Topic 17 Hardware – Summary

Vocabulary

stored-program computer (n): a computer where the program instructions are stored in memory – whether the same memory as data, or separate memory.

Von Neumann architecture (n): a computer in which instructions are stored in memory with the data. Modern computers use this architecture at the level of RAM.

Harvard architecture (n): a computer in which instructions are stored in memory, but not in the same memory as the data. Modern high-performance computers use this architecture for level 1 cache.

central processing unit (**CPU**) (n): a hardware device that manages and executes computer instructions (general purpose arithmetic, logic, and control operations) and functions as the core coordinator of the computer system

volatile (adj): memory that is erased when the power is turned off

non-volatile (adj): memory that is retained (not lost) when the power is turned off.

random access memory (**RAM**) (n): a temporary, volatile store for data and instructions.

bus (n): a group of connections between electrical devices in a computer

memory address (n): a number that uniquely identifies a memory storage location

bottleneck (n):

microprocessor (n): page 169

arithmetic logic unit (**ALU**) (n): the part of the CPU that performs calculations and logic operations

floating point unit (**FPU**) (n): in addition to an ALU, advanced processors may also contain a floating-point unit that performs calculations on floating point numbers (real numbers).

Concepts

Memory Access Categories

sequential	Data must be accessed in a fixed, linear sequence. To get to a specific piece of data, you must pass through all the data that comes before it.
random	Any storage location can be accessed directly and in approximately the same amount of time, regardless of its physical location.
direct	This is the hybrid category for devices like hard drives. Access is not purely sequential, but it's also not truly random because the access time varies depending on the physical location of the data.

cassette tape
(sequential access)

memory chips
(random access)

hard drive
(direct access)

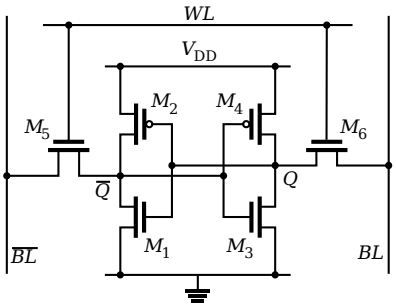
Topic 17 Hardware – Summary

Read-Only Memory (ROM) – The History

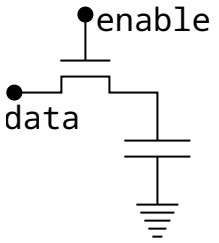
ROM	The first ***read-only memory***, or ROM, was truly read-only as it was hard-wired during its assembly.
PROM	***Programmable read-only memory*** was developed that allowed users to burn fuses in the chip, storing the data they wished: write one time, read many.
EPROM	***Erasable programmable read-only memory*** was erasable using ultraviolet light, allowing the ROM to be written multiple times, saving time and cost when developing software
EEPROM	***Electrically erasable programmable read only memory*** finally made it possible for non-volatile memory to be erased and re-written in-circuit. Initially, these chips required high voltage input, but advancements have removed this requirement
Flash	While traditional EEPROM is byte-erasable, Flash memory is a type of EEPROM where memory can only be erased in blocks. Block size has been increasing with increased memory capacity.

The Structure of Random Access Memory (RAM)

The two main architectures of memory are ***static RAM*** (**SRAM**)and ***dynamic Ram*** (**DRAM**).



static random access memory (SRAM) cell



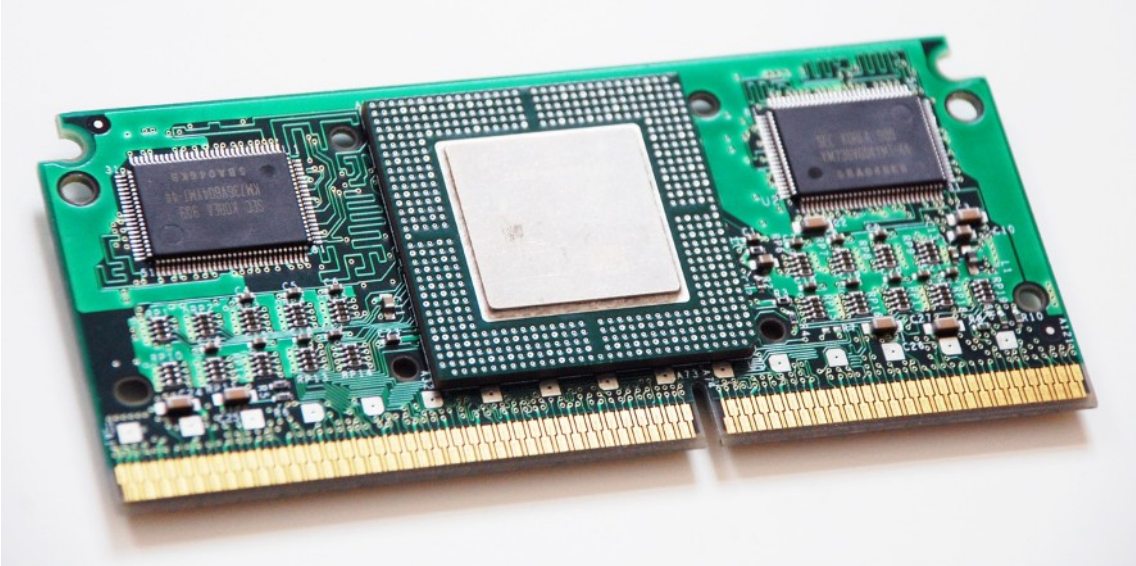
dynamic random access memory (DRAM) cell

