Cahners
MICROPROCESSOR REPORT
THE INSIDERS' GUIDE TO MICROPROCESSOR HARDWARE
VOLUME 13, NUMBER 14
OCTOBER 25, 1999

Coppermine Outruns Athlon

Single-Chip Pentium III Boosts Performance, Cuts Cost

MICROPROCESSOR FORUM 1999

by Linley Gwennap

Hidden behind the same brand name as its predecessor, Intel's 0.18-micron Pentium III processor adds features and performance optimizations to strengthen its market position. The new chip, code-named Coppermine, was announced today in desktop, mobile, and workstation/server versions at clock speeds of up to 733 MHz. At this speed, the chip delivers better integer performance than any other, including new rival Athlon, as Figure 1 shows.

At this month's Microprocessor Forum, Intel architecture manager Jim Wilson described the new chip's design. Enhancements over the previous Pentium III, known as Katmai (see MPR 3/8/99, p. 1), include an on-die 256K level-two (L2) cache, an improved bus interface, and, for the mobile segment, a variable voltage and clock-speed feature called SpeedStep (formerly known as Geyserville). Along with boosting clock speed, the 0.18-micron version reduces cost by bringing the die size down to a highly manufacturable 106 mm^2, despite the addition of the on-chip cache. This cache eliminates the need for the bulky, expensive module used for Katmai-based Pentium IIIs.

Another benefit of the 0.18-micron process is lower power. Coppermine enables Pentium III to reach the mobile market for the first time, providing the benefits of both higher clock speeds and the SSE multimedia extensions. Intel today announced mobile versions at speeds up to 500 MHz, and the mobile parts will exceed 700 MHz next year in the higher-power SpeedStep mode. The integrated cache offers the further benefit of a reduced footprint for the smallest desktop and mobile systems.

Integrated Cache Reduces Cost

Approximately 30% of the Coppermine die is devoted to the 256K L2 cache. Intel did a full re-layout of the original Pentium III design to take advantage of the sixth metal layer available in the 0.18-micron P858 process (see MPR 1/25/99, p. 22) and to improve yield at higher speeds. (Note that, despite the product code name, the P858 process does not use copper.) The focus of the re-layout was on speed paths limited by RC delay, not transistor speed. As the company expects to produce faster P858 transistors over time, this effort should ensure continued clock speed increases. We expect the 0.18-micron design to reach 866 MHz by the end of 2000.

At 106 mm^2, Coppermine is 17% smaller than the 0.25-micron Katmai, despite the addition of the sizable L2 cache. Although the chip uses C4 (flip-chip) bonding, it still uses a perimeter pad ring to better support a variety of package types. According to the MDR Cost Model, the manufacturing cost of the new Pentium III is less than $40, a huge decrease from $65 for the 0.25-micron version. About half of this saving comes from the efficiency of the 0.18-micron process and the other half from integrating the L2 cache, eliminating the external SRAM and other module costs.

Why didn't Intel integrate the L2 cache sooner? In the 0.25-micron process, the cost savings are only half as much, because adding 256K of cache pushed the Pentium II (Dixon) die size to 180 mm^2, well beyond the "sweet spot" of

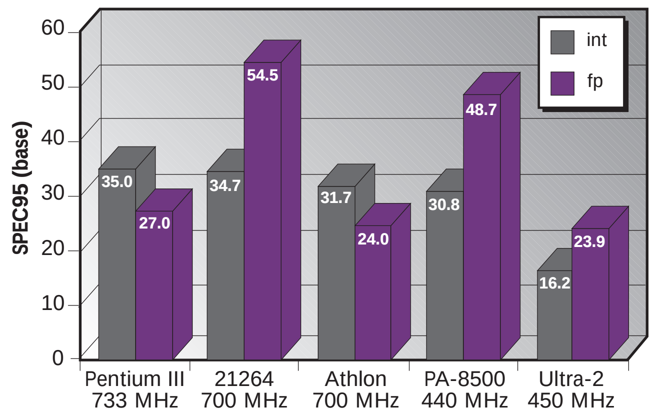


Figure 1. The 733-MHz Pentium III leads all shipping processors on SPECint95 (base) but not in floating-point performance. (Source: Intel, AMD, SPEC)

about 100 mm^2. Furthermore, Intel didn't have enough 0.25-micron capacity to shift all Pentium II shipments to such a large die, despite the theoretical cost savings. Thus, Intel offered its 0.25-micron integrated-cache designs only in the mobile and value segments, where the cost and power savings of integration are most valuable. With the 0.18-micron process, Intel can integrate the L2 cache in all segments.

