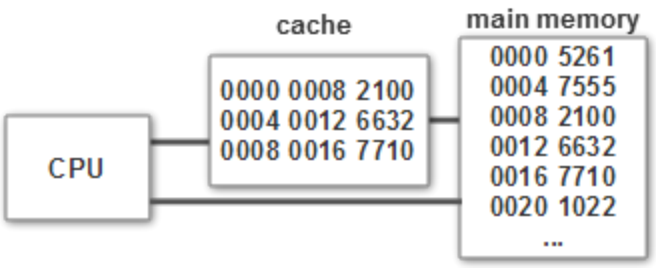
Caching

The first use of *cache* (also *casshe*) was for a hiding place, especially of goods, treasure, and so forth. The 1971 compressed OED (*Oxford English Dictionary*) quotes this sentence from 1595: "The inhabitants havinge intelligence of our cominge had .. hid theyre treasure in casshes." The word comes from the French word *casher*, meaning "to hide".

Cache is pronounced the same as *cash*; they are homophones. The brash moustached crook rashly stashed the stolen cash in a trash can full of ash; the police found the cache in a flash, a blink of an eyelash, and the crook gnashed his teeth in dispair.

A CPU memory cache

In a computer with a CPU (Central Processing Unit), it takes time to fetch data or instructions from the main memory. In order to speed up the process at least some of the time, the CPU may have its own memory —small and close by, so that accessing it is efficient— where it keeps copies of recently accessed parts of the main memory. It's a cache, a small treaure trove, hidden from the user's view.



We illustrate this using a hypothetical example, shown to the right. Main memory consists of words, each containing 4 bytes. The word at location 0008 contains the number 2100. The CPU has retrieved the word at location 0008 —and also the next two words because neighboring words will often be accessed— and stored them in its small cache. If asked to access word 0008, 0012, or 0016, the CPU gets the word from the cache. If asked to access some other word, the word (and some of its surrounding words) will be retrieved from main memory and stored in the cache, replacing other words in the cache because the cache space is limited.

If the CPU has to write a number to main memory location 0008, the CPU writes it quickly to the cache and then goes on about its business while, in parallel, the number is written from the cache to the main memory.

That, in a, nutshell, is how a memory cache works. The MacBook Pro on which this document was written (in 2018) has 16 GB of memory. Each of its 4 Processing Units (CPUs) has a 256 KB cache; these are called *L2 caches*. There is also an *L3 cache* of 8 MB, which serves all four Processing Units. A processor looking for a word in memory looks first in its L2 cache; if not there, it looks in the L3 cache; if not there, it gets it from memory.

Other caches on computers

The CPU memory cache is a piece of hardware. Caches are used in other places in computing systems, and the cache might be in software as well as hardware. For example,

- Graphics processor units (GPUs) often have several kinds of caches to increase efficiency.
- Hard disk drives (HDDs) often come with built-in caches, perhaps of size 64MB or 256MB. These affect read performance more than write performance, making it quicker to retrieve data.
- Web browsers use caches to save both images and web pages that are expected to be used again, so they don't have to be retransmitted across the network. These are software caches.