Pentium Competitors Go Head to Head

AMD, Cyrix Prepare to Join NexGen in Battle Against Intel

by Linley Gwennap

The time has come for PC makers to consider their options. Until recently, the only alternative to Intel's Pentium was a 486. But with both Cyrix's M1 and AMD's K5 now sampling into systems due late this year, system vendors have a choice: stick with Intel or buy a Pentium alternative. Not to be forgotten in this fray is NexGen, which was the first to debut a Pentium competitor and plans to rapidly increase the performance of its chips over the next several months.

According to their vendors' claims, these three Pentium competitors will eventually match the performance of Intel's fastest Pentium chips. Although all carry higher manufacturing costs than Pentium, we expect them to offer significantly better price/performance than Pentium at all but the lowest performance points, taking advantage of Intel's high margins. A single motherboard can support Pentium, the K5, and the M1, giving system vendors flexibility in selecting a processor.

Neither AMD nor Cyrix has published benchmarks for its processor, leaving their ability to meet their performance goals unproven. NexGen has benchmarked its current systems, but these results have been disputed by published reviews. NexGen is also hampered by its system interface, which is different than Pentium's. Based on an examination of these chips' designs, we can make some predictions about their ability to meet their goals and penetrate the market.

Improving on Pentium's Efficiency

With the dual-pipeline design shown in Figure 1(a), Pentium introduced the concept of superscalar execution
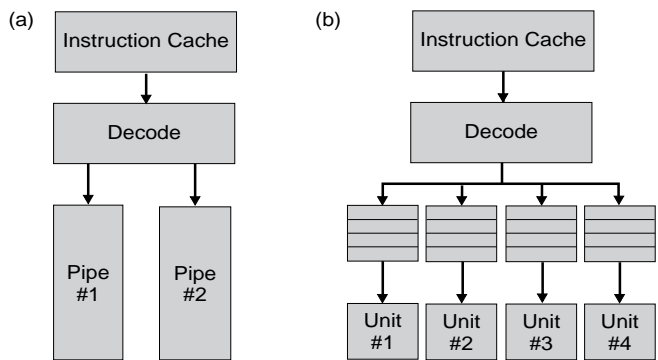

Figure 1. A dual-pipeline design (a) issues instructions to pipelined execution units, with little or no opportunity for reordering. A decoupled design (b) issues instructions to multiple queues, allowing more extensive reordering.

to the x86 world *(see 070402.PDF)*. At first, analysts were so amazed that the bear could dance that most didn't critique the dance. It became apparent, however, that the Pentium design is not very efficient, as many situations stall one or both pipelines.

Some pairs of instructions that access the same register must be processed one at a time to avoid conflicts, a problem exacerbated by the small x86 register set. The small number of registers also increases the number of memory references; unfortunately, instructions that reference memory take extra cycles to execute in Pentium. Floating-point instructions cannot be paired with integer instructions. Finally, any instruction that stalls one pipeline stalls the entire processor.

Many of these problems can be avoided by carefully arranging instructions, a task that Intel's Pentium compilers perform admirably. But most applications available today are compiled for the 486 (or even 386) processors that form the bulk of the PC installed base. These programs typically run 10–20% slower than recompiled programs *(see 070401.PDF)*.

Starting with a similar dual-pipeline design, Cyrix's M1 *(see 071401.PDF)* adds several features to reduce the bottlenecks in Pentium. Register renaming avoids stalls due to register reuse. Extra pipeline stages allow instructions that contain memory references to flow smoothly through the pipeline without delays. Out-of-order execution lets one pipeline continue even if the other is stalled.

For recompiled code, this feature set provides little advantage over Pentium's, perhaps 10% or less in integer performance. The major benefit is that recompilation is not required to take advantage of the M1's added features. Assuming that the M1 will perform nearly as well on older code as with recompiled code, the Cyrix design could deliver perhaps 15–30% better performance on older code than Pentium at the same clock speed. These figures assume a small baseline improvement plus the 10–20% Pentium penalty noted above.

Cyrix claims the M1's performance is 35% better than Pentium's, a figure at the high end of our estimate. Being more conservative, we expect the benefit to be closer to 20%. This advantage will be diminished on Pentium-optimized programs.