New Design Improves Performance

The advantages of Coppermine's integrated cache go beyond cost. The 256K on-chip cache actually delivers better performance than the external 512K cache. The big win comes from lower latency. It takes far less time to move data across the die than to read it from an external SRAM. Intel would not disclose Coppermine's L2 cache hit latency, but Mendocino, Intel's previous part with integrated L2 cache, takes 8 cycles from load to use, versus 18 for Katmai.

In addition, the new Pentium III increases the width of the L2 bus to 256 bits (plus ECC), supplying an entire cache line in one cycle. This change does not reduce the load-use latency, as previous designs fetched the critical word first, but it allows the cache to service a new request every other cycle, delivering a peak bandwidth of 11.7 GBytes/s at 733 MHz.

Although technically this is a half-speed cache, it delivers twice the bandwidth of Xeon's external full-speed cache and four times the bandwidth of Katmai's external half-speed cache. Staying at half the CPU speed in Coppermine still supplies plenty of bandwidth without creating a critical timing path that might limit future speed increases.

Although the integrated cache holds half as much as Katmai's external cache, Coppermine's cache is eight-way set-associative, whereas the external cache is only four way. As a result, the hit rate of the 256K cache is nearly as good as

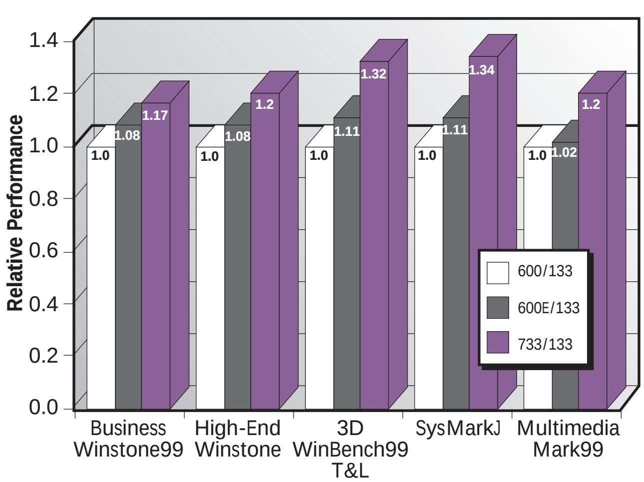


Figure 2. On most PC benchmarks, the Coppermine core (600E) offers at least one speed grade improvement over the Katmai core (600), with additional gains at 733 MHz. The benchmarks were run on an Intel VC820 board with an 820 chip set, 128M of RDRAM, an IBM 371800 ATA-66 disk, a Diamond Viper 770 Ultra TNT2 AGP4× card with 32M, and Windows NT 4. (Source: Intel)

for the 512K cache. Greater associativity is easier to achieve on die because routing wider buses is more feasible.

For large systems, the cache supports a 36-bit physical address space and MESI coherency. To improve yield, the cache includes redundant columns. The entire array can be quickly validated during manufacturing using a built-in self-test (BIST). A final benefit of the integrated cache is lower power dissipation: the Coppermine-733 dissipates 24.1 W (maximum), about 43% less per MHz than Katmai.

In addition to the improved cache, a faster system bus increases Coppermine's performance. The new processor supports a 133-MHz bus, 33% faster than in most Katmai systems. Although Intel recently released 533- and 600-MHz versions of Katmai that use the 133-MHz bus (see MPR 10/6/99, p. 30), Coppermine goes a step further by reworking the system-bus controller to better utilize the faster bus.

The new design includes six fill buffers (versus Katmai's four), eight bus queue entries (versus four), and four write-back buffers (versus one). These changes allow the processor to support more simultaneous bus transactions. Without them, the faster bus spends more time waiting for DRAM, which has not improved its latency. Intel's Wilson presented data showing a sustainable memory bandwidth of 1,010 MBytes/s, 95% of peak capacity, for Coppermine, versus only 64% of capacity for Katmai with a 133-MHz bus. These system-bus improvements will be most useful in applications with heavy memory traffic, as is the case with many 3D and multimedia programs.

