Memory Caching

So as not to confuse you with information overload, I have deliberately left out the role of the processor memory caches in all of this.

Basically, the contents of the most frequently accessed memory addresses will be stored on the CPU itself (or just off) inside a multilevel data cache, the purpose being to speed up memory access with fast memory hardware and shorter access paths.

The data cache is split into three levels called L1, L2 and L3. The L1 and L2 caches are on the CPU, with L1 allowing for the fastest access, and typically storing between 8 and 64 KB. The L2 cache store is larger than the L1 cache (typically up to 2 MB), but access is slower.

The L3 cache is stored on the motherboard and is the largest cache store; for example, the Intel Core I7 processor has an 8 MB L3 cache. It's the slowest of all three caches, but still faster than direct memory access.

When a memory address is requested, the caches are checked first, before an attempt is made to access physical RAM, and the OS only actually has to deal with the paging file at the end of that checking sequence.

The access order is L1–>L2–>L3–>physical memory–>page file.

The intricacies of the data cache could fill a whole new chapter in themselves, but we'll not go there in this book, as I think we've covered enough for now.