data and machine code. It is usually a volatile type of memory, which means it holds data while the computer is running; this is done so the processor can access the needed data quickly. It is worth mentioning non-volatile RAM has also been developed.

"A random-access memory device allows data items to be read or written in almost the same amount of time irrespective of the physical location of data inside the memory, in contrast with other direct-access data storage media (such as hard disks, CD-RWs, DVD-RWs and the older magnetic tapes and drum memory), where the time required to read and write data items varies significantly depending on their physical locations on the recording medium, due to mechanical limitations such as media rotation speeds and arm movement."[3]

Today's modern technology allows RAM to be implemented through integrated circuit chips with metal-oxide-semiconductor cells.

5.1.1. Dynamic RAM

DRAM is widely used for building a computer's main memory. As seen in Figure 3, its memory cell is made using and transistor and a capacitor which make up an integrated circuit, with the bit of data being stored in the capacitor. The capacitor being charged represents the binary digit 1, and when the capacitor is discharged the state of the memory cell is 0. However, the transistor always leaks a small amount, resulting in the capacitor slowly discharging. This results in the DRAM needing to be refreshed (resupplied with charge) every few milliseconds.

Dynamic RAM is used for applications which require a large RAM capacity, such as in personal computers.

3- Source: https://en.wikipedia.org/wiki/Random-access_memory#Addressing

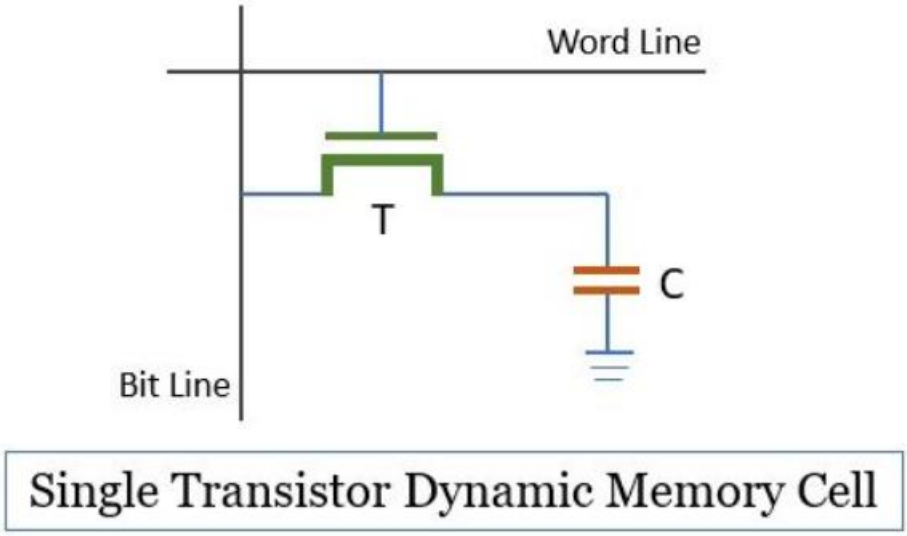

Figure 3: DRAM cell

Source: https://binaryterms.com/internal-memory-in-computer-architecture.html

5.1.2 Static RAM

An SRAM cell stores a bit of information using a circuit made of two inverters which are cross-connected, forming a latch, as seen in Figure 4. This latch is connected to two-bit lines which are connected to two transistors. The transistors play the role of a switch that opens and closes as ordered by the control line. They are switched on when that particular memory cell is selected as a read/write operation is being performed on it.

Data in an SRAM cell is retained as long as the system is supplied with power, unlike DRAM which has to be periodically refreshed. Because of this SRAM is faster, but it is also more costly, making DRAM a more popular choice in typical computer systems. SRAM is typically used for applications that don’t require a too large memory capacity.

