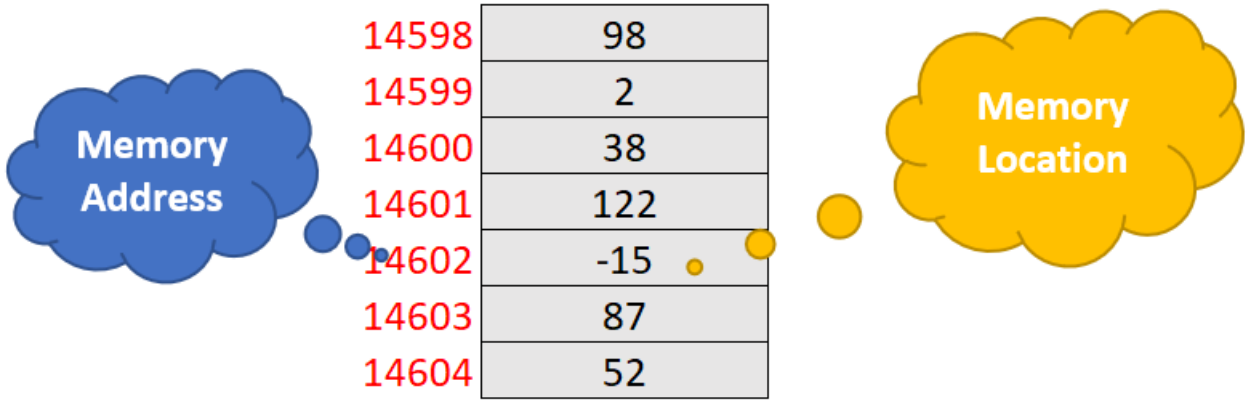
Primary Memory

Primary storage, also known as '**main storage**' or '**memory**', is the main area in a computer in which data is stored for quick access by the computer's processor. On today's smaller computers, especially personal computers and workstations, the term '**random access memory**' (**RAM**) - or just 'memory' - is used instead of primary or main storage, and the hard disk, diskette, CD, and DVD collectively describe 'secondary storage' or auxiliary storage.

Main memory, to be more exact, consists of all the (electronic) memory that is directly accessible by the processor i.e. RAM, ROM and the various kinds of Cache. Main memory can be considered as a sequence of locations where each location has an address that identifies it. This is shown in the following diagram.

Each memory location is identified by means of an address.



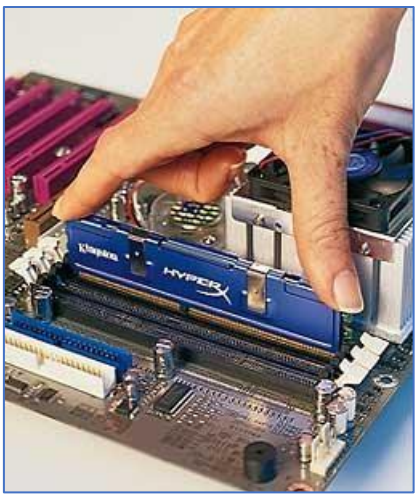
A small part of memory

RAM

RAM is an acronym for 'random access memory', a type of computer memory that can be accessed randomly or directly. RAM is the most common type of memory found in computers and other devices, such as printers.

There are two different types of RAM: **DRAM** (Dynamic Random-Access Memory) and **SRAM** (Static Random-Access Memory). The two types differ in the technology they use to hold data, with DRAM being the more common type. In terms of speed, SRAM is faster. DRAM needs to be refreshed thousands of times per second while SRAM does not need to be refreshed, which is what makes it faster than DRAM. DRAM supports

access times of about 60 nanoseconds, SRAM can give access times as low as 10 nanoseconds. Despite SRAM being faster, it is not as commonly used as DRAM because it is so much more expensive. Both types of RAM are volatile, meaning that they lose their contents when the power is turned off. DRAM is used as normal RAM while SRAM is used as cache.



A RAM module being placed in a socket

Kilo, Mega, Giga, Tera

The following table shows the terms used to describe the amounts of memory.