Instruction Translation Improves Reordering

Both NexGen's 586 and AMD's K5 use a different architectural approach, instruction translation, that Intel has also adopted for its P6 processor. These chips

convert awkward x86 instructions into more RISC-like operations. A single x86 instruction turns into one or more of these reduced operations, which NexGen calls RISC86 instructions and which AMD calls ROPs *(see 081401.PDF)*.

As Figure 1(b) shows, these ROPs are dispatched to their respective execution units, where they are queued until the necessary data and function units are available. Because the queues are decoupled, ROPs can execute out of program order. The processor then "retires" instructions in their proper order to ensure correct execution.

The M1 implements restricted reordering. If a single instruction stalls, subsequent instructions continue executing in the other pipeline. If a second instruction requires data from the stalled instruction or is itself stalled for any reason, no new instructions can be executed. Even when the second pipeline continues, the M1 is reduced to only a single pipeline.

The decoupled designs of the K5 and Nx586 allow more extensive instruction reordering. In the K5, up to 16 ROPs can be pending at once. If the first few ROPs in that group stall, the remainder can be executed speculatively, provided that they use different function units than the stalled instructions. Similarly, the Nx586 can handle up to 14 of its RISC86 instructions at once. Both chips provide dual integer units to help avoid stalls. This reordering ability allows the two chips to continue processing instructions in situations where the M1 (or Pentium) would stall.

As Table 1 shows, the Nx586 and K5 both implement register renaming, like the M1. In addition to the 8 software-visible registers, the M1 includes 24 extra registers, while the K5 has 16 renamable registers and the NexGen chip just 14. This may give the M1 a slight advantage in some situations.

Since the K5 implements essentially all the microarchitecture features of the M1 and adds more extensive reordering, we believe that the AMD chip will have a small advantage in performance at the same clock rate. AMD claims a 30% advantage over Pentium at the same clock speed, a figure we believe is achievable.

NexGen Decode Is a Bottleneck

The Nx586 could probably match the clock-for-clock performance of the K5 if not for a major bottleneck in the design: it can decode only a single x86 instruction per cycle, translating it into one or more RISC86 operations *(see 081403.PDF)*. The K5, in contrast, can decode up to four x86 instructions per cycle if each maps to a single ROP. Since some instructions usually require more than one ROP, the average decode rate for the K5 is two or

		Intel Pentium	Cyrix M1	AMD K5	NexGen Nx586	Intel P6
x86 dispatch rate		2 instr	2 instr	2–3 instr	1 instr	3 instr
Internal execution rate		2 instr	2 instr	4 instr	4 instr	5 instr
Pipeline stages		5 stages	7 stages	6 stages	7–9 stages	12 stages
Instruction translation?		no	no	yes	yes	yes
Out of order window		none	limited	16 instr	14 instr	40 instr
Rename registers		none	24 regs	16 regs	14 regs	40 regs
Branch history table		256 × 2	256 × 2	1,024 × 1	128×16×2	512×4×*
Return address buffer		none	8 entries	none	8 entries	16 entries
FPU location		on-chip	on-chip	on-chip	external	on-chip
FP multiply latency		3 cycles	4–9 cyc	7 cycles	2 cycles	5 cycles
Cache size (I/D)		8K/8K	16K	16K/8K	16K/16K	8K/8K
L2 cache bus?		no	no	no	yes	yes
Pentium pinout?		yes	yes	yes	no	no
APIC on chip?		yes	no	no	no	yes
Burst order		Intel	linear	Intel	linear	Intel
SMM software model		Intel	Cyrix	Intel	NexGen	Intel

Table 1. Intel's competitors have improved upon Pentium's microarchitecture in various ways, but none delivers the full feature set of the P6. AMD's system interface is most like Pentium's, while NexGen uses a completely different bus design. Rename registers exclude the eight needed to hold the architected x86 register state. *Intel has not fully specified P6's branch table. (Source: vendors)

three x86 instructions per cycle, still much better than the Nx586.