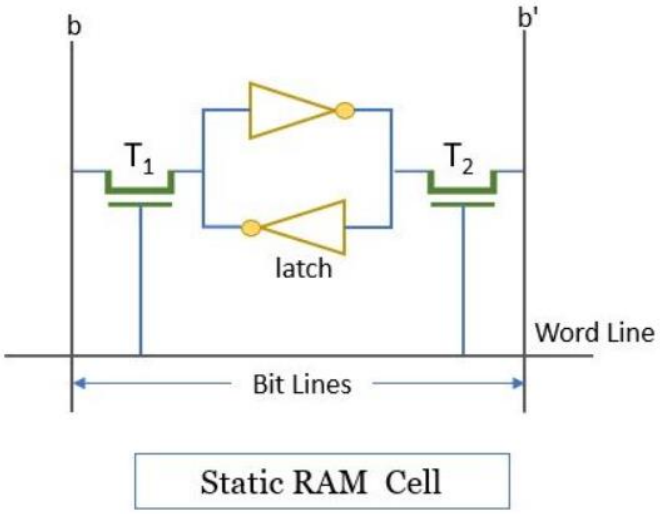

Figure 4: SRAM cell

Source: https://binaryterms.com/internal-memory-in-computer-architecture.html

5.1.3. DRAM vs SRAM

The following table, Figure 5 shows comparison of the two types of RAM.

Feature	DRAM	SRAM
Cost	Cheaper	More expensive
Performance	Slower: Off-chip memory with longer access time	On-chip memory with minimal access time; can run at the speed of the host microprocessor
Use case	Main memory	Level 1 and Level 2 microprocessor caches
Density	Less density per cell (1 transistor per chip) but can pack more cells into space	Denser (6 transistors per chip) but can fit fewer cells into space
Power	Generally higher: Capacitors leak power thanks to imperfect insulation, requiring regular power refreshes.	Generally lower: No charge leakage since it changes direction of current through switches instead of leaking power through the capacitor. However, this depends on the application environment and SRAM can consume as much or more power as DRAM.
Storage capacity	Larger: Connects directly to CPU bus, volatile storage measured in GBs	Smaller: Acts as cache; storage measured in MBs
Volatility	Volatile: Must have active power supply plus frequent charges while active.	Volatile: Does not require additional charges while it is receiving power, but eventually loses data without it.
Physical placement	Motherboard	Processors or between processor and main memory

Figure 5: DRAM and SRAM comparison

Source: https://www.enterprisestorageforum.com/hardware/sram-vs-dram/

5.2 Read-Only Memory

Read-Only Memory, or ROM, is a non-volatile type of memory, meaning that data written into it is retained even when the system's power supply is cut off.

The structure of a ROM cell is shown below in Figure 6. The bit value of the memory cell is 0 if the transistor is at ground level, and 1 otherwise.

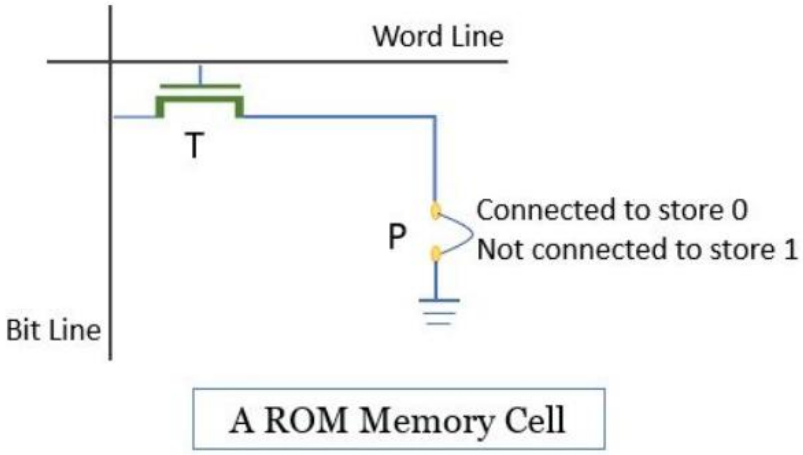

Figure 6: ROM cell

Source: https://binaryterms.com/internal-memory-in-computer-architecture.html

The bit line and the power supply are connected through the resistor. When reading the value of the memory cell, the word line is activated and the transistor is then connected to the ground. This makes the voltage of the bit line drop to 0 in case the transistor is connected to the ground. Otherwise, the bit line retains its voltage and indicates the value of the memory cell is 1. The most important characteristic of ROM is the fact that the state of the memory cell is set at the time of manufacturing, so only a reading mechanism is needed (and implemented, as seen)- which can be concluded from the memory type's name alone.

It is useful for storing software which is rarely changed during the system's lifecycle (this type of software is called firmware). It can also be used "for microprogramming, such as for storing library subroutines, system programs, function tables."[4]

4- Source: https://binaryterms.com/internal-memory-in-computer-architecture.html

ROM also has subtypes, some of which being: programmable ROM, erasable programmable ROM, electrically erasable programmable ROM and flash memory.