The World's Fastest Integer Processor

On SPECint95, a 600-MHz Coppermine outscores a 600-MHz Katmai by 12%, or one full CPU speed grade. Because SPECint95 executes mainly out of cache, this increase shows the advantage of the reduced cache latency. The system-bus improvements come into play on the more memory intensive SPECfp95 tests, boosting performance by 20% over Katmai, even when both use a 133-MHz bus. This gain represents nearly three speed grades on this benchmark.

At 733 MHz, the new Pentium III delivers an estimated 35 int/27 fp on SPEC95_base. This integer figure is the highest yet posted for any shipping processor, RISC or x86, topping the 34 reported for a 700-MHz 21264 and 32 for an Athlon-700, as Figure 1 shows (see MPR 10/25/99, p. 1). Much of this advantage, however, is due to Intel's rapid deployment of its 0.18-micron process. AMD is already sampling a 0.18-micron Athlon that is likely to match Coppermine's performance, and Samsung's 0.18-micron 21264 is due early next year with significantly better performance.

Despite impressive improvements on the floating-point side, the new Pentium III still lags high-end RISC processors such as the 21264 and HP's PA-8500 by a nearly 2:1 margin. The new design does again push Pentium III ahead of Athlon, however. AMD hopes to regain the lead by improving its compiler optimizations and pushing its bus speed to 266

MHz, but the actual impact and timing of these changes remain to be seen.

On PC benchmarks, the increase in per-clock performance is more moderate, but still impressive, ranging from 8% to 11% in most cases, as Figure 2 shows. Coppermine delivers a gain of just 2% on MultimediaMark99, but this benchmark is CPU-bound and doesn't reflect the gains from the faster cache and bus. In all other cases, the increase is at least one full speed grade; that is, it is greater than the increase from 600 MHz to 667 MHz on the same core.

Coppermine to Displace Katmai Quickly

Given the cost and performance advantages of Coppermine, Intel hopes to quickly obsolete the older Katmai parts. To support this plan, the company is rapidly building 0.18-micron capacity, with four fabs expected to be producing parts using the new process by the end of this year. (In the past, Intel has brought up a new process in one fab at a time, but P858 is coming up nearly simultaneously across most of Intel's fabs.) With its compact die size, Coppermine yields more than 250 good die per wafer, according to our estimates, so Intel should be able to quickly deliver vast quantities of the new parts.

Today's (10/25/99) product announcement includes desktop Pentium III processors at 533, 600, 667, and 733 MHz using the 133-MHz bus. This unusually large number of speed grades covers price points from $305 to $776, allowing Coppermine to instantly occupy most of the performance (non-Celeron) desktop segment. We expect the price of the 533 to drop below $200 by 2Q00, driving Katmai completely out of the product line.

Unlike in previous transitions, Intel will continue to support the 100-MHz bus at higher CPU clock speeds, even while encouraging OEMs to move to the 133-MHz bus. Part of the reason for this unusual stance is that most OEMs don't like making platform transitions in the fourth quarter, the biggest selling season in both consumer and business segments. Intel had planned to launch both Coppermine and the 820 chip set (see MPR 10/6/99, p. 30) last spring, giving OEMs plenty of time to convert before the fourth quarter, but technical problems delayed both products.

As a result, many PC makers are temporarily sticking with the 440BX chip set and its slower bus, particularly for business systems. For these vendors, Intel also announced Coppermine processors at 550, 600, 650, and 700 MHz for the 100-MHz bus. To distinguish the various versions where there is clock-speed overlap, the 133-MHz versions carry a "B" designation, and the Coppermine parts have an "E" at the end. For example, there are 600, 600B, 600E, and 600EB versions, as described in Table 1.

Once the new year begins, OEMs should be ready to adopt the new bus speed, and the 820 should be available to support it in style. We don't expect Intel to announce any additional PC speed grades with a 100-MHz bus unless the 820 is severely delayed. But the company will continue to support the current 100-MHz parts well into next year to ensure a smooth transition.

