No. 63, Feb 1999

THE COMPUTER CORNER

Cache as Cache Can

- by Stan Kaplan, WB9RQR
105 Martin Drive
Port Washington, WI 53074-9654
(414) 284-9346
WB9RQR @ N9PBY.EN63BI.WI.USA.NA
skaplan@mcw.edu

What is all this stuff about caches? Well, the word comes from the French word cacher, which means to hide, and a dictionary definition says it is a place for hiding or storing anything. In the computer world, a cache is a buffer or holding place, but a holding place with some intelligence. For example, data flowing between the faster Central Processing Unit (CPU) and slower peripherals (either way) is watched. That data that is accessed most often is placed in the cache. How is the data to be cached selected? By an intelligent mathematical algorithm (set of formulas), which actually scoops out copies of data from the rivers of data flowing past. .

Well, so what? What good does that do for a computer system? The answer is, it can significantly speed up how fast the computer system operates. For example, in memory caching, special fast memory chips keep copies of the most recent memory accesses. If the CPU asks for this data a second time, it is supplied by the super fast memory chips rather than the slower system memory (RAM). Get the general picture?

Here is an analogy. Suppose you, Dick and I are sitting in a room full of phone books from every city in the US. You ask me to read you all the names and phone numbers. I begin, taking book after book from the pile, and reading at a rate of 10,000 names and numbers per second (please allow that this speed is possible!). Meanwhile, Dick takes a wild but educated guess that you might later be interested in names beginning with the letter Q, and phone numbers starting with any digit followed by 89 (as in 789-1234). Therefore, Dick records each name and phone number fitting either criterion in a notebook; he is very fast and can do that three times the speed at which I read. After a bit over 100 hours, we are done.

Next, as Dick partially guessed, you ask me to give you both the full name and phone number of each person whose name begins with Q, and whose phone number begins with 989. Normally, I would have to look through every phone book again to find the data. However, Dick pipes up, and says he already has the data written down. Indeed, he does - all names beginning with Q and all numbers starting with 189, 289, 389 … 989. Dick actually has more data than you requested, but for sure he has everything you want among it. Dick has a 100% hit rate, and he only has to look through his written notes to find it, rather than each and every phone book. On top of all that, Dick is 33% faster than I am at finding and calling out the data. Dick and his notebook represent a very efficient, very fast cache.

Now the neat thing is that some CPUs, starting from the 486 chip, have a memory cache built right into the chip! The 486 has what is known as a Level 1 Cache, right inside the CPU package itself. The cache is either 8,000 or 16,000 bytes in size (depending on the particular model of 486 CPU), and the engineers designed it so well that it has a "hit" ratio of between 90 and 95%. That means that between 90 and 95% of the data it copies from the river of data going by is asked for a second time, and therefore supplied by the very fast cache!

Lets explore why this is important in a little more depth. Ordinary RAM (30 or 72 pin SIMMs, for example) is plugged directly into the motherboard. Since the memory chips on these SIMMs must both receive and transfer data through the motherboard, they can operate only at the speed of the motherboard. The CPU, on the other hand, is often running at some multiple of the speed of the

motherboard. For example, a 333 MHz Pentium II CPU chip runs at 333 MHz, but its motherboard runs at only 66 MHz, one-fifth the speed of the CPU. Therefore, if the CPU needs data from the SIMMs, it has to wait around for it to be delivered through the more slowly operating motherboard. On the other hand, that CPU's L1 cache (which is right inside the CPU itself) can also run at or near 333 MHz. Therefore, the CPU doesn't have to wait around if the data is in the L1 cache. L1 cache memory is the only memory that can keep up with the speed of the CPU in all but the latest computers.

Most 486 systems incorporate a small amount of very high speed (and very expensive) memory on the motherboard known as L2 cache which can help to speed things up, too. However, because it is on the motherboard it cannot outperform the L1 cache. Furthermore, there is a cost/benefit ratio to be considered. Author [1]Scott Mueller suggests that you install the middle amount of L2 cache your computer can accept. Thus, for a machine, which can take up to, 128 kb of L2 cache, installation of 64 kb would probably give you the best bang for your buck.

The latest Pentium chips have taken us to a new height in memory caching. They incorporate the Level 2 cache right inside the same chip as the CPU. That is, the L2 cache is a separate die (chip), but it is enclosed within the CPU package and can therefore be made to run at speeds up to that of the CPU itself. Chipmakers are exploring new vistas in memory design even as you read this. Look for new announcements within months concerning the way memory is designed and packaged.

Indeed, things are moving so fast that I feel compelled to make a new prediction. Look for talk of a fully functional computer on a chip. It will come in a 4 x 4-inch box with the single chip inside. The rest of the little box's space will be taken up with sockets - one for a monitor, one for a mouse/microphone/keyboard, one for a power cord and one for any other peripheral or combination of peripherals you might care to plug in. The step after that will be disappearance of the plugs, as the unit becomes totally wireless and battery powered. Following that, the unit will need no manual input at all. It will tune to your brain waves and respond to your thoughts. Happy computing!

---

[1]Mueller, Scott ***Upgrading and Repairing PCs***, 10th edition. ISBN 0-7897-1636-4. Que Corporation. This is the best general book on PCs available today; a must have if you wish to learn about computer hardware and how it is all put together. This edition was published in September 1998.