5.2.2 PROM

"The programable ROM (PROM) is used when few ROMs are required with a specific memory content. PROMs can be written only once using electric signals."[5]

5.2.3 EPROM

Erasable Programmable ROM, or EPROM, is a type of ROM which can be read and written via electric signals. Before the write operation is done, ultraviolet rays are used to set the memory chip back to its original state, thus deleting the current content of the chip. This writing and erasing of the chip can be done repeatedly, and, being a type of ROM, the data is retained even when there is no power supply. It is fairly more costly than regular ROM and PROM.

5.2.4 EEPROM

Electrically Erasable Programmable ROM, or EEPROM, is a type of ROM that can be selectively deleted and rewritten several times. Unlike EPROM, which has to be wiped whole using ultraviolet rays, parts of data stored on a EEPROM device can be wiped by applying a higher voltage than normal. This makes EEPROM more complex than EPROM, consequently making it more expensive.

5.2.5 Flash memory

One could say that the flash memory is somewhere in between EPROM and EEPROM, functionality- and cost-wise. In flash memory, blocks of cells are written, which are erased before the write operation, unlike in EEPROM where erasure is done on a byte level. Erasing flash memory is also faster.

5- Source: https://binaryterms.com/internal-memory-in-computer-architecture.html

5.3 Cache memory

Cache memory is a special type of memory which is very fast, very small and expensive. It is a volatile type of memory- it can't retain data without constant power supply. It is used to store copies of data from the main memory- it acts as a buffer between RAM and the processor. It is used to store frequently used data and instructions so that it is easily accessible by the CPU as needed. This majorly enhances the performance of the whole system. Figure 7 shows the (simplified) process flow between different parts of a computer's memory and CPU.

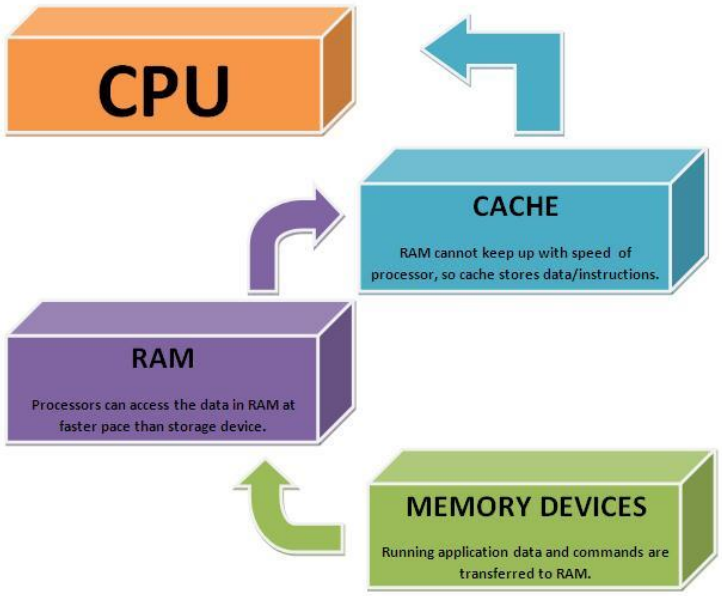

Figure 7: Process flow between memories

Source: https://www.engineersgarage.com/how-cache-memory-works/

It is more expensive than main memory or disk memory, but it is more affordable than CPU registers. Based on these trade-offs, computers are designed by implementing several types of memory, balancing out the system's performance, size and cost. This solution is referred to as the memory hierarchy and will be discussed later on.

5.3.1. Levels of cache

It is worth mentioning a system can have more than one level of cache memory. These levels simply mean that the system's overall cache memory is split onto different hardware segments with different processing speed and sizes.

For example, we will assume a system has three levels of cache memory: L1, L2 and L3. L3 cache is faster and smaller than RAM but larger and slower than L2 cache, which is placed in close vicinity of the processor (or built-in inside the processor in some modern devices). At the core level is the L1 cache which is smallest, fastest and the most expensive of the cache levels.

The process flow in one such system is described as followings:

Whenever a processor requires some data, it first checks the registers inside the CPU. If the data is not present there, it looks at the first level of cache memory- L1, and if the data is also not present there, it goes to the 2nd and further 3rd level of cache memory. When the required data is not located in the cache, this is referred to as a cache miss, and when the data is found in the cache- it is referred to as a cache hit. Furthermore, if the data isn't found in any level of the cache, other layers of the computer's memory are checked, first the RAM and then other storage devices. This process is depicted in Figure 8 below.