Adding further complexity, Coppermine is currently available in two packages: the standard Slot 1 module and a PGA compatible with Celeron's Socket 370. The PGA package is much smaller, of course, and the socket is less expensive than a slot connector, so PC makers are likely to switch to it over time. Currently, however, Intel is charging a $20 premium for this package. Thus, it is mainly being used in small-form-factor PCs to save space. Early next year, the company should remove this premium and begin to get rid of the Slot 1 module once and for all, ridding all but the workstation/server line of the costly modules.

Low Prices Drive Transition

With such a complex product line, shown in Table 1, there are a few pricing anomalies to point out. The Coppermine-533 delivers much better performance than a Katmai-550 or even a Katmai-600, but at $305 it costs less than either, making it an excellent buy. The Coppermine-based 600E and 600EB carry the same price, for simplicity, but the latter will have better performance. Both will far outperform the Katmai-600 but cost $10 less. Even for OEMs looking solely at clock speed, Intel wants to make it clear that Coppermine should be preferred at all frequencies where it overlaps with Katmai.

The high-end Pentium III-733 carries an unusually high list price of $776. Intel has kept the high end of its line around $600 for most of this year, reflecting a lack of demand for the most expensive PCs. Sources indicate that Intel was

DESKTOP	CPU Core	Bus Speed	List Price
Pentium III-733	Coppermine	133 MHz	$776
Pentium III-700	Coppermine	100 MHz	$754
Pentium III-667	Coppermine	133 MHz	$605
Pentium III-650	Coppermine	100 MHz	$583
Pentium III-600EB	Coppermine	133 MHz	$455
Pentium III-600E	Coppermine	100 MHz	$455
Pentium III-600B	Katmai	133 MHz	$465
Pentium III-600	Katmai	100 MHz	$465
Pentium III-550E	Coppermine	100 MHz	$348
Pentium III-550	Katmai	100 MHz	$348
Pentium III-533E	Coppermine	133 MHz	$305
Pentium III-533	Katmai	133 MHz	$316
Pentium III-500	Katmai	100 MHz	$229
Pentium III-450	Katmai	100 MHz	$173
WORKSTATION/SERVER			
Pentium III Xeon-733	Coppermine	133 MHz	$826
Pentium III Xeon-667	Coppermine	133 MHz	$655
Pentium III Xeon-600	Coppermine	133 MHz	$505
MOBILE			
Mobile Pentium III-500	Coppermine	100 MHz	$530
Mobile Pentium III-450	Coppermine	100 MHz	$348
Mobile Pentium III-400LV	Coppermine	100 MHz	$348

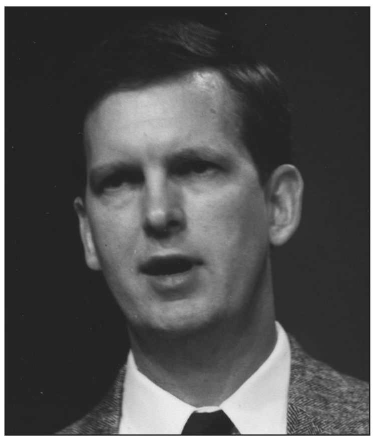
Table 1. Current 1,000-piece list pricing for new Coppermine-based processors (purple) and older Pentium III processors. (Source: Intel)

previously planning to hold off on the 733 speed grade until next year. We believe the pull-in is a response to AMD's aggressive Athlon introduction, trumping the Athlon-700, which AMD introduced a few weeks ago (see MPR 10/6/99, p. 4). The high price, however, will ensure that there are relatively few orders for the flagship speed grade.

The pull-in may have its costs. We expect Coppermine to reach at least 800 MHz in 1H00 and 866 MHz in 2H00, but these increases will probably not match Intel's usual pace of one new speed grade per quarter. On the other hand, today's announcement represents a huge leap forward in performance—essentially three speed grades at once—that should satisfy the market for a while.

Mobile Line Gains SSE, Faster Bus

Coppermine is a boon to the mobile space, raising performance through a combination of higher CPU speeds, faster L2 cache, faster bus speeds, and Pentium III's SSE extensions. Previously, the high end of the mobile line was a 400-MHz Mobile Pentium II (Dixon) with 256K of integrated L2 cache but no SSE. Intel is now offering Mobile Coppermine at 400, 450, and 500 MHz with a 100-MHz bus.


MICHAEL MUSTACCHI

Jim Wilson, architecture manager for Coppermine, shows off that chip's performance gains at the Forum.

