Week 5: Cache-3

As the processor is often many times faster than the memory system, if left to its own the processor would need to wait a (relatively) long time for data to be retrieved. This is obviously quite inefficient: why bother having a fast processor if nothing can keep up with it? A processing system can only be as fast as its slowest component.

The concept of processor cache memory is that it is a small, ultra-fast pool of memory that sits between the processor and the standard memory system. Some support circuitry monitors how the processor is using the memory, and attempts to fill the cache memory up with information that the processor is most likely to need.

> Fast memory is expensive; that's why the entire memory system isn't made up of cache.

Such predictions are never perfect, but most computing tasks have a distinctive pattern in how they use memory. A correct prediction is called a cache "hit", an incorrect prediction is a "miss". Depending on the task, modern cache control systems can achieve cache hit rates exceeding 90%.

The disparity between processor and memory speed has become so great that modern computer systems include a number of cache "levels" of increasing size and decreasing speed. The idea behind this is that the most frequently- accessed data will end up in a faster, higher-level cache pool, while still providing much of the benefit of large cache sizes.

Level 1 cache

"Level 1" cache memory is a very small data cache that exists right on the processor itself, running at the same break-neck speed as the processor's execution units. (Some designs have level-1 cache that runs at half speed.)

As it is running at the same speed as the processor, there is no delay when data is fetched from the cache.

Level 2 cache

Due to the cost constraints of high-speed memory, the level 1 cache is often backed up by a larger "level 2" cache: this is a slower bank of memory, still residing on the CPU itself.

Level 3 cache

(Can you see where this is heading?) modern desktops, servers and high-end workstations, include a third level of cache. This is an even larger, slower than Level 2 (but still faster than main memory) cache pool. The size of these caches can dictate that it is on a separate chip from the processor.

Miniaturisation of transistors on silicon has now got to the point where Level 3 caches can reside on the CPU itself.