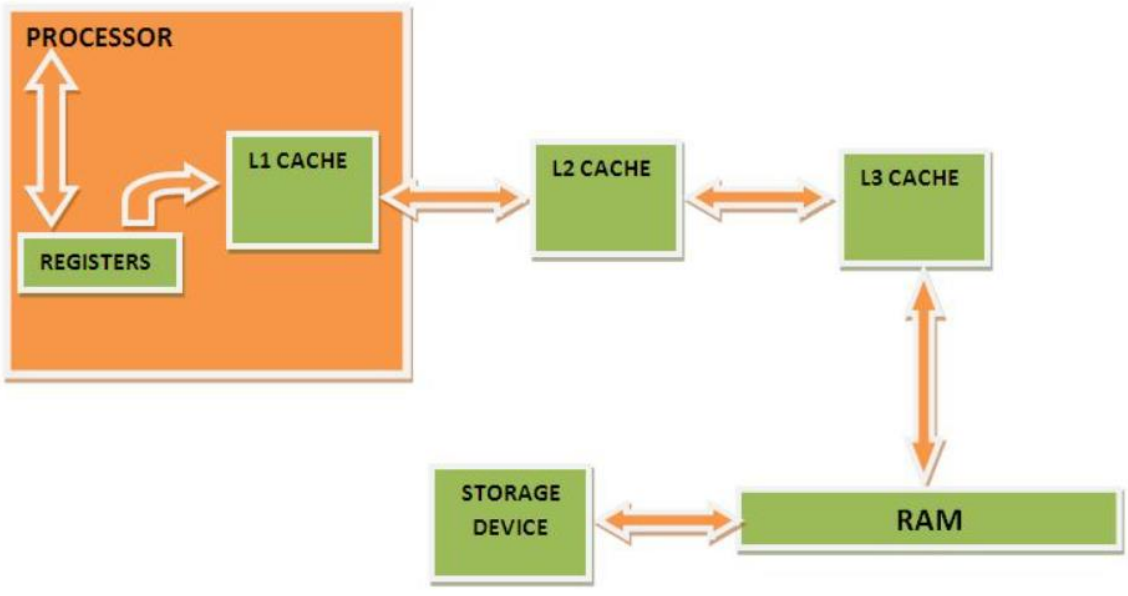

Figure 8: Data searching through different memory parts

Source: https://www.engineersgarage.com/how-cache-memory-works/

5.3.2. Cache mapping

Currently in use are three different types of cache mapping, all of them described below:

- Direct mapping- the simplest mapping technique. “In Direct mapping, assign each memory block to a specific line in the cache. If a line is previously taken up by a memory block when a new block needs to be loaded, the old block is trashed. An address space is split into two parts index field and a tag field. The cache is used to store the tag field whereas the rest is stored in the main memory.”

- Associative mapping- considered to be the fastest and most flexible mapping technique. “In this type of mapping, the associative memory is used to store content and addresses of the memory word. Any block can go into any line of the cache. This means that the word id bits are used to identify which word in the block is needed, but the tag becomes all of the remaining bits. This enables the placement of any word at any place in the cache memory. It is considered to be the fastest and the most flexible mapping form.”

- Set-associative mapping- an enhanced form of direct mapping which addresses the problem of possible thrashing[8]. “It does this by saying that instead of having exactly one line that a block can map to in the cache, we will group a few lines together creating a set. Then a block in memory can map to any one of the lines of a specific set..Set-associative mapping allows that each word that is present in the cache can have two or more words in the main memory for the same index address. Set associative cache mapping combines the best of direct and associative cache mapping techniques. In this case, the cache consists of a number of sets, each of which consists of a number of lines.”

6- Source: https://www.geeksforgeeks.org/cache-memory-in-computer-organization/

7- Source: https://www.geeksforgeeks.org/cache-memory-in-computer-organization/

8- Thrashing- In computer science, thrashing occurs when a computer's virtual memory resources are overused, leading to a constant state of paging and page faults, inhibiting most application-level processing

9- Source: https://www.geeksforgeeks.org/cache-memory-in-computer-organization/

6. Memory hierarchy

Now that we are familiar with different types of memory in a computer system, we will see how they are combined to be as efficient as possible. This efficiency enhancement is achieved through the implementation of what we call a memory hierarchy. It is a way to organize the memory so that access time is minimized. It was developed based on a behavior called locality of references[10].

