Cahners
MICROPROCESSOR REPORT
www.MPRonline.com
THE INSIDER'S GUIDE TO MICROPROCESSOR HARDWARE

MIPS 20Kc Is Fastest Licensable Core

New MIPS64/MIPS-3D High-Performance Core Targets IP Market

By Keith Diefendorff {7/3/00-01}

Like many microprocessor companies, MIPS Technologies sees digital-entertainment and networking as large and rapidly growing opportunities. Answering the call for the high-performance embedded processors these markets need, MIPS has designed a new microarchitecture, called the 20K, based on the MIPS64 architecture and enhanced with the MIPS-3D instructions for better geometry processing (see *MPR 8/23/99-en*, "MIPS Adds a New Dimension to MIPS64"). Speaking at last month's **Embedded Processor Forum**, Victor Peng, engineering director at MIPS, laid out the details of his company's newest and fastest 64-bit design.

The new design, code-named Ruby, will be sold commercially in two forms: a complete standalone microprocessor, called the R20K, and a licensable core, called the 20Kc, for use in ASICs, ASSPs, and SoCs. The new offerings extend MIPS's current family of licensable 32-bit low-end and 64-bit midrange devices, the 4Kc (see *MPR 5/31/99-05*, "Jade Enriches MIPS Embedded Family") and 5Kc (see *MPR 10/25/99-05*, "MIPS Plays Hardball With Soft Cores"). R20K silicon is expected next quarter (3Q00), and the 20Kc core will see first silicon in 1Q01.

Despite the fact that Peng was working on the "H2" processor at SGI when MIPS spun out as a separate company (see *MPR 4/20/98-01*, "Silicon Graphics, MIPS Part Ways"), the new 20K design inherits little from that earlier design. The H2 was a wide-issue superscalar out-of-order machine aimed at high-end workstations and servers, whereas the new 20K is a far simpler dual-issue in-order design for the embedded space, as Figure 1 shows.

As typical embedded processors go today, however, the 20K is still quite aggressive, and it will probably fulfill MIPS's claim of being the highest-performance licensable

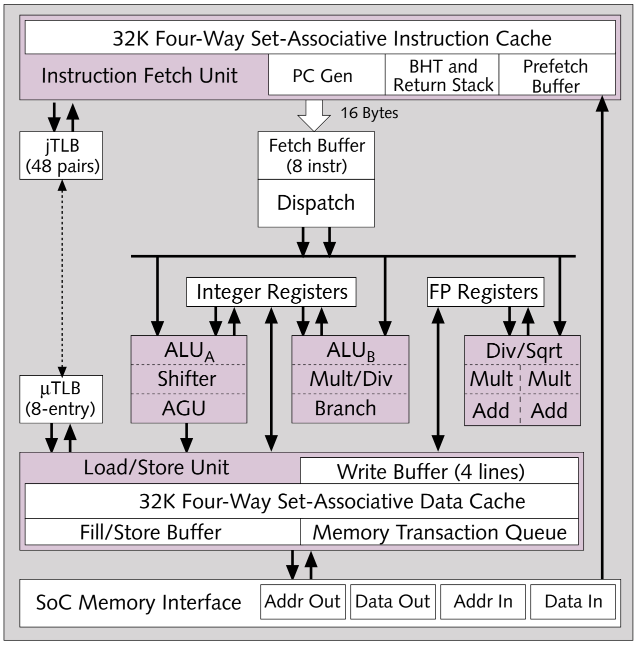


Figure 1. The 20Kc has a dual-issue in-order microarchitecture and implements a seven-stage pipeline. With this design, MIPS believes it is hitting a sweet-spot design space that gives the best possible instruction-level parallelism and frequency with- out the complexity of wide-issue out-of-order or long-pipeline designs.

	MIPS R20K	
Feature	0.18 Micron	0.15 Micron
Architecture	MIPS64 w/MIPS-3D	
Issue Width	2 instructions	
Issue Order	In order	
Pipeline Length	7 stages	
Execution Units	2 integer, 1 floating-point/3D	
On-Chip Cache	32K/32K, 4-way set-associative	
Data TLB	8-entry µTLB, 48-pair jTLB	
Instruction TLB	48-pair jTLB	
Transistors	7.2 million (≈2 million logic)	
Core Size	9mm^2	6mm^2
Core w/caches	34mm^2	24mm^2
Voltage	1.5–1.8V	1.5V
Frequency	450–600MHz	600–750MHz
Performance At:	600MHz	750MHz
Dhrystone 2.1 (est)	1,200	1,500
Peak Floating Point	2.4 GFLOPS	3.0 GFLOPS
3D Transforms	30 Mpolygons/s	37 Mpolygons/s
Power	1.8W	2.0W

Table 1. The 20Kc microarchitecture offers a good balance of performance, frequency, power, and die size.

core. As Table 1 shows, the 20K will operate at up to 600MHz in a 0.18-micron six-layer-metal process (at nominal channel lengths), delivering an integer performance of around 1,200 Dhrystone 2.1 mips and a floating-point 3D-geometry performance of 2.4 GFLOPS and 30 million polygons per second. Yet at only 34mm^2, including dual 32K caches, the core is small enough to be a candidate for even low-cost embedded systems.

