MICROPROCESSOR REPORT

THE INSIDERS' GUIDE TO MICROPROCESSOR HARDWARE

VOLUME 9 NUMBER 14 OCTOBER 23, 1995

Nx686 Goes Toe-to-Toe with Pentium Pro

NexGen Rolls Out First Competitor to Intel's High-End Chip

by Linley Gwennap

Launching an aggressive campaign to bring Pentium Pro performance to the masses before Intel does, NexGen unveiled its Nx686 processor in a surprise announcement at the recent Microprocessor Forum. The new chip includes most of the key features of Intel's Pentium Pro (P6) design but avoids that chip's problems on 16-bit code. Although Intel has yet to announce any processor with multimedia extensions, NexGen has jumped the gun with its own enhancements. The tiny company claims that its chip costs less to build than Pentium Pro as well.

NexGen's processor appears to be about 6–9 months behind Pentium Pro (PPro); Nx686 systems should ship around the middle of 1996. But while initial PPro processors will be restricted to servers and workstations, the upstart startup has immediately targeted the volume PC market for its product. According to NexGen architect Greg Favor, the 0.35-micron Nx686 will be smaller than a 0.35-micron PPro (which will ship initially in a 0.5-micron process), allowing the company to price its chip for mainstream PCs. Pentium Pro is not expected to reach these price points until 1997.

While based on the lessons of the Nx586, the new chip is a ground-up redesign that significantly improves the performance and the efficiency of NexGen's original product. If the company delivers on its promises, it will establish a clear performance lead over Intel's other x86 competitors. Despite its incompatible pinout, the 686 should significantly raise NexGen's market profile.

What Is a P6-Class Processor?

It is impossible to define a set of features that separates a Pentium-class processor from a P6-class device. But on any feature comparison with Pentium Pro, the Nx686 stacks up well. Both chips decode multiple x86 instructions per cycle and convert them to RISC-like operations (ROPs), which are executed in a decoupled, out-of-order core. Both devices support large windows for reordering instructions, and both reorder memory references as well. Both have a high-speed interface to a pipelined L2 cache along with a separate 528-Mbyte/s interface to main memory and the rest of the system. But other x86 processors, particularly AMD's K5, also bear a passable resemblance to Pentium Pro.

We classify x86 processors based on performance, not feature set. NexGen expects its initial 686 parts to reach 180 MHz. Because of its deeper pipeline, PPro should have a 10–20% better clock speed once it reaches the 0.35-micron level. The deeper pipeline, however, reduces efficiency by 10–20%: on 32-bit code, PPro is rated at about 1.35× Pentium on a clock-for-clock basis, while the Nx686 averages 1.6×, according to the vendors. Based on these estimates, NexGen's part should be in the same performance class as Pentium Pro.

On 16-bit code, the Nx686 should outperform Pentium Pro. NexGen has taken care to handle situations that sap 16-bit performance on PPro *(see **091001.PDF**)*. For example, the Nx686 handles misaligned loads with a one-cycle penalty, versus six cycles in PPro, and performs segment-register writes and partial register accesses efficiently, without draining the pipeline. As a result, NexGen estimates that its new chip will deliver, clock-for-clock, about 1.4× the performance of a Pentium on 16-bit code; PPro, in contrast, offers less 16-bit performance than a Pentium of the same clock speed.

The Nx686 should also beat Pentium Pro on multimedia applications. NexGen has added new instructions that perform parallel calculations on the 8- and 16-bit data typically used in audio and video algorithms. The company did not provide additional details, but claims that its extensions are comparable to Sun's VIS instruction set *(see **081604.PDF**)*. NexGen claims that these instructions will increase performance on certain multimedia applications by 2–4× compared with a non-enhanced processor such as Pentium Pro.

Intel is developing a similar set of instruction extensions, which it will add to its Pentium line next year and, we believe, to Pentium Pro in 1997. NexGen appears set to ship processors with multimedia extensions before Intel does, and certainly before the enhanced PPro chip. It is unlikely that NexGen's instructions will be compatible with Intel's, but they can be buried in drivers and other low-level software without modifying application code. NexGen says that it has several unnamed backers for its multimedia extensions, and the large increase in performance will help attract more interest.