In this regard, the NexGen part is inferior even to Pentium, which can decode two x86 instructions per cycle. The Nx586 and Pentium cores achieve similar performance in very different ways. Pentium peaks at two instructions per cycle but spends many cycles completing one or even zero instructions, resulting in an average of about one instruction per cycle. The Nx586 issues no more than one instruction per cycle but makes heroic efforts to keep the pipeline operating at its peak rate.

While NexGen's Winstone testing indicates that its Nx586 delivers clock-for-clock-performance similar to Pentium's, some independent studies have failed to confirm these results, instead showing as much as a 20% degradation. NexGen claims that these lower results are caused by the use of nonoptimized VL-bus drivers, and that this problem will disappear when PCI systems begin shipping in the next few months.

It is fair to criticize NexGen's system vendors for not delivering optimized drivers to their customers. But this mistake does not disprove the underlying performance of the CPU core. If NexGen's claims are valid, we expect to see significantly better performance from second-generation Nx586 systems—performance that matches up well, clock for clock, with Pentium's.

Predecoding Speeds K5 Decode

The variable length of x86 instructions makes it difficult to determine the starting point of the second instruction in a sequence before decoding the first. This problem helped convince NexGen to stick with a single x86 decoder. Most superscalar x86 processors—including Pentium, the M1, and the P6—partially decode the

first instruction before starting the second, but this method adds time to the decoding process.

The K5 is unique among x86 processors in predecoding instructions as they are loaded into the cache. When the instructions are later fetched, the predecode bits tell the decoders where the next instruction begins. This design allows up to four instructions to be decoded in a single cycle without extending the decoding time, and it is extensible to more instructions. One problem is that the instruction cache must be expanded to store the extra information; in the K5, this overhead enlarges the instruction cache's die area by about 50%. The predecoding also adds to the instruction-cache miss penalty, but this increase should have a negligible performance impact on most applications.

Performance Depends on Clock Speed

Clock-for-clock comparisons are interesting, but actual performance depends on clock speed as well. To match Pentium's performance, the M1 and K5 must get to within 20–30% of the Intel chip's clock speed. The Nx586, with its decode bottleneck, must come even closer to Pentium's clock speed.

All things being equal, clock speed is mainly a function of pipeline design. In this regard, the Cyrix and NexGen designs have an advantage, as they use deeper pipelines than the K5 or Pentium. The M1 spreads the work of instruction processing across seven stages, compared with five for Pentium. Specifically, the complex task of decoding and dispatching two x86 instructions per cycle is spread across two full stages in the M1, while Pentium crams most of this work into a single stage. The M1 also provides two stages for segmentation checks, which Pentium completes in one stage.

The Nx586 pipeline is more complex, allocating from seven to nine stages depending on the task. Like the M1, it allocates two stages each to instruction decoding and address calculation, including segmentation checks. Other added stages deal mainly with the overhead of instruction translation and reordering.

Assuming that decoding or address calculations limit the speed of Pentium, the M1 and Nx586 should achieve faster clock rates. Indeed, in similar 0.5-micron processes, both the M1 and the Nx586 should reach 133–140 MHz, while the Pentium reaches 120 MHz.

The K5 pipeline is more similar to Pentium's. It consists of six stages, but the final stage is for retiring instructions, which Pentium doesn't need to do. Another difference is the K5's extra stage to generate ROPs.

To compensate for this extra decode stage, AMD chose to combine address calculation and data-cache access into a single stage. No other pipelined x86 processor combines these tasks; the Nx586 allows three stages for them. The K5 design goes through great pains to speed this path, but we expect that the chip will not match Pentium's clock frequency in comparable processes—although AMD disagrees. The company hopes to get the K5 to 150 MHz in a 0.35-micron process, whereas Intel expects Pentium to reach 150 MHz and beyond in its 0.35-micron fabs.

With clock speed, all things are never equal. Intel has been more aggressive than its competitors in process development, bringing a 0.35-micron process into production in 2Q95, six to nine months ahead of AMD and even further ahead of IBM, which builds the M1 and Nx586. Intel also has the time and resources to carefully compress its designs, tuning speed paths to achieve maximum performance. In short, Intel's advantages prevent its competitors from realizing the full clock-speed edge present in their designs.