Figure 9 shows different levels of a computer hierarchy.

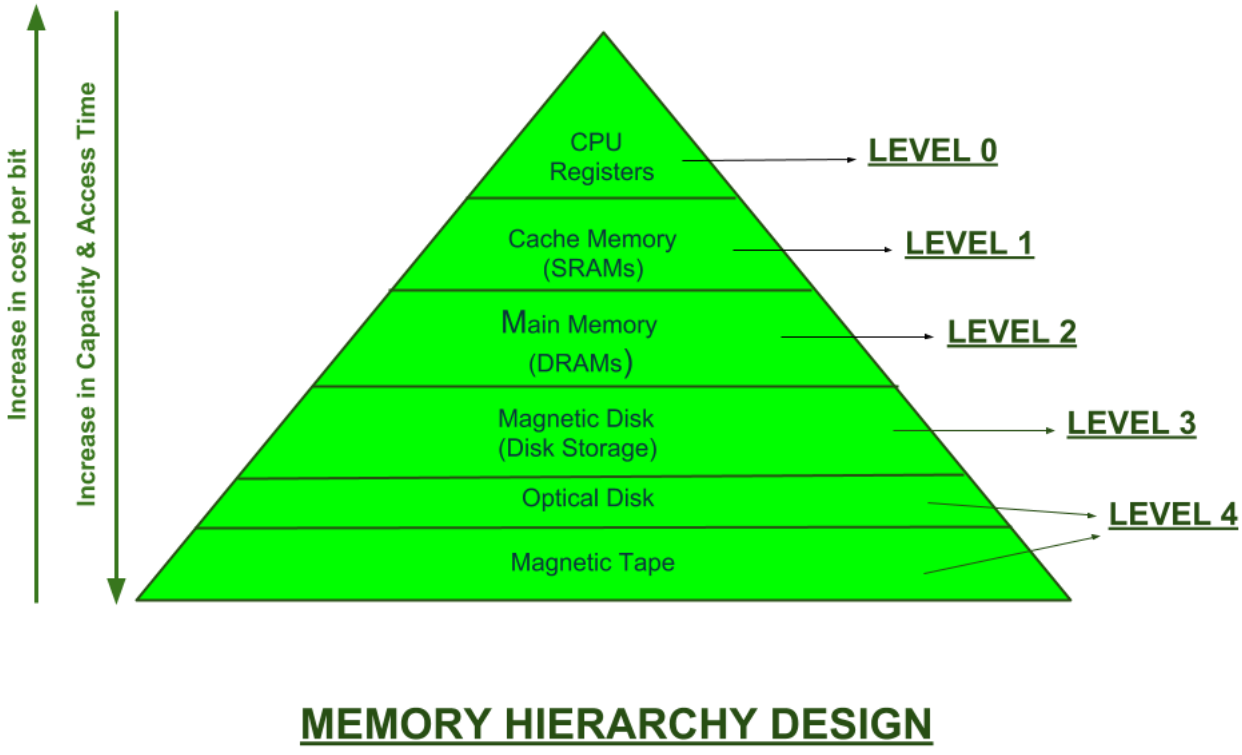

Figure 9: Memory hierarchy design

Source: https://www.geeksforgeeks.org/memory-hierarchy-design-and-its-characteristics/

As seen in the figure above, as we move downwards from the top to the bottom in the hierarchy, there is an increase in the memory type's capacity and access time, and a decrease in cost per bit.

10- Locality of reference- in computer science, locality of reference, also known as the principle of locality, is the tendency of a processor to access the same set of memory locations repetitively over a short period of time

"One of the most significant ways to increase system performance is minimizing how far down the memory hierarchy one has to go to manipulate data"[11]

7. In-memory computing

In-memory computing, synonymous with in-memory processing/computation, is a newly emerged technology in which data in processed entirely in the computer's memory. Older systems use disks for data storage and relational databases (which use SQL query language), which have shows to be inadequate for various business intelligence needs. As a result, in-memory computing has been invented, as stored data is accessed much faster when it is located in RAM or flash memory. In-memory computation represents a breakthrough in data-processing because it eliminates all the slow data accesses by relying exclusively on data stored in RAM.