One area where Pentium Pro holds a performance advantage is in bus throughput. Although both processors have a 66-MHz, 64-bit bus to main memory, the PPro bus is fully pipelined *(see **090701.PDF**)*; by overlapping requests, it can sustain its peak bandwidth for extended periods. Using a bus similar to its predecessor's, the Nx686 can perform only one memory access at a time, limiting its throughput. The NexGen bus should be adequate for uniprocessor PCs, the company's target market. The PPro bus is superior for servers and high-end desktop systems that use multiple processors or otherwise demand high memory throughput.

Breaking the Decode Bottleneck

Conceptually, the Nx686 is similar to the Nx586 *(see **080403.PDF**)*. The biggest changes are in the instruction decode and dispatch. Whereas the Nx586 decodes only one x86 instruction per cycle, limiting its performance, the 686 can decode up to two, doubling the issue bandwidth. The dispatch logic eliminates the strict queues of the 586 design in favor of a flexible instruction pool that can dispatch instructions in any order.

The Nx686 and Pentium Pro use different decode strategies, due in part to Intel's emphasis on 32-bit performance. The NexGen part can generate up to four ROPs (which NexGen calls RISC86 instructions) per cycle; if the first x86 instruction produces one or two ROPs, the second x86 instruction can also be decoded, provided that it also produces only one or two ROPs. If either instruction produces more than two ROPs, only the first can be decoded in that cycle.

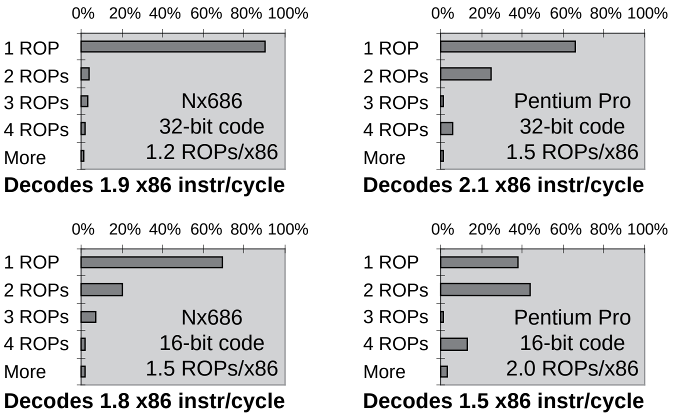

Figure 1. On typical 32-bit code, NexGen's 686 comes within 10% of the instruction-decode bandwidth of the Pentium Pro. The Nx686 does much better than PPro on 16-bit code. (Source: MDR estimates based on vendor data)

NexGen says that x86 instructions generate an average of 1.2 ROPs for 32-bit code, which tends to use only RISC-like x86 instructions; on older programs that run in 16-bit mode, the average increases to 1.5 ROPS per x86 instruction.

Pentium Pro can decode up to three x86 instructions per cycle, but while the first can generate up to four ROPs (which Intel calls uops), the second and third must be single-ROP instructions or they cannot be decoded in tandem. This method works best if most x86 instructions convert to a single ROP. Unfortunately, all writes to memory convert to two ROPs in PPro (but only one in the Nx686). This means that, even on 32-bit code, PPro generates 1.5 ROPs per x86 instruction; on 16-bit code, the average is 2.0 ROPs, according to Intel.

Figure 1 compares these two decoding strategies for both 16- and 32-bit code. One might think that the Intel chip would have much better instruction-decode bandwidth, due to its third decoder, but it does not. In the 32-bit case, the number of multiple-ROP instructions puts PPro at 2.1 x86 instructions per cycle, far from its peak. The Nx686, in contrast, comes close to its maximum of two instructions per cycle, achieving 1.9.

