

CACHE MEMORY

Most PCs are held back not by the speed of their main processor, but by the time it takes to move data in and out of memory. One of the most important techniques for getting around this bottleneck is the memory cache.

The idea is to use a small number of very fast memory chips as a buffer or cache between main memory and the processor. Whenever the processor needs to read data it looks in this cache area first. If it finds the data in the cache then this counts as a 'cache hit' and the processor need not go through the more laborious process of reading data from the main memory. Only if the data is not in the cache does it need to access main memory, but in the process it copies whatever it finds into the cache so that it is there ready for the next time it is needed. The whole process is controlled by a group of logic circuits called the cache controller.

One of the cache controller's main jobs is to look after 'cache coherency' which means ensuring that any changes written to main memory are reflected within the cache and vice versa. There are several techniques for achieving this, the most obvious being for the processor to write directly to both the cache and the main memory at the same time. This is known as a 'write-through' cache and is the safest solution, but also the slowest.

The main alternative is the 'write-back' cache which allows the processor to write changes only to the cache and not to main memory. Cache entries that have changed are flagged as 'dirty', telling the cache controller to write their contents back to main memory before using the space to cache new data. A write-back cache speeds up the write process, but does require a more intelligent cache controller.

Most cache controllers move a 'line' of data rather than just a single item each time they need to transfer data between main memory and the cache. This tends to improve the chance of a cache hit as most programs spend their time stepping through instructions stored sequentially in memory, rather than jumping about from one area to another. The amount of data transferred each time is known as the 'line size'.

(Adapted from Cache Memory, PC Plus)

How a Disk Cache Works

Disk caching works in essentially the same way whether you have a cache on your disk controller or you are using a software-based solution. The CPU requests specific data from the cache. In some cases, the information will already be there and the request can be met without accessing the hard disk.

If the requested information isn't in the cache, the data is read from the disk along with a large chunk of adjacent information. The cache then makes room for the new data by replacing old. Depending on the algorithm that is being applied, this may be the information that has been in the cache the longest, or the information that is the least recently used. The CPU's request can then be met, and the cache already has the adjacent data loaded in anticipation of that information being requested next.

(Adapted from How a Disk Cache Works, PC Magazine)