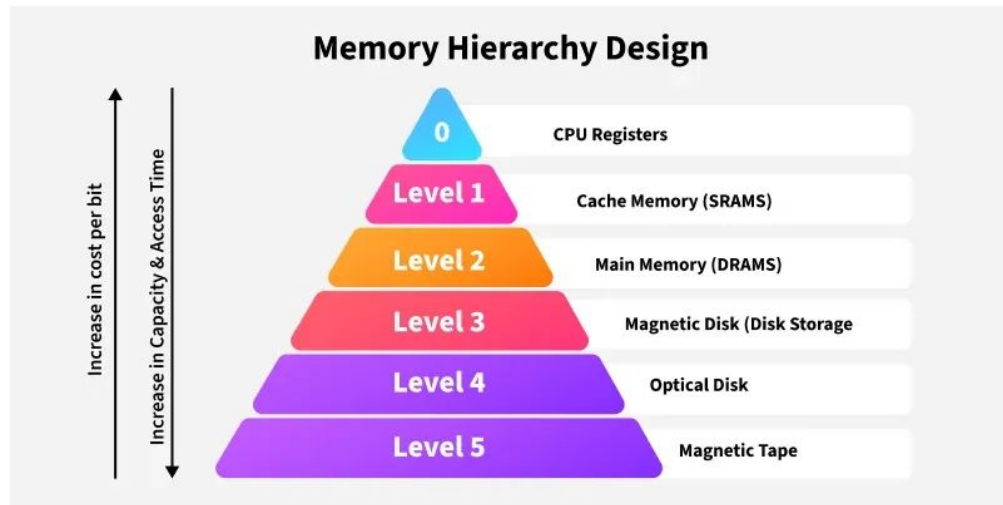

Memory Hierarchy

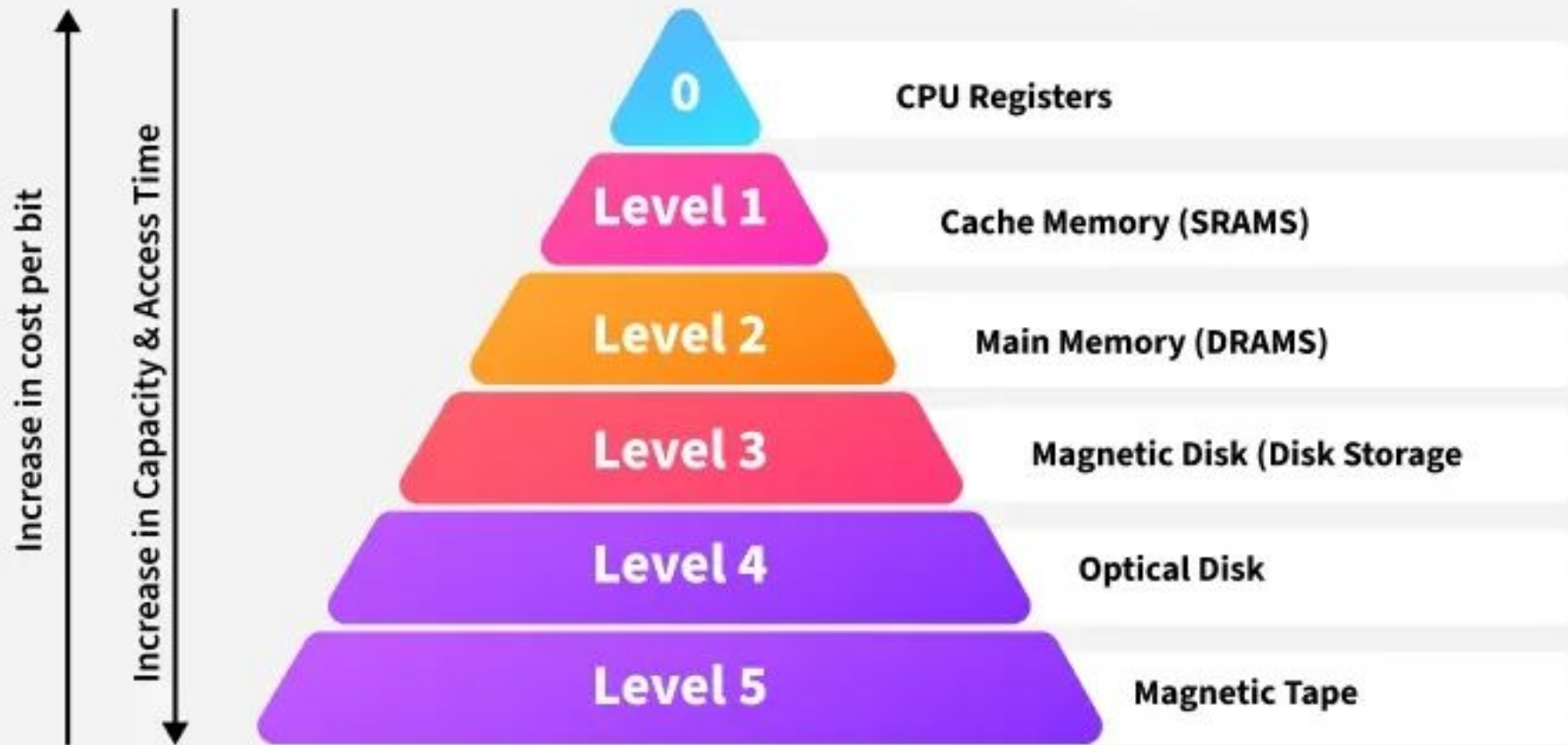
— Comp. Architecture Final Project —

What is Memory Hierarchy?

- A way of organizing memory types based on speed, cost, and capacity, with the goal of