On 16-bit code, the Nx686 sees a little degradation, decoding 1.8 instructions per cycle, but PPro's strategy fares much worse, falling to 1.5. Thus, on 32-bit code the NexGen design comes within 10% of PPro's decode bandwidth while clearly surpassing it on 16-bit code.

The Nx586's single x86 decoder kept that processor from ever sustaining more than one instruction per cycle, limiting it to Pentium-class performance despite its complex RISC-like core. Two decoders are enough to unleash the performance of the Nx686's core. The designers looked at adding a third decoder but found that the higher decode bandwidth increased per-clock performance by only 2–5%, indicating that the current design is not limited by the decoders. The third decoder would have either extended the cycle time by 15–20% or added another pipeline stage, resulting in an overall decrease in performance.

Flexible Instruction Dispatch

The second major change is the dispatch logic. In the Nx586, if a function unit stalls (typically due to a data dependency), that unit is blocked from executing subsequent ROPs until the stall is resolved. (The K5 uses a similar strategy.) Thus, the function units are often not operating at their peak efficiency.

Like Pentium Pro, the Nx686 uses a more flexible dispatcher. All ROPs are held in a central holding tank, which NexGen calls the instruction control unit (ICU). Instructions are dispatched to function units only when they are ready to execute, allowing the function units to operate more efficiently. Thus, the Nx686 achieves higher performance than its predecessor despite having much the same selection of function units.

The ICU can store 24 ROPs, but only 12 x86 instructions (an average of 15–18 ROPs) can be active at any time. Pentium Pro's reservation station holds 20 ROPs, with as many as 40 active in its reorder buffer (ROB). PPro's ROB must be larger because of the chip's deeper pipeline and greater number of ROPs per instruction. We estimate that a typical stream of instructions without stalls will consume about 25 of 40 ROB entries, whereas the same stream will fill 9.5 of NexGen's 12 ICU entries. This comparison still leaves the PPro with more slots available to deal with stalls caused by L1 cache misses or long dependency chains.

Nx686 ROPs can access 48 physical registers, providing extensive register renaming. This is more than twice the size of the physical register file in the Nx586 and is comparable to Pentium Pro, which combines a 40-entry reorder buffer with 8 standard registers.

Function Units Have Low Latency

The new chip has two integer units, a floating-point unit, a multimedia unit, a load unit, and a store unit, as Figure 2 shows. Compared with its predecessor, the Nx686 adds the multimedia unit, which is used only for the new instructions, and separates the old memory unit into load and store units. In the older design, the FPU was physically on a second chip; the Nx686 integrates the FPU on the processor chip.

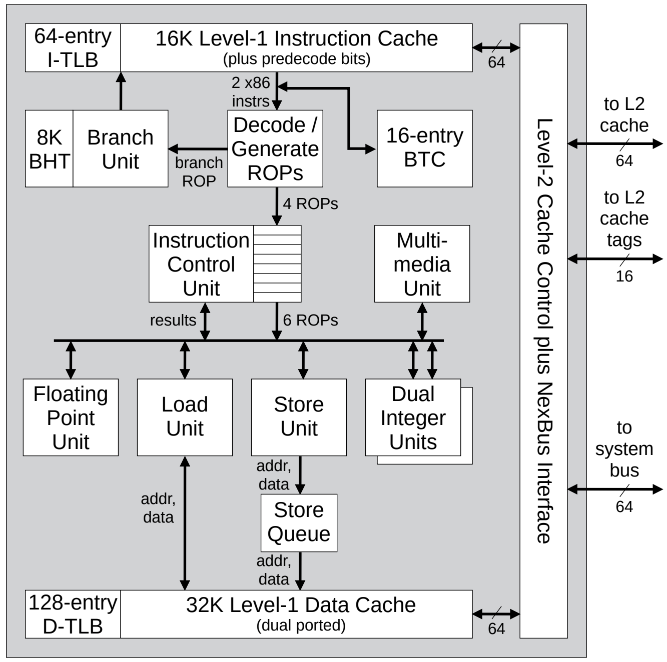

Figure 2. The Nx686 converts x86 instructions into ROPs that are executed out of order by seven function units (including the branch unit) under control of the ICU.