NexGen Excels at Branch Prediction

The bugaboo of long pipelines is branching. The more stages in the pipeline, the more instructions that must be invalidated on a mispredicted branch. For the M1 and Nx586 in particular, mispredictions must be minimized to achieve good performance.

Both Cyrix and NexGen paid extra attention to this problem. The M1 implements a 256-entry branch target buffer (BTB) similar to Pentium's, to which Cyrix adds a 256-entry not-taken-branch table and an eight-entry return-address stack. This stack predicts subroutine returns, which are otherwise difficult to predict because they often have different targets on successive iterations. Intel estimates that Pentium's BTB delivers about 80% accuracy on the SPECint92 suite; with its added features, the M1 should push that figure to perhaps 85%.

The M1 also has a four-set prefetch buffer that holds instructions. On a mispredicted branch, this buffer usually contains the correct path as well as the predicted path, reducing the mispredicted branch penalty by one cycle on a hit.

The Nx586 goes beyond any other Pentium-class chip by implementing two-level branch prediction similar to that in the P6. In fact, the Nx586 was the first chip to implement this advanced algorithm, although NexGen only recently revealed this innovation. The Nx586 implements the GAs method, according to Yeh and Patt's taxonomy *(see **090405.PDF**)*, using a 7-bit global history register to index into a 2,048-entry BTB. This BTB consists of 16 sets, which are selected using four bits from the program counter. As in Pentium, each entry contains two bits of history plus a predicted target.

In total, the Nx586 has far more BTB entries than either the M1 or Pentium. The two-level arrangement is known to produce better predictions than a single-level structure of the same size. The Nx586 also includes an eight-entry return stack like the M1's. This combination of features should give the NexGen design better prediction accuracy than any of its peers; the company

estimates SPECint92 accuracy at 92%, cutting the number of mispredictions in half compared with the M1.

The K5, with its shorter pipeline, has devoted less circuitry to branch prediction. It relies on a 1,024-entry BTB, but each entry has only one history bit instead of the two bits used by its competitors. Also, because it shares its tags with the cache, the K5's BTB can store only one prediction per cache line. Although this BTB has more entries than Intel's or Cyrix's, its prediction accuracy should be about the same as Pentium's. Without a return stack, the K5 will have a lower prediction accuracy than the M1.

Competitors Downplay Floating Point

Previous Intel products had unimpressive floating-point performance, but Pentium devotes significant die area to the FPU. For example, its FP multiplier can generate a result in three cycles, five times faster than the 486. To resolve problems with the x86's FP register stack, Pentium allows parallel execution of an FP operation and an FXCH instruction. These improvements have helped Pentium in the workstation market and also benefit some 3D graphics and multimedia applications.

Intel's competitors have chosen to skimp on FP performance, targeting the mainstream PC market. The K5 has a seven-cycle multiplier, less than half the speed of Pentium's, although it does support parallel FXCH execution. The M1 takes four to nine cycles for an FP multiply and does not support the parallel FXCH. Both the M1 and K5 can pair FP instructions with integer instructions, which Pentium cannot, but we expect both chips to have lower performance on FP benchmarks than Pentium.

The Nx586 has no FPU at all: its partitioning puts the FPU on a separate chip. NexGen chose to defer work on the FPU chip, focusing on getting its CPU to market sooner. The company will combine the CPU and FPU chips in a single multichip module; this "DX" version is due this summer, but no pricing has been released.

Once this product is available, however, it should rival Pentium's FP performance. In fact, NexGen's FPU performs most FP operations, including multiplies, in two cycles, one fewer than Pentium. One drawback is that FP adds and subtracts are not pipelined, reducing throughput compared with Pentium. The greater transistor count allowed by NexGen's two-chip design enables it to devote more transistors to floating-point math.

Cache Design Favors Nx586

Although the biggest differences among these designs are in the CPU cores, all use slightly different cache designs. The Pentium design is the most basic, with 8K of instruction cache and 8K of data cache, both on-chip. Like the 486, Pentium uses a single external bus to connect to the rest of the system, including secondary cache, main memory, and I/O. This 64-bit bus,

which is limited to 66 MHz, can be a performance bottleneck, particularly as CPU speeds increase.