The 0.18-micron process, which is designed to operate as low as 1.1 V, reduces power dissipation considerably from the older 0.25-micron parts, allowing Intel to offer much higher clock speeds within the standard thermal envelope. Intel could have used P858 to extend Pentium II's clock speed beyond 400 MHz but chose to wait for Coppermine to avoid overlap with the new Mobile Pentium III line.

The 400-MHz Coppermine is a low-voltage design that operates at 1.35 V and uses just 7 W (TDP). This is the only Coppermine speed suited to emerging mininotebook systems, which cannot dissipate as much heat as standard notebooks.

The faster Mobile Pentium IIIs operate at 1.6 V, only slightly lower than the desktop Coppermines at 1.65 V. At 1.6 V, the Mobile Pentium III-500 dissipates 11.2 W (TDP), about the same as a 0.25-micron Dixon at 400 MHz. The faster bus will add a bit more power to other devices that connect to it, but the big boost in performance should more than make up for it. With the 500-MHz Coppermine already at the power limit for standard notebooks, however, Intel will have to find a way to yield faster parts at a lower voltage to extend the Mobile Coppermine line.

The mobile products are available in BGA-495 and microPGA packages that are about 20% smaller than the equivalent Mobile Pentium II packages (see MPR 6/21/99, p. 4). For an extra $49, the 450 and 500 are also available in the MMC2 mobile module with AGP.

Coppermine gives Intel some much-needed differentiation between its high-end mobile processors and its low-cost Mobile Celeron line. Earlier this year, these two lines contained exactly the same speed grades using the same processor core at the same power levels. The only distinction was a barely noticeable performance drop due to Celeron's smaller L2 cache. With the Celeron parts listing for less than half the price of the Mobile Pentium II chips, the potential for cannibalization has been strong.

The new Mobile Pentium III offers all of the advantages noted above—more powerful CPU, faster bus, SSE, and lower power—while the current Mobile Celeron does not, topping out at 466 MHz with a 66-MHz bus. Furthermore, the Celeron parts continue to use the 0.25-micron process, pushing their power as high as 15.6 W (TDP). For oversized notebooks with bigger fans, this power level is acceptable, but OEMs will seek out the more expensive Pentium III for their premium thin-and-light products.

At some point, Intel will need to move Mobile Celeron to the 0.18-micron process to deliver further speed increases. Even then, the company can manipulate the voltage to keep the power level of the Celeron parts higher. Intel also does not seem to be in a rush to introduce SSE or the 100-MHz bus to the Celeron line, retaining those features as differentiators for some time.

SpeedStep Aids Mobile—Soon

Even the 500-MHz Mobile Pentium III is far behind the desktop Pentium III-733 in performance. Intel will reduce this gap with a new technology called SpeedStep, formerly known as Geyserville. This design allows the processor to operate in two modes: a low-power mode for battery operation and a high-power mode when the system is plugged in. Notebook systems are frequently used near a wall socket, and SpeedStep allows the processor to operate at a higher clock speed when AC power is available. Theoretically, a mobile Coppermine should be able to match desktop performance in the high-power mode.

The SpeedStep "technology" is actually fairly simple. Because the bus speed remains constant in both modes, the chip simply switches to a higher bus ratio in the high-power mode. To switch modes, the system logic puts the processor into a deep-sleep state, with the PLL stopped, while it changes the CPU's voltage. The system logic then restarts the CPU at the new voltage, and the processor powers up with the new bus ratio. Presumably, Intel's future mobile chip sets will automatically perform this transition.

Other than adding the variable bus ratio, Intel's challenge was to develop a manufacturing process to test the chip at two different voltage/speed combinations. This extra testing raises cost slightly, but the value to the end user is well worth it. We expect SpeedStep processors to run two or three speed grades (100–150 MHz) faster in the high-power mode. Even in the faster mode, however, the mobile processors are likely to lag slightly behind Intel's fastest desktop chip at any given time, both in CPU speed and in bus speed.

Unfortunately, the initial Mobile Pentium III processors lack SpeedStep. The company plans to ship SpeedStep parts early next year. We expect the company to offer SpeedStep in only the Pentium III line, not the Mobile Celeron line, creating further differentiation between the two.