Although non-volatile memory is also random access (any byte on the device can be read without the need to go through the device sequentially)) called ***random access memory***, or **RAM**.

Beyond Registers: The Computer Storage Hierarchy

Cache is very fast *volatile* storage in close proximity to the processor. There can be multiple levels of cache – the lower the level, the closer to the processor.

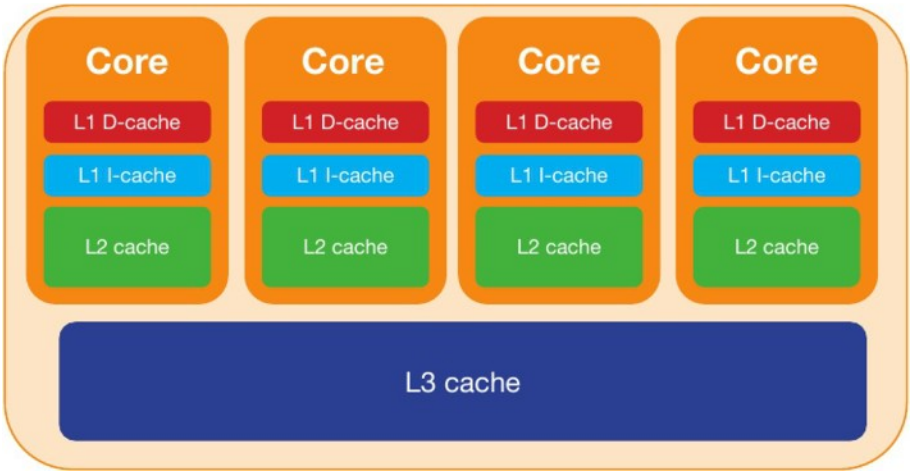
Historically, in the 1990s to early 2000s, single-core processors had ***level 1*** cache on-die, while ***level 2*** cache was on a separate silicon die – first in a separate chip or chips on a circuit board that slotted into the motherboard (as shown in the image below), then later, the two dies were placed in a single chip package.



Slot 1 PCB with Pentium II CPU and Level 2 Cache

Currently cache is as follows:

- ***level 1 cache***: inside the CPU, private to each CPU core; fastest
- ***level 2 cache***: larger than level 1; may be private to a core, or shared
- ***level 3 cache***: largest on-die cache; shared between all cores



In the diagram, above, Level 1 cache is separated into ***instruction cache*** and ***data cache***. Recall that having instructions and data in the same memory space is called the ***von Neumann architecture***. Separating instruction and data memory is called the ***Harvard architecture***. This separation allows for increased performance, and is implemented in most high-performance processors, such as x86 and Apple M-series processors.

Topic 17 Hardware – Summary

Current approximate cache speeds and sizes (as of 2025) are shown in the table:

	Size	Access Time (clock cycles)	Access Time (ns)
Level 1	32-64 kB	4-5	1-2ns
Level 2	256kB – 1MB	~12	4-10ns
Level 3	Multi-MB	40-50	20-50ns
RAM (DDR4)	8GB-1TB	140	50-80ns

RAM (Random Access Memory): very fast memory (but slow compared to cache).

Note that the cache has it's own controller – the central logic of the CPU simply requests a memory address, and the cache controller determines if it's in cache, and when and how to load it if it is not.

**** cache miss***

Storage Hierarchy Summary:

- ***primary***: volatile; fastest – registers, cache, RAM
- ***secondary***: non-volatile, immediately accessible by the OS – magnetic storage (HDD, tape drives), solid-state storage (SSD), optical drives (CD, LaserDisc, DVD, Blu-Ray), cloud storage
- ***tertiary***: mass non-volatile storage for backup or archive – off-site tape backups, or even an optical storage jukebox.

Virtual Memory

swap, page; "disk thrashing"