The Nx686's FPU is quite fast for an x86 chip, delivering lower latencies than PPro's, particularly for multiplication. The NexGen chip executes most FP operations (adds, compares, multiplies) in two cycles; data movement (loads, stores, FXCH) take one cycle. The Intel design needs three cycles for an FP add or compare, five for a multiply. As Table 1 shows, PPro has one advantage: single-cycle throughput for adds. The Nx686 is not pipelined for any floating-point operation.

	Nx686		Pentium Pro	
	Throughput	Latency	Throughput	Latency
Integer multiply	2 cycles	2 cycles	1 cycle	4 cycles
Integer divide	23 cycles	23 cycles	12–36 cyc	12–36 cyc
FP load	1 cycle	1 cycle	1 cycle	2 cycles
FXCH	1 cycle	1 cycle	1 cycle	1 cycle
FP add	2 cycles	2 cycles	1 cycle	3 cycles
FP multiply	2 cycles	2 cycles	2 cycle	5 cycles
FP divide	16–36 cyc	16–36 cyc	18–38 cyc	18–38 cyc
FP sq root	16–36 cyc	16–36 cyc	29–69 cyc	29–69 cyc

Table 1. The Nx686 has low-latency integer and floating-point operations, but the Pentium Pro has better throughput on some operations. (Source: vendors)

Although Pentium and other superscalar x86 microprocessors can execute two loads in a single cycle, the Nx686 (like Pentium Pro) is limited to one load and one store per cycle. On typical PC applications (e.g., Word), x86 instructions generate 0.5–0.6 loads each. Thus, in many cases a simple superscalar processor like Pentium could not achieve its peak rate of two instructions per cycle with a single load unit.

Out-of-order processors like the Nx686 can reorder loads, so the important issue is the sustained average load rate. At 0.6 loads per x86 instruction, a single load unit can sustain 1.7 x86 instructions per cycle, which is about the best speed of the Nx686 on PC applications.

Short Pipeline Simplifies Design

The NexGen chip uses a six-stage basic pipeline, as Figure 3 shows. It decodes two instructions in a single clock cycle; the decode time is reduced by predecoding instructions as they enter the instruction cache, a technique used by the K5 and several RISC processors.

NexGen claims to have done better than these competitive products. Other processors require an extra cycle to predecode instructions, extending the cache-miss penalty; NexGen's Favor says his design avoids this extra cycle. Due to pending patent applications, he would not offer details. Favor also would not reveal the exact number of predecode bits but hinted that it is about three per byte, less than the five bits per byte required by AMD's K5.

Decoded instructions and their ROPs are stored in the ICU. During the third pipeline stage, the ICU checks dependencies for all ROPs; those that have their data available are dispatched to the appropriate function

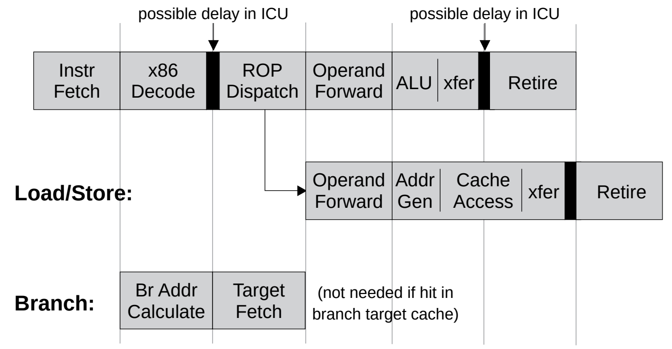

Figure 3. The Nx686 uses a six-stage pipeline with an extra stage for loads and stores. Branches are handled early.

unit. The ICU can dispatch one ROP per cycle to each unit, for a maximum dispatch rate of six.

In the next stage, each ROP reads its operands from the register file. Most ROPs then take one cycle in the execute stage; floating-point calculations take two. Load and store ROPs take two cycles to execute but are fully pipelined, as the figure shows. Once ROPs complete, they remain in the ICU until all previous ROPs have completed; they are then retired in order. The Nx686 can retire four ROPs per cycle, matching its peak issue rate.