improving performance by minimizing access time.
- Fastest and most expensive at the top, and slowest and cheapest at the bottom.

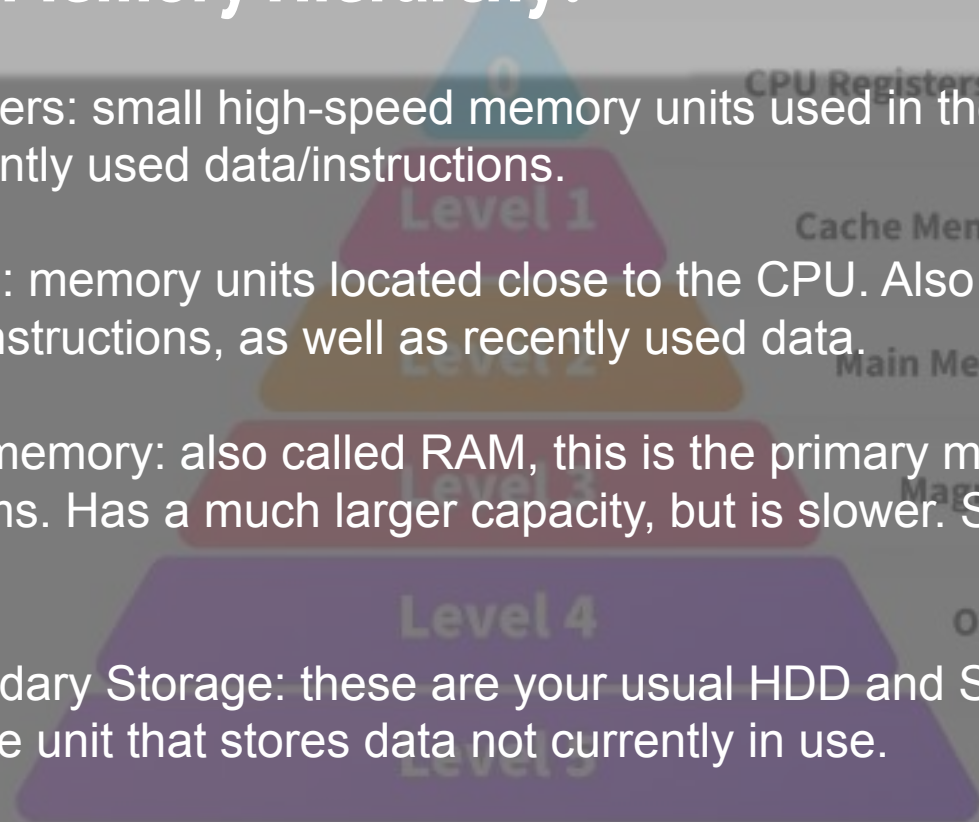


Memory Hierarchy Design



Memory Hierarchy Design

Who is Memory Hierarchy?