For compatibility, AMD retained the single-bus design but increased the size of the K5's on-chip instruction cache to 16K, not including the predecode information. This increase, coupled with a smaller line size, increases the hit rate and reduces traffic on the system bus. Unlike its competitors, the K5 uses virtual cache tags to speed accesses. A virtually tagged cache has a lower hit rate than a physical cache, but AMD included a set of physical tags on the K5 as well. These additional tags maintain the hit rate of a physical cache but add die area.

Cyrix also wanted to reduce the load on the system bus, but without adding significantly to the die area. The M1 contains 16K of on-chip cache, the same amount as Pentium, but keeps it in a single unified structure, giving it a higher hit rate than Pentium's split-cache design.

The downside of a unified cache is its inability to supply instructions and data in the same cycle. To get around this problem, the M1 features a small, 256-byte instruction buffer that it accesses in parallel with the main cache. A prefetch engine tries to fill the instruction buffer whenever the main cache is not occupied with data accesses. This design echoes the 486's cache, which also is unified and uses a prefetch buffer.

The Nx586 includes two 16K caches, more cache than any other Pentium-class chip. In addition, NexGen decouples the L2 cache from the system bus, improving performance in two ways. First, the L2 cache bus can operate at speeds greater than 66 MHz, increasing bandwidth to the external cache. Second, it removes cache traffic from the system bus, allowing memory accesses to flow more smoothly. The Nx586's slight per-clock performance edge over Pentium stems from these features.

The data caches in the K5, Nx586, and Pentium all support two accesses per cycle, as does the M1's unified cache. The Nx586, however, has just one load/store unit, so it can execute only one load or store per cycle. In that chip, the second cache access is used for snooping or cache reloads, keeping these activities from blocking the CPU. The other processors can all complete two loads per cycle, increasing throughput on programs with frequent memory references.

K5, M1, Pentium Are Socket Compatible

Both Cyrix and AMD have aimed at making their chips fully compatible with the P54C Pentium at both the hardware and software level. For the most part, they appear to have succeeded, but a few minor differences remain, particularly with the M1.

All three chips use essentially the same pinout, but neither the M1 nor the K5 includes the signals associated with Intel's APIC. This lack is not a problem for uniprocessor systems, which don't use the APIC. For multiprocessor systems, both AMD and Cyrix have endorsed the OpenPIC standard *(see 0905MSB.PDF)*. This choice is due to fears that Intel has system-level patents pending that the company could assert against system makers that use non-Intel APICs.

For similar reasons, Cyrix does not support Intel's patented nonlinear burst order on the M1's system bus, substituting a linear burst order. This change has little or no performance impact but limits designers to system-logic chip sets that support linear bursting. Fortunately, most new chip sets do, with the exception of Intel's.

For system makers that insist on Intel chip sets (or others that don't support linear bursting), the M1 includes a compatibility mode that uses two bus accesses to simulate an Intel-ordered access *(see 081601.PDF)*. This mode, however, carries a 5–10% performance penalty, reducing the M1's performance edge.

The K5 avoids this problem by matching Intel's burst order, but this decision opens the possibility of Intel legal action against AMD's customers. AMD does not expect Intel to pursue this issue in the courts.

Both the K5 and M1 are compatible with Intel's power-management signals, such as STPCLK and SMI. At the software level, the M1's system-management mode (SMM) differs from Intel's, retaining compatibility with earlier Cyrix designs, while the K5 is compatible with Pentium. Differences in SMM are handled by the BIOS and are not visible to operating systems or applications.

Another software-visible deviation concerns Pentium's "secret" Appendix H, which details a few extensions to the x86 architecture, including block address translation and improvements to virtual 8086 mode. Cyrix chose not to implement the Appendix H features, although it does provide a noncompatible method for block address translation. AMD, on the other hand, claims to have reverse-engineered Appendix H and implemented those features in the K5. Since these features are mainly for operating-system use, and Microsoft says that it has not used them in any of its software, this difference is probably moot.

Nx586 Requires Different Motherboard

The NexGen processor implements a completely different system interface than its competitors, as its design was firm well before Intel published its Pentium specifications. As a result, the chip is not compatible with existing Pentium motherboard designs or chip sets. The Nx586 is compatible with but a single chip set, NexGen's own product that is based on VL-bus. The company expects to deploy a PCI chip set, which will also be sold by VLSI, this summer.