The pipeline is improved from the Nx586, which requires three cycles to perform a load or store; the Nx686 takes two cycles. The mispredicted branch penalty remains five cycles, more if instructions are executing out of order. There is a single-cycle load-use penalty, but it is usually hidden by the out-of-order execution.

This pipeline is half the length of Pentium Pro's. The shorter pipeline greatly reduces the penalties for branches and loads. This, in turn, results in improved efficiency (net instructions per cycle), supporting NexGen's claim to deliver better performance than PPro at the same clock speed.

The shorter pipeline also simplifies the chip's design by reducing bypassing and synchronization logic. NexGen believes these issues were critical to producing a die smaller than Intel's.

The advantage of a longer pipeline is its clock speed. Intel expects that Pentium Pro will be about 40% faster than Pentium in the same IC process. Instruction predecoding should help the Nx686 run a bit faster than Pentium, but it will probably still be slower than PPro in comparable manufacturing processes. For example, NexGen's target clock rate is 180 MHz in a 0.35-micron process; we expect PPro to reach 200–220 MHz in a similar process, giving it a 10–20% clock-speed edge. NexGen, however, has yet to demonstrate that the Nx686 can meet its clock-speed goals.

Superior Branch Handling

As with any out-of-order processor, branches must be handled quickly and accurately to prevent pipeline-draining mispredictions. NexGen placed significant emphasis on this area, using some unusual strategies.

To start, the Nx686 retains from its predecessor an advanced two-level branch-prediction algorithm. The new chip uses an 8,192-entry branch history table, four times the size of the BHT in the Nx586 and far bigger than that of any other announced microprocessor. The table is indexed using the GAs method *(see* ***090405.PDF****)* by combining 9 bits of global branch history with 4 bits of PC. This 13-bit index selects one of the 8,192 entries, each of which contains the standard two history bits. This structure should predict branches with an accuracy of about 95% on programs such as SPEC92.

To accommodate such a large BHT, the chip does not store predicted target addresses, except for a 16-entry return-address stack. For other types of branches, the target address is calculated on the fly during the decode stage, as Figure 3 shows. To calculate the address this quickly, parallel adders calculate all possible target addresses before the instructions are fully decoded; by the end of the decode cycle, the chip chooses which (if any) of these results is valid.

If the branch is predicted taken, the target address is known right away, and the target instructions can be fetched on the following cycle. This strategy, also used in the R10000, causes a one-cycle taken-branch penalty, but this penalty is hidden as long as there are enough instructions already waiting in the ICU.

To avoid this one-cycle penalty, the Nx686 includes a branch-target cache (BTC) similar to the one in the Nx586. The BTC contains 16 entries, each holding 16 instruction bytes. It is indexed by the branch address and so is accessed during the decode stage. If there is a hit in the BTC and the BHT predicts a taken branch, the instructions from the BTC are sent to the decoder. NexGen estimates that this 256-byte BTC hits about 70% of the time, eliminating the branch penalty in those cases.

Thus, the high accuracy of the BHT avoids the lengthy mispredicted branch penalty in nearly all cases, but many branches have a single-cycle penalty. As this penalty is often hidden by the instruction queue, NexGen did not spend as much die area to eliminate it, leaving more area for the large BHT. Note that the BHT consumes about as much space as a 2K cache.

Other than NexGen's chips, Pentium Pro is the only announced processor to implement the two-level prediction algorithm. Although Intel has not revealed the full details of its two-level BTB, it appears to have far fewer entries than the Nx686's BHT. We expect that the NexGen processor will have better branch-prediction accuracy, although it will encounter occasional single-cycle branch penalties not seen by Pentium Pro.