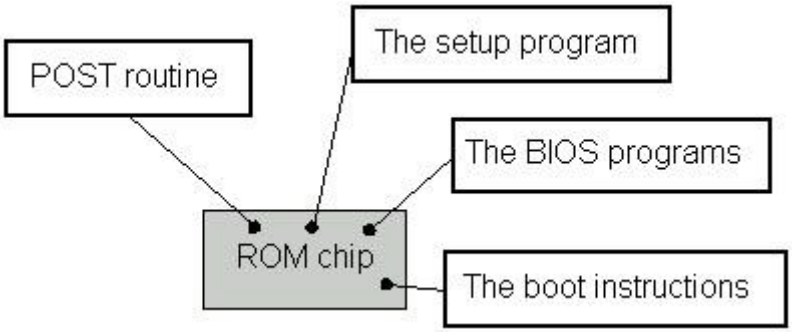
Abbr.	Prefix name	Decimal size	Size in thousands	Binary approximation	Address variable size
K	kilo-	10^3	1,000	$1{,}024 = 2^{10}$	10
M	mega-	10^6	$1{,}000^2$	$1{,}024^2 = 2^{20}$	20
G	giga-	10^9	$1{,}000^3$	$1{,}024^3 = 2^{30}$	30
T	tera-	10^{12}	$1{,}000^4$	$1{,}024^4 = 2^{40}$	40
P	peta-	10^{15}	$1{,}000^5$	$1{,}024^5 = 2^{50}$	50
E	exa-	10^{18}	$1{,}000^6$	$1{,}024^6 = 2^{60}$	60

ROM

The '**Read-Only Memory**' is non-volatile. ROM contains the data needed to start up the computer. The computer is designed in such a way that when it is switched on it executes programs contained in the ROM.

Different ROM-type memories contain these essential start-up data, i.e.:

- The **BIOS** is a program for controlling the system's main input-output interfaces, hence the name BIOS ROM.
- The **bootstrap loader** is a program for loading the operating system from secondary storage to RAM and launching it.
- The **CMOS** (Complementary metal–oxide–semiconductor) holds data like the time, date, and configuration information. CMOS requires constant (low) power to keep the information.
- The **Power-On Self-Test** (POST), a program that runs automatically when the system is booted, thus allowing the system to be tested.



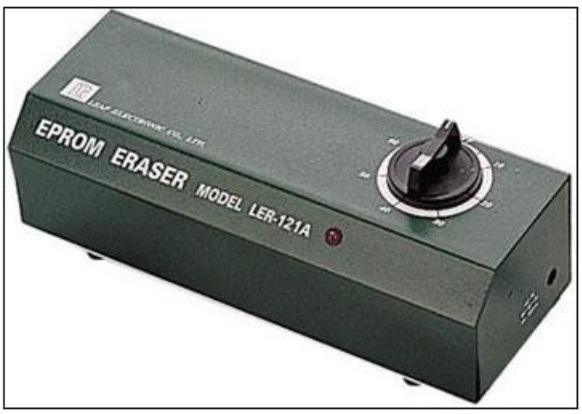
The contents of the ROM chip of a personal computer

A CMOS ROM chip containing the BIOS

Given that ROMs are much slower than RAM memories (access time for a ROM is around 150 ns whereas for SDRAM it is around 10 ns), the instructions given in the ROM are sometimes copied to the RAM at start-up; this is known as **shadowing**, though is usually referred to as **shadow memory**).

ROM memories have gradually evolved from fixed read-only memories to memories than can be programmed and then re-programmed:

- ROM: The first ROMs had their programs and data written at the manufacturing stage and the information in them could not be altered.
- **PROM**: PROM (Programmable Read Only Memory) memories were developed at the end of the 70s. Information in them could "burnt" using a device called a "ROM programmer". Once the information was written in them it could not be altered.
- **EPROM**: EPROM (Erasable Programmable Read Only Memory) memories are PROMs whose data can be deleted, and they could be reprogrammed. However, to do this process the chips have to be pulled out of the computer.
- **EEPROM**: EEPROM (Electrically Erasable Read Only Memory memories are also erasable PROMs, but unlike EPROMs, they can be erased by a simple electric current, meaning that they can be erased even when they are in position in the computer.
    - There is a variant of these memories known as **flash memories** (also Flash ROM or Flash EPROM).



EPROM programmer and eraser

Cache

Cache memory is a chip-based computer component that makes retrieving data from the computer's memory more efficient. It acts as a temporary storage area that the computer's processor can retrieve data from easily. This temporary storage area, known as a cache, is more readily available to the processor than the computer's main memory source, typically some form of DRAM.