Although it is different, the NexGen motherboard is not inherently more or less expensive than a Pentium motherboard. The biggest issues concern features and performance. With more chip sets to choose from, system vendors can design Pentium (and M1 or K5) products for a range of features and cost points. Also, at any given time, Pentium chip sets may deliver better system performance than NexGen's. These issues restrict the competitiveness of NexGen's design.

For system makers that buy motherboards instead of processors, there is less difference between using the NexGen part or Pentium. A few companies supply complete Nx586 motherboards that are compatible with Pentium motherboards. The major difference today is the lack of PCI support, but this will change shortly. The many PC companies that purchase motherboards from third parties are thus good candidates for the Nx586.

Extended Roll-Outs Expected

All three Intel competitors are using a similar roll-out strategy: first get a working product to market, then shrink the die to improve speed and cost. Cyrix has taken this strategy to the extreme: the initial M1 die is 394 mm^2, more than twice the size of a P54C Pentium die. Despite its extended pipeline, it runs at 100 MHz, the same speed as the P54C. This version is built in a hybrid process that combines 0.65-micron transistors with the three metal layers of IBM's 0.8-micron process.

The M1 design is finalized, and wafers of these initial parts are in production, with shipments expected next month. Quantities of these chips will be limited, however, and most vendors will use them as samples for developing systems. Cyrix is already nearing completion on a redesign to IBM's 0.5-micron, five-layer-metal CMOS-5S process; this version should cut the die size to less than 225 mm^2 while boosting speeds to 120 MHz. The company says this shrink will be in production in 4Q95, in time for initial system shipments.

An optical shrink to IBM's CMOS-5X process is

expected to push the clock speed to 133 MHz; the target for this version is a die size of 169 mm^2, still somewhat larger than the 143-mm^2 P54C. Because the optical shrink requires no additional design work, Cyrix hopes that this version will reach production in 1Q96. The company expects yet another shrink by late 1996.

AMD will start in 4Q95 with a 0.5-micron version of the K5 that will probably run at just 75 MHz. While this version consumes 251 mm^2, the company plans a quick shrink to 0.35-micron in 1Q96, reducing the die size to 161 mm^2 and increasing the speed to 100 MHz. The 0.5-micron version will be used mainly for samples, although Compaq may ship some systems with this part. AMD's plans for volume production rest with the smaller die.

The company still claims that the K5 will achieve clock frequencies of 133 MHz "and beyond," but these faster parts will roll out over the course of 1996 as the company adjusts the layout and circuit design of its processor to speed critical paths. We find it difficult to believe that AMD can improve timing by 33% or more with such fine-tuning; process improvements may be required to achieve the faster speeds, delaying their introduction.

Another risk factor for AMD is its reliance on its new fab. Volume production for the K5 will come from Fab 25, which is now completing final qualification runs of 0.5-micron wafers. Fab 25 has run a few 0.35-micron test wafers, but AMD must qualify this new process in the new fab by Christmas to start wafers for 1Q96 shipments. Any snags could delay production or cause shipments to fluctuate through mid-1996.

NexGen is ahead in this game, as its Nx586 design is verified and in production at speeds up to 93 MHz. The company will soon deploy a new version of the chip that is optimized for IBM's 0.65-micron CMOS-5L process, shrinking the die to 118 mm^2 and increasing performance by about 20%. A shrink to CMOS-5S is expected to ship by the end of this year, further increasing performance and shaving a bit off the die size. By mid-1996, an optical shrink to the CMOS-5X process will bring additional improvements.

Figure 2 compares the roll-out plans of the Pentium-class chips. We project that Intel will keep Pentium performance ahead of the competition through at least mid-1996. If Cyrix and NexGen meet their goals, they will come close to Intel's pace; AMD's K5 lags Intel's performance curve by 6–9 months. The M1 and Nx586 may eventually exceed Pentium's performance when IBM gets 0.35-micron versions into production, but by that time, Intel will already be into the second generation of the P6, far surpassing their performance.