Large On-Chip Caches

The Nx686 features 48K of on-chip cache, at least

50% more than any other x86 processor. Unlike AMD, which emphasizes instruction-cache size in the K5, NexGen chose to double the size of the data cache to 32K while leaving the instruction cache at 16K. Both are two-way set-associative. The data cache is pipelined, performing two accesses per cycle. To reduce pipeline delays, the load is performed first, followed by the store one-half cycle later. This technique, similar to one used in IBM's Power2 processor, eliminates bank conflicts.

These caches are backed by a dedicated interface to a unified direct-mapped L2 cache. This cache requires standard pipelined synchronous SRAMs, which are becoming increasingly common in PCs. The cache can run at submultiples of the CPU clock, although a full-speed cache can be used for maximum performance. With a half-speed cache, the first 64-bit word is returned in seven cycles (7-2-2-2 access rate). A full-speed cache, which requires expensive 180-MHz SRAMs, delivers 5-1-1-1 accesses. These accesses are pipelined, although only the full-speed cache is fully pipelined.

All cache-control logic is on chip; only data and tag SRAMs are needed to complete the L2 cache. (Unlike the Nx586, the new chip uses separate SRAMs to store data and tags.) The external cache operates in write-back mode and has a 32-byte line size. The Nx686 supports up to 2M of cache, although most desktop systems will probably use a 256K or 512K cache. The relatively large on-chip caches provide good performance even with smaller L2 caches.

MICHAEL MUSTACCHI

NexGen CPU architect Greg Favor describes the advanced capabilities of the Nx686.

In contrast, Pentium Pro has only 16K of on-chip cache but adds a four-way associative 256K level-two cache in the same package. This L2 cache operates at the CPU clock speed, is fully pipelined, and has a three-cycle latency. NexGen believes that, on many applications, the higher hit rate of its on-chip caches will compensate for the longer latency of its external cache.

Of course, the performance of the Nx686 will vary widely, depending on the size and speed of the L2 cache. This gives the system vendor flexibility in choosing a price/performance point. It also means that Nx686 systems with a small half-speed cache may not match Pentium Pro's 32-bit performance. On the other hand, systems with a large, full-speed cache might outrun Pentium Pro.

Limited Chip-Set Options

The separate system bus is also 64 bits wide. This bus, called NexBus, runs at a submultiple of the CPU clock and is limited to 66 MHz. This speed is quite adequate, given that the bus need not support the L2 cache.

The Nx686 uses a slightly improved version of NexBus; the recently released NexGen PCI chip set *(see **0913MSB.PDF**)* supports both the 586 and 686 processors. No other vendor currently sells NexBus chip sets, although VLSI has an option to market the PCI chip set (which it builds for NexGen).

The Nx686 uses a 365-pin package that is not compatible with the Nx586 pinout. The new chip does not support the external floating-point unit or asynchronous cache used by its predecessor. For upgrade purposes, it would be possible to design a version of the 686 to plug into a 586 socket.

NexGen's Favor would not reveal the power dissipation of the Nx686 but claimed that it would be comparable to that of Pentium (10 W max, 4 W typ) and much less than that of Pentium Pro. To save power, the NexGen part disables function units that are not being used, much as Pentium does. It operates at a core voltage of 2.5 V, reducing power usage, with 3.3-V–compatible I/O. The Nx686 includes a system-management mode (not compatible with Intel's) and stop-clock features. Favor believes the chip will be used in some notebook systems, although the primary market for the part is desktops.

Without revealing the die size of the six-million-transistor chip, he says that the Nx686 will compare favorably with the 0.35-micron Pentium Pro, which we believe will be roughly 200 mm^2. This would put the Nx686's manufacturing cost at about $150, not much more than a K5 or a Cyrix 6x86. IBM will build the NexGen device in its 0.35-micron five-layer-metal CMOS-5X process, a fairly mature process that has already been in production for about a year.

NexGen expects the initial version of the Nx686 to operate at 180 MHz. As the Nx586, which uses a similar pipeline, should reach 150–160 MHz in CMOS-5X, the new chip has an aggressive but achievable target. Even at just 150 MHz, the Nx686 should deliver better performance than any version of Pentium, the K5, or 6x86 expected to ship in 1996. Nx686 performance will continue to increase as the chip moves to IBM's 0.27-micron CMOS-6S and future processes.