Help for Workstations Now, Servers Later

Intel also rolled the new Coppermine into its Pentium III Xeon line, matching the desktop PC speeds of up to 733 MHz. These parts replace the current Xeons that have a 512K full-speed cache; these parts are used in most x86-based workstations. The faster CPU speeds, faster bus, reduced cache latency, and improved system bus are all major benefits for these workstations, as shown by the impressive SPEC scores posted by the new Pentium III.

With its integrated 256K L2 cache, the new Coppermine Xeon, code-named Cascades, greatly reduces manufacturing cost from the previous Xeons, which use expensive full-speed cache chips to deliver better performance than the non-Xeon Pentium III. Unfortunately, this strategy also reduces differentiation between the two product lines: the new Pentium III Xeon is identical to the new Pentium III, except that the Xeon part is packaged in a Slot 2 module for compatibility with current Xeon systems.

This compatibility is actually unnecessary; the new parts require new motherboards that support the faster bus. In fact, these processors will typically use the new 840 chip set (see MPR 10/25/99, p. 28) instead of the 440BX or 440GX used in earlier systems. Many OEMs will design these systems to use the Slot 1 Coppermine, which delivers the same performance, supports dual-processor systems, and costs $50 less than the Xeon parts, as Table 1 shows.

One possible advantage to staying with Slot 2 is that it includes a voltage regulator on the module, providing tighter voltage tolerances at the CPU. This factor, coupled with the greater thermal area of the larger module, could allow Intel to offer slightly higher clock speeds in Slot 2 than in Slot 1. Some workstation makers will pay $50 per system to be ready for a potential future speed boost.

The new Xeon can be used in small servers with only one or two processors, but larger servers can't effectively use Coppermine because of its small cache. These systems typically require processors with up to 2M of full-speed cache. The larger cache is particularly important because four-processor servers can't use the 133-MHz bus and will remain limited to 100 MHz through the remainder of Pentium III's lifetime. The larger cache reduces bus traffic, which is critical with the slower bus.

Intel is developing a large-cache version of Cascades for just this purpose. That chip will have up to 2M of on-die L2 cache, but it is not slated to be available until 2Q00. Until then, large Xeon servers will muddle along with the current 550-MHz Katmai parts, which offer solid performance today but will be long in the tooth by next spring.

Price & Availability

Intel's Pentium III and Pentium III Xeon/256K are available now at clock speeds of up to 733 MHz. Mobile Pentium III is also available now at speeds of up to 500 MHz. See Table 1 for pricing. For more information, access *http://developer.intel.com/design/pentiumiii/.*

Improvements in All Dimensions

Wilson's presentation left no doubt that Coppermine is an impressive design achievement, bringing together the fast Pentium III core with a low-latency on-die cache and an advanced 0.18-micron IC process. The combination of the faster cache and faster bus boosts performance by at least a full speed grade, compared with the original Katmai Pentium III, and the 0.18-micron process delivers higher CPU speeds on top of that increase. Beating the integer performance of the fastest RISC processor is outstanding, and regaining the x86 lead from AMD was imperative. Boosting performance by three full speed grades has enabled Coppermine to grab the performance lead.

The new process also provides cost and power advantages, the latter being most helpful in the mobile market. The forthcoming SpeedStep parts will give mobile users new options for higher performance when using AC power, bringing notebook systems closer to parity with desktop PCs. The chip's modest die size and integrated cache significantly reduce Intel's manufacturing cost, boosting margins, and will enable Intel to ramp production of the new design in record time.

Coppermine creates some transition issues for PC makers, namely the switch to the 133-MHz bus and the eventual move from Slot 1 to Socket 370. These issues are exacerbated by the delays that resulted in the awkward timing of Coppermine's introduction. Until OEMs are ready to adopt a new platform, Intel's product line will be overly complex. But within six months, these transition issues will ease. Unfortunately, Coppermine does not immediately help either the high-end server or the low-end Celeron segments. These segments should see the benefits of the new design by the middle of next year.

Facing surprisingly stiff competition from AMD's Athlon, Intel now has a new weapon that offers much lower manufacturing cost while erasing Athlon's nascent performance edge. The game is far from over, but Coppermine puts AMD back on the defensive. While watching this battle play out, PC users can sit back and enjoy Coppermine's strong performance. M