Pentium Holds Cost Advantage

Because of the great number of shrinks in these plans, a fair cost comparison is difficult. From a design standpoint, one could compare die sizes in similar man-

ufacturing processes. For example, at 0.5 micron, the K5 is by far the largest, at 251 mm^2, and the M1 is about 200 mm^2. Pentium is 143 mm^2, while the Nx586 is just 100 mm^2. These processes are not quite comparable because AMD uses just three metal layers, whereas Intel uses four and Cyrix and NexGen five. With the same number of metal layers, the K5 and M1 would be about the same size, while Pentium would be about a third smaller and the FPU-less Nx586 a bit smaller yet.

There is no apparent technical reason why the Cyrix and AMD chips should be so large. The M1 actually has fewer transistors than Pentium; although the K5 has about a million more transistors than Intel's chip, most are in the added cache memory, which adds little area to the die. For the most part, these differences in die size are due to Intel's focus on packing the transistors as densely as possible, maximizing fab output. With fewer engineers and more time-to-market pressure, AMD and Cyrix have not devoted as much attention to compacting their chips, although they may in the future.

The Nx586, on the other hand, has about 30% fewer logic transistors than Pentium, mainly due to its lack of an FPU. Some of the extra space is taken up by the additional 16K of cache, but the NexGen design should still be significantly smaller than Pentium. NexGen gains an additional size advantage by using IBM's flip-chip mounting, eliminating the pad ring. These advantages are partially offset by Intel's superior circuit packing.

The overall cost of the Nx586 is increased, however, by the larger package required by its two-bus design. The part uses a 463-pin PGA rather than the 296-pin package
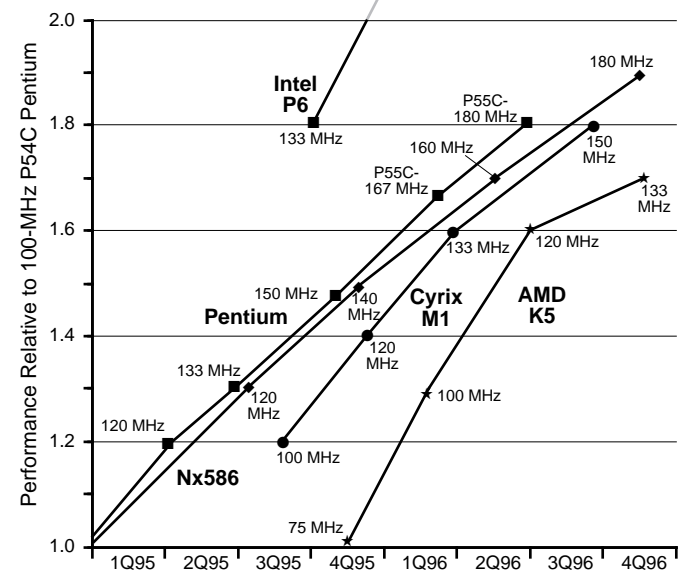
Figure 2. Pentium should retain a slight performance lead over its competitors while Intel's P6 sets new performance standards. Performance is for typical (unrecompiled) integer PC applications. (Source: MDR estimates based on vendor information)

	Intel Pentium	Cyrix M1	AMD K5	NexGen Nx586	Intel P6
Clock speed (1Q96)	167 MHz*	133 MHz	100 MHz	140 MHz*	133 MHz
FPU included?	yes	yes	yes	no	yes
IC process	0.35μ 4M BiCMOS	0.5μ 5M CMOS	0.35μ 3M CMOS	0.5μ 5M CMOS	0.5μ 4M BiCMOS
Logic transistors	2.4 million	2.1 million	2.4 million	1.6 million†	4.5 million‡
Total transistors	3.3 million	3.0 million	4.3 million	3.5 million†	5.5 million‡
Package type	296-pin CPGA	296-pin CPGA	296-pin CPGA	463-pin CPGA	387-pin MCM-C
Die size	90 mm²	169 mm²	161 mm²	100 mm²*†	306 mm²‡
Est mfg cost	$80*	$120*	$125*	$120*†	$350*‡
Per-clock perf**	same	+20%*	+30%	+7%	+35%
Overall perf**	1.6*	1.6*	1.3*	1.5*	1.8*

Table 2. Comparing processors expected in 1Q96, we see that Intel's competitors may come close to Pentium's performance but not its manufacturing cost, while the P6 outperforms them all. **Performance is relative to a 100-MHz Pentium on typical integer PC applications. †FPU chip not included ‡L2 cache chip not included (Source: vendors except *MDR estimates)

used by the other parts. NexGen's costs are further increased if one includes the FPU chip, which is needed to match the feature set of Pentium. This increase will be counterbalanced somewhat by a forthcoming reduction in pin count enabled by removing the external FPU bus from the package. Even so, we estimate that Pentium has a cost advantage even when all chips use similar manufacturing technology.