NexGen received first silicon during the past summer and, at the Microprocessor Forum, demonstrated an Nx686 system running common PC applications. These initial prototypes are built in a 0.5-micron process and do not run at 180 MHz, but the company believes that the final version will meet its clock-speed goal.

Price & Availability

NexGen has not formally announced the Nx686, so pricing is unavailable. The company expects volume production to begin in mid-1996. Contact NexGen (Milpitas, Calif.) at 408.435.0202; fax 408.435.0262, or access the Web at *www.nexgen.com.*

Beating Pentium Pro to the Mainstream

Even if the Nx686 just comes close to its goals, it will outperform any Pentium-class processor, leaving Pentium Pro as its only competition. Throughout 1996, however, PPro will be restricted to high-end desktops and servers for two reasons. First, it will be expensive, although by the end of 1996, PPro prices may drop as low as $600. In addition, the first- and second-generation PPro processors will be restricted to Windows NT and other 32-bit operating systems, due to relatively poor performance on Windows 95.

Thus, if the Nx686 can debut as planned, it will be in a powerful position as the fastest processor in the world for Windows 95 applications, putting it in the thick of the mainstream PC market. Even on Windows NT, the NexGen product should offer better price/performance than Pentium Pro. The Nx686's multimedia extensions should give it a further advantage.

Intel's plan is to deliver Pentium Pro to the masses with a third-generation part sometimes called the P68. We believe this device will use a 0.28-micron process to reduce the cost of Pentium Pro to moderate levels. This design will include Intel's multimedia extensions, as in the P55C, and is likely to offer improved performance on Windows 95. It will probably move to a single-chip design with an external L2 cache, eliminating the current MCM package.

In short, the P68 might look an awful lot like the Nx686. We don't expect the P68 to ship until early 1997, giving NexGen as much as a nine-month head start in the market. When the P68 does arrive, it will probably outperform the Nx686, given Intel's more advanced manufacturing process, until NexGen can deploy its 0.27-micron version.

At best, NexGen could have the only processor offering better-than-Pentium performance on Windows 95 for a period of several months. Even after the P68 rolls out, NexGen should be able to match, or at least approach, the performance of Intel's fastest processors. Cyrix and AMD will eventually roll out P6-class devices (as opposed to the Pentium-class 6x86, for example), but NexGen will have the advantage of being first to market.

A Complete Product Strategy

The Nx686 will not immediately eclipse the current 586. The company plans to continue that product to fill the low end of its line, as it has significantly lower manufacturing costs than the 686. When the 686 debuts, it will probably sell for several hundred dollars, while the 586 covers the $100–$300 range. Even to meet these prices, the 586 clock speed must be improved from its current 93-MHz limit through process shrinks, and its lack of an FPU must be addressed.

Individually, NexGen's product plans seem reasonably achievable, but to meet its goals the small company must execute better than it has recently. The combination of faster 586 parts, the 686, and its PCI chip set would put the vendor in a strong position to challenge AMD and Cyrix.

One problem for NexGen is the incompatible pinout and bus design of its parts, which make it impossible to drop a NexGen processor into a Pentium motherboard or connect it to a Pentium chip set. Although many PC makers buy complete motherboards, not CPUs, the limited chip-set selection prevents NexGen from offering the same range of features and performance as Pentium motherboards. Other Intel competitors are likely to face the same challenge when they introduce P6-class parts.

Another problem for NexGen is its size: the fledgling company is well behind AMD or even Cyrix in its ability to attract and support high-volume PC makers. (One exception is Compaq, a major investor, which is likely to sell Nx686-based PCs.) Over time, manufacturing capacity and support issues can be improved. In the short term, NexGen needs to hold only a few percent of the x86 market to become profitable. By being the first to challenge Pentium Pro, the Nx686 offers the potential for performance leadership that could push NexGen to profitability and beyond. ♦