"Besides reducing latency and energy cost associated with data movement, in-memory computing also has the potential to improve the computational time complexity associated with certain tasks due to the massive parallelism afforded by a dense array of millions of nanoscale memory devices serving as compute units. By introducing physical coupling between the memory devices, there is also a potential for further reduction in computational time complexity. Memristive devices, such as PCM, ReRAM, and MRAM are particularly well suited for in-memory computing."[12]

As stated in the previous quote, memristors[13] represent a group of devices which have high potential for realizing and further advancing the in-memory computation technology. This is just one of the technologies which remain yet to be explored further and exploited to their full potential.

11- Source: https://www.geeksforgeeks.org/memory-hierarchy-design-and-its-characteristics/

12- Source: Mehonic, A., Sebastian, A., Rajendran, B., Simeone, O., Vasilaki, E. and Kenyon, A.J. (2020). Memristors—From In-Memory Computing, Deep Learning Acceleration, and Spiking Neural Networks to the Future of Neuromorphic and Bio-Inspired Computing. Advanced Intelligent Systems)

13- Memristor- the memristor is a new building block for electrical circuits whose resistance can be programmed (resistor function) and subsequently remains stored (memory function). The advantage of memristors is their fast processing and energy efficiency in the computational process.

As the technology keeps evolving every day, more unexplored concepts and device designs are bound to emerge in data storing as well as other fields of computing, making computer memory cheaper, larger and faster, resulting in computer performance being greatly improved overall.

8. Conclusion

It is evident that a modern computer system cannot be designed successfully without the implementation of several different types of internal memory, including both static and dynamic RAM, different types of ROM, as well as cache memory which is arguably the most impactful memory type in a computer. When talking about a computer's RAM, as previously mentioned, both SRAM and DRAM are used, with DRAM being a more affordable option due to it requiring periodical refreshing in order to retain data. This is a consequence of the DRAM memory cell's design, which uses a transistor and a capacitor, with the transistor constantly leaking, making the capacitor's charge drop every few milliseconds. On the other hand, SRAM, because of its higher price, is used in applications which don't require a too large memory space, but where speed is of importance. Another major type of internal memory is ROM, a type of memory which is written at manufacture and mostly cannot be changed (although there exist technologies such as PROM, EPROM, EEPROM and flash memory which can be edited). As such, ROM is used to store software which rarely needs to be changed during the life of the computer- firmware.

Yet another crucial type of internal memory is the cache memory, which is absolutely vital to a computer system. It is much faster than other types of memory and one could say that eliminating a system's cache would be a trip back in time to the days when computers took hours to complete a task. Based on which memory-mapping technique the cache uses to store copies of data from the main memory, modern cache memory architectures allow the percentage of cache hits to be very high, enhancing the whole system's overall performance by majorly reducing the time it takes for the processor to find the information it requires.

As technology keeps evolving, it is quite possible and probable that we will be introduced to new memory types, which will likely be more affordable per bit, allowing the average computer to be faster and overall better, as well as some special-purpose computers to be 'stronger', paving the way for humanity to make new discoveries in various scientific fields.

References:

1. Mehonic, A., Sebastian, A., Rajendran, B., Simeone, O., Vasilaki, E. and Kenyon, A.J. (2020). Memristors—From In-Memory Computing, Deep Learning Acceleration, and Spiking Neural Networks to the Future of Neuromorphic and Bio-Inspired Computing. Advanced Intelligent Systems
2. https://www.tutorialspoint.com/computer_fundamentals/computer_memory.htm (accessed 3rd of January, 2022)
3. https://binaryterms.com/internal-memory-in-computer-architecture.html (accessed 7th of January, 2022)
4. https://www.britannica.com/technology/computer/Main-memory (accessed 7th of January, 2022)
5. https://www.quora.com/How-does-Computer-Memory-work (accessed 4th of January, 2022)
6. https://www.enterprisestorageforum.com/hardware/sram-vs-dram/ (accessed 3rd of January, 2022)
7. https://bohatala.com/types-of-internal-memory/ (accessed 4th of January, 2022)
8. https://www.atpinc.com/blog/computer-memory-types-dram-ram-module (accessed 4th of January, 2022)
9. https://www.engineersgarage.com/how-cache-memory-works/ (accessed 5th of January, 2022)
10. https://www.geeksforgeeks.org/cache-memory-in-computer-organization/ (accessed 5th of January, 2022)
11. https://www.geeksforgeeks.org/memory-hierarchy-design-and-its-characteristics/ (accessed 5th of January, 2022)