- 
- The diagram illustrates the memory hierarchy as a pyramid with five levels. From top to bottom, the levels are: Level 0 (CPU Registers), Level 1 (Cache Memory (SRAMS)), Level 2 (Main Memory (DRAMS)), Level 3 (Magnetic Disk (Disk Storage)), and Level 4 (Optical Disk and Magnetic Tape). To the left of the pyramid, two vertical arrows indicate trends: 'Increase in cost per bit' pointing upwards and 'Increase in capacity & access time' pointing downwards.
- Registers: small high-speed memory units used in the CPU to store the most frequently used data/instructions.
 - Cache: memory units located close to the CPU. Also stores frequently used data/instructions, as well as recently used data.
 - Main memory: also called RAM, this is the primary memory used in computer systems. Has a much larger capacity, but is slower. Stores currently used data.
 - Secondary Storage: these are your usual HDD and SSDs; a non-volatile storage unit that stores data not currently in use.

Why is Memory Hierarchy?

- Performance: Frequently used data is stored in faster memory (like cache), reducing access time and improving overall system performance.
- Cost Efficiency: By combining small, fast memory (like registers and cache) with larger, slower memory (like RAM and HDD), the system achieves a balance between cost and performance. It saves the consumer's price and time.
- Optimized Resource Utilization: Combines the benefits of small, fast memory and large, cost-effective storage to maximize system performance.
- Efficient Data Management: Frequently accessed data is kept closer to the CPU, while less frequently used data is stored in larger, slower memory, ensuring efficient data handling.

Why not Memory Hierarchy?

- Complex Design: Managing and coordinating data across different levels of the hierarchy adds complexity to the system's design and operation.
- Cost: Faster memory components like registers and cache are expensive, limiting their size and increasing the overall cost of the system.
- Latency: Accessing data stored in slower memory (like secondary or tertiary storage) increases the latency and reduces system performance.
- Maintenance Overhead: Managing and maintaining different types of memory adds overhead in terms of hardware and software.

How is Memory Hierarchy?

- Principle of Locality refers to the tendency of a computer system to access the same set of memory locations repeatedly over a short period of time.
 - Temporal locality - an accessed location is more likely to be accessed again in the near future
 - Spatial locality - nearby locations are likely to be accessed as well
- The memory paging management scheme that loads the parts of data as is needed takes advantage of this.

Where is Memory Hierarchy?

- Consider the intel core i9;
 - Registers in each core
 - CPU Caches L1, L2 and L3
 - L1 is the fastest cache
 - L2 is private per core and stores data that doesn't fit in L1
 - L3 is shared across cores and is the largest of the three cache levels and can be up to 36MB
 - RAM handles cache misses that don't fit in L1-L3



References

GeeksforGeeks. (2022, December 6). *Memory hierarchy design and its characteristics*.

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

GeeksforGeeks. (2023, January 11). *Magnetic disk memory*. <https://www.geeksforgeeks.org/magnetic-disk-memory/>

GeeksforGeeks. (2023, February 8). *Locality of reference and cache operation in cache memory*.

<https://www.geeksforgeeks.org/locality-of-reference-and-cache-operation-in-cache-memory/>

Wikipedia contributors. (n.d.). *Locality of reference*. Wikipedia. Retrieved June 10, 2025, from

https://en.wikipedia.org/wiki/Locality_of_reference

Wikipedia contributors. (n.d.). *Memory paging*. Wikipedia. Retrieved June 10, 2025, from

https://en.wikipedia.org/w/index.php?title=Memory_paging

Intel Corporation, "Intel® Core™ i9-13900K Processor (36M Cache, up to 5.80 GHz) - Specifications," *Intel ARK*, [Online]. Available:

<https://www.intel.com/content/www/us/en/products/sku/230496/intel-core-i913900k-processor-36m-cache-up-to-5-80-ghz/specifications.html>