Cache memory is sometimes called **CPU memory** because it is typically integrated directly into the CPU chip or placed on a separate chip that has a separate bus interconnect with the CPU. Therefore, it is more accessible to the processor, and able to increase efficiency, because it is physically close to the processor.

To be close to the processor, cache memory needs to be much smaller than main memory. Consequently, it has less storage space. It is also more expensive than main memory, as it is a more complex chip that yields higher performance.

What it sacrifices in size and price, it makes up for in speed. Cache memory operates between 10 to 100 times faster than RAM, requiring only a few nanoseconds to respond to a CPU request.

The actual hardware that is used for cache memory SRAM.

Types of cache memory

Traditionally, it is categorized as "levels" that describe its closeness and accessibility to the microprocessor. There are three general cache levels:

L1 cache, or primary cache, is extremely fast but relatively small, and is usually embedded in the processor chip as CPU cache.

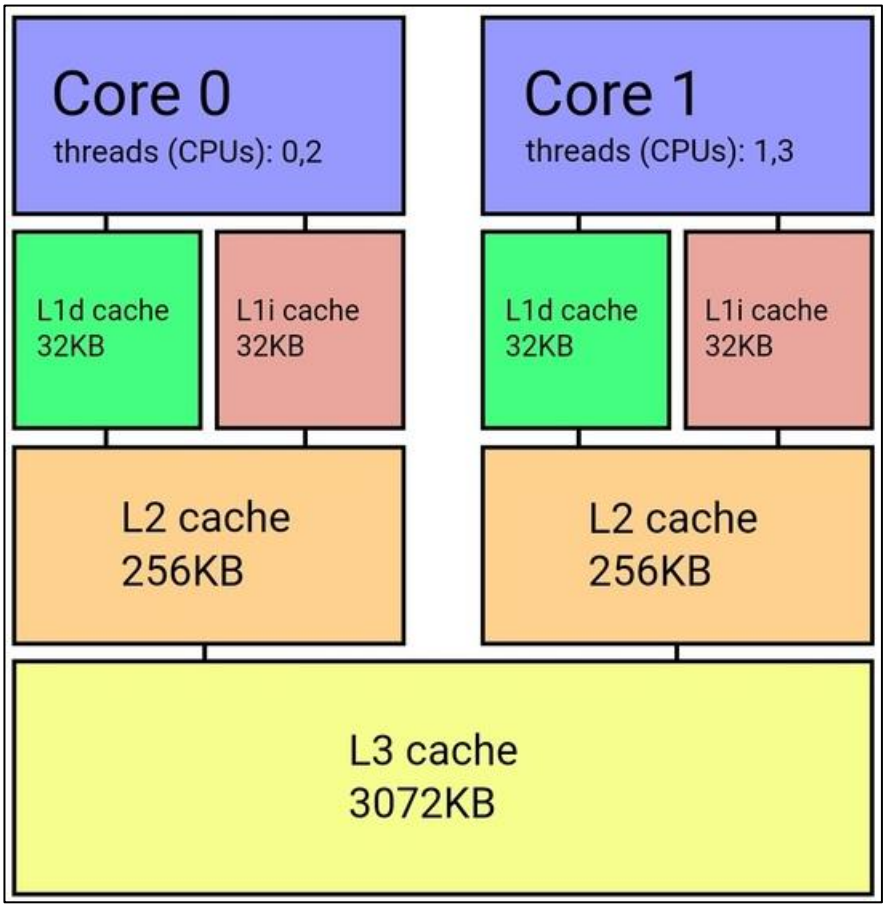
L2 cache, or secondary cache, is often more capacious than L1. L2 cache may be embedded on the CPU, or it can be on a separate chip or coprocessor and have a high-speed alternative system bus connecting the cache and CPU. That way it does not get slowed by traffic on the main system bus.

L3 cache is specialized memory developed to improve the performance of L1 and L2. L1 or L2 can be significantly faster than L3, though L3 is usually double the speed of DRAM. With multicore processors, each core can have dedicated L1 and L2 cache, but they can share an L3 cache. If

an L3 cache references an instruction, it is usually elevated to a higher level of cache.

In the past, L1, L2 and L3 caches have been created using combined processor and motherboard components. Recently, the trend has been toward consolidating all three levels of memory caching on the CPU itself.

Another type of cache is **memory caching** which is a DRAM buffer for the hard disk.



L1, L2, L3 caches in a dual core processor