In the real world, things aren't that fair. Intel has moved to 0.35-micron technology ahead of its competitors. In this process, the Pentium die is 90 mm², far smaller than the K5 or M1. Table 2 shows the estimated manufacturing cost of Pentium-class chips projected for 1Q96. These projections assume that AMD moves K5 into its 0.35-micron process by that time, and that the CMOS-5X version of M1 is ready as well. We project that the competitors' chips will all cost at least 50% more to build than Pentium does.

For Intel's competitors, cost is not an overwhelming issue. If their parts perform as promised, they will sell for $300–$500, delivering solid profits even with a manufacturing cost of $120 or so. The low end of the Pentium market will be a bigger problem. MDR projects that the price of a 75-MHz Pentium will be well below $200 in 1996. Intel can deliver these parts profitably, but its competitors cannot. AMD will compete in this space with fast 486 processors *(see 0908MSB.PDF)*, while Cyrix will use the cost-reduced M1sc *(see 081601.PDF)*.

Competing with the P6 Could Be Tough

The high end of the x86 market will also be a problem for Intel's competitors. With the P6 rolling out this year, these competitors must already consider how they can compete with this next-generation device. None of the Pentium-class competitors can match the performance of the P6; these vendors will need to redesign their parts to compete in this space. Some have more work to do than others.

NexGen is the closest, and the company expects to deliver its P6-class processor before any other Intel competitor. The company has not released any details of its 686, but Table 1 showed that only a few changes are needed to bring the Nx586 to P6 standards. The biggest is the addition of more x86 instruction decoders, allowing the chip to process more than one x86 instruction per cycle. NexGen also needs to expand the reorder buffer to be comparable to the 40 entries in the P6, and add a second address unit to allow the CPU to generate two cache accesses per cycle, which the cache itself can already handle. These changes are relatively straightforward.

The resulting chip would have more on-chip cache than the P6 while retaining the local bus for the L2 cache. The Nx586 already offers branch prediction and out-of-order execution similar to the P6's. NexGen could even skip the complex P6 bus interface designed for MP servers, simplifying its design. One drawback would be the static ordering of NexGen's instruction queues; the P6's dynamic instruction dispatching is more flexible. NexGen says that its 686 will be in systems in 1H96.

AMD expects its K6 processor to reach volume production a year later, in 1H97. This schedule does not allow for a complete redesign of the K5. We expect the K6 CPU core to be much like the K5's, but with a larger reorder buffer and a few other minor changes. AMD must address the system interface, however. Either the K6 must significantly expand the on-chip caches, add a local L2 cache bus, or use a much faster system bus such as Intel's P6 bus or a similar design.

Cyrix has not announced plans for a P6-class processor. The company faces the biggest design effort, as its M1 does not implement the decoupled architecture used by the P6, K5, and Nx586. During 1996, Cyrix will probably boost M1 performance with a larger cache and/or a new bus interface, but these changes will not deliver P6-class performance. Sources indicate a complete redesign, code-named M3, is also in the works; this chip should be a true P6-class device but probably won't ship until 1997 or 1998, about the same time as Intel's P7.

Thus, it appears that Intel's competitors will remain a half or even a full generation behind for the foreseeable future, giving Intel sole claim to the premium price points in the x86 processor market. Its competitors, however, can make a profitable business selling mid-range processors. With projected sales of 40 million units in 1996 and 50 million in 1997, the Pentium market is big enough for these smaller companies to survive and even flourish. But in the end, Intel's advantages in design and manufacturing will allow it to retain an 80% share of the markets in which it is most interested. ♦