Cache Strategies Key to Future CPUs

On-Chip L2 Cache to Become Widespread by 1999

In the PC microprocessor market, the biggest changes this year will be not in the CPU cores but in the cache subsystems. Driven by the need for higher cache bandwidth at low cost, and enabled by 0.25-micron process technology, all the x86 suppliers except Cyrix have announced plans for processors with on-chip L2 caches. These caches will be essential to extending the life of Socket 7.

Intel was the first x86 supplier (not counting NexGen's ill-fated 586) to boost L2 cache bandwidth with a dedicated cache bus, first with Pentium Pro and then with Pentium II. Both processors use off-chip, but in-module, L2 caches. This was essential in 0.35-micron technology, and even in a 0.25-micron process, it enables Intel to provide relatively large caches. By putting the processor chip and the L2 cache in a module, Intel is free to use various L2 cache architectures without affecting the motherboard design.

By the end of the year, Intel will offer Pentium II modules with 512K L2 caches running at half the CPU speed; larger caches running at the full CPU speed; 128K–256K on-chip L2 caches that eliminate the SRAMs from the module (code-named Mendocino); and, at the very low end, a module with no L2 cache at all (code-named Covington).

All these products except Mendocino will be based on the same Deschutes CPU die. Intel will also offer Deschutes on a Mobile Module that will be pin-compatible with the Pentium/MMX Mobile Module as well as on a new mini-cartridge that is essentially a compact Slot 1 module. This approach gives Intel a very wide range of products, covering the spectrum from the least-expensive PCs to high-end servers, and from notebooks to workstations, all using one CPU chip. This will enable Intel to achieve extraordinary economies of scale, focusing the vast majority of its silicon fabrication efforts this year on the Deschutes die.

Covington is presumably a stopgap product driven by a sense of urgency to serve the low-cost PC market with a Slot 1 product so PC makers can focus all their efforts on Slot 1 motherboards. Eliminating the cache will cut the performance dramatically; sources indicate that on typical business applications, a Covington-266 will be slower than a Pentium/MMX-233 with L2 cache but faster than a cacheless Pentium/MMX-233. Thanks to the P6's faster FP and MMX units, Covington should be faster than Pentium/MMX on applications that stress those functions.

Covington will compare poorly with the products expected from Intel's competitors, which will run at 266 MHz in lower-cost Socket 7 motherboards and are likely to deliver higher performance with low-cost external L2 caches. If Covington boosts the fortunes of alternative Socket 7 processors by making them look good by comparison, it could be a mistake. Assuming Intel prices Covington aggressively, however, it could force the competitors to price their processors lower than they would like.

Mendocino, with an integrated L2 cache, should perform far better than Covington and scale to higher clock speeds. Intel will need Mendocino to compete against Socket 7 processors with on-chip L2 caches from AMD and IDT.

Mendocino's cache will be smaller than that in today's Pentium II, but it will be faster. The cost of the SRAMs it replaces is modest, but eliminating the off-chip cache simplifies the module, lowers the cost of the CPU chip package by eliminating the cache bus, and makes it practical to sell the CPU chip by itself.

For Intel's competitors, an on-chip L2 cache gives them a way to gain the performance benefit of a backside cache bus while retaining the Socket 7 interface. The die-size premium is not insignificant, but the die size of today's CPUs with 256K of cache in a 0.25-micron process will be manageable. AMD, for example, says that the K6+ 3D will be only 135 mm^2, compared with 81 mm^2 for the version without L2 cache.

Freed from the constraints of the system bus, an on-chip L2 cache can run at the CPU speed, regardless of the speed of the system bus. And while the 66-MHz cache RAMs used in today's PCs cost only a few dollars, cache RAMs that run at 300 to 400 MHz are another matter entirely—even if there is a way to connect these RAMs to the processor.

All the initial products with on-chip L2 caches are likely to maintain 64-bit interfaces between the CPU and the cache to minimize the design changes to the CPU core. Future designs could implement wider on-chip interfaces, providing much higher bandwidth at little incremental cost. Although leading-edge CPUs probably will continue, for another generation or two, to push the limits of die size without on-chip L2 caches, once they have gone through one process shrink on-chip L2 caches are likely to be the best approach.

When Intel introduces Katmai in 1999, for example, we expect it will debut without on-chip L2 cache in 0.25-micron technology. When this processor moves to 0.18-micron technology, however, a 512K on-chip cache should be possible. From this point forward, modules with off-chip L2 caches probably will be limited to the high end of the product line. M

See www.MDRonline.com/slater/cache *for more on this subject. I welcome your feedback at* mslater@mdr.zd.com.