One feature curiously missing from the 20K, considering its target market, is SIMD instructions for media processing. The 20K doesn't even implement MIPS's relatively modest MDMX extension (see *MPR 11/18/96-06*, "Digital, MIPS Add Multimedia Extensions"), which was an optional extension to the MIPS V architecture for exactly this purpose. MIPS says the omission was intentional, to save silicon area (cost), and was made on the assumption that the scalar performance of the core would be adequate for most multimedia algorithms. This may be true, but in a world where processors will increasingly be called upon to execute multiple near-real-time tasks in parallel, the lack of SIMD capability reduces the capacity of the core, and this will inevitably limit the core's potential scope.

MIPS-3D Mnemonic	Data Types	Description	Use
ADDR	.ps	SIMD reduction add	Vertex transforms
MULR	.ps	SIMD reduction mult	Vertex transforms
CABS	.s/.d/.ps	FP compare absolute	Clip tests
BC1ANY2T	n/a	Cond br on any 2 true	Clip tests
BC1ANY2F	n/a	Cond br on any 2 false	Clip tests
BC1ANY4T	n/a	Cond br on any 4 true	Clip tests
BC1ANY4F	n/a	Cond br on any 4 false	Clip tests
RECIP1	.s/.d/.ps	Reciprocal estimate	Perspective, lighting
RECIP2	.s/.d/.ps	Reciprocal refinement	Perspective, lighting
RSQRT1	.s/.d/.ps	Recip square root est	Normalization
RSQRT2	.s/.d/.ps	Recip sq root refine	Normalization
CVT.ps.pw	.ps	Convert ints to floats	General
CVT.pw.ps	.ps	Convert floats to ints	General

Table 2. The MIPS-3D extension adds 13 instructions to the MIPS V architecture. The extensions have a big effect on 3D-geometry processing performance but require only 3% more silicon than a standard double-precision floating-point unit. MIPS-3D extensions use the normal MIPS floating-point registers. n/a = not applicable.

Bucking a Trend

In the PC-3D-graphics arena, the trend is clearly toward moving the floating-point geometry-processing task onto dedicated hardware in the rendering chips. The reasons behind this trend, however, are that the floating-point performance of Intel x86 microprocessors has traditionally been weak, and hardware-rendering performance is outpacing the growth of CPU performance by a factor of two, leaving geometry processing as the bottleneck in 3D applications. As a result, 3D-chip companies have been forced to implement their own floating-point geometry-processing pipelines.

MIPS argues, however, that as long as performance is adequate, geometry processing is better performed by the processor, because of its superior flexibility and programmability. Furthermore, in the embedded space, where cost is king, doing geometry processing on the processor is a more cost-effective solution. MIPS contends that high-end embedded processors need good floating-point capability for reasons other than 3D anyway, and that it costs very little to upgrade the FPU for good 3D-geometry processing. MIPS says its MIPS-3D solution doubles geometry performance while adding only about 3% to the size of the 20K's FPU (and the FPU occupies only about 10% of the core).

MIPS's solution for 3D-geometry processing is to divide the basic double-precision floating-point unit into two halves, providing two-wide single-precision SIMD processing. This solution reuses the transistors in the double-precision multiplier, adder, and shifters, and adds only the cost of a second exponent pipe and some control logic. The double-precision divide/square-root hardware in the FPU is not sliced, since iterative division is not well suited for SIMD execution. In the place of full division and square root, reciprocal-estimate and reciprocal-square-root-estimate instructions are provided, which are well suited to SIMD execution.

The 20K's SIMD operations are carried out by nine new paired-single (.ps) instructions and four conditional branches for clip tests. The new instructions, shown in Table 2, implemented by a small amount of additional silicon, make a huge difference in 3D performance. As Figure 2 shows, MIPS-3D increases 3D-transform performance by 45%, and transform-and-lighting performance by almost 85%. The 20K's 3D-transform rate is more than six times that of a Pentium III with SSE at the same frequency.

Short Pipe Avoids Complexity

Bucking another trend—the trend toward very long pipelines—the 20K implements a moderately short, seven-stage in-order pipeline, as Figure 3 shows. Peng says the MIPS architects looked closely at longer pipelines but rejected them for several reasons. Although a longer pipeline would have permitted a higher clock frequency, it would have added a large number of transistors for pipeline latches, and it would not have increased performance unless even more transistors were added for forwarding stages and powerful branch-prediction logic. MIPS decided that the silicon area and power consumption were not good trade-offs for the embedded markets it seeks. The decision is probably a good one, considering that in embedded markets, unlike in the PC market, performance, cost, and power are generally more important than frequency.

Although the 20K's pipeline dispatches, executes, and completes instructions only in strict program order, the architects pulled a few tricks to eliminate as many stalls as possible. For one, the 20K implements a decoupled fetch unit that uses dynamic branch prediction to fetch the predicted execution path into an eight-instruction fetch buffer. This technique, which has been used by other processors, such as the WinChip 4 (see MPR 12/7/98-05, "WinChip 4 Thumbs Nose at ILP"), allows the instruction-fetch unit to run ahead of the rest of the pipeline to keep the dispatch unit supplied with instructions.

The fetch unit comprises the first two stages of the pipeline. During the fetch stage, an aligned block of four instructions is fetched from the 32K four-way set-associative instruction cache and deposited in the instruction buffer. The instruction cache uses way-prediction, so that only one-quarter of the cache is powered up each cycle. A way mispredict costs one cycle in the pipeline, but this occasional bubble is usually squashed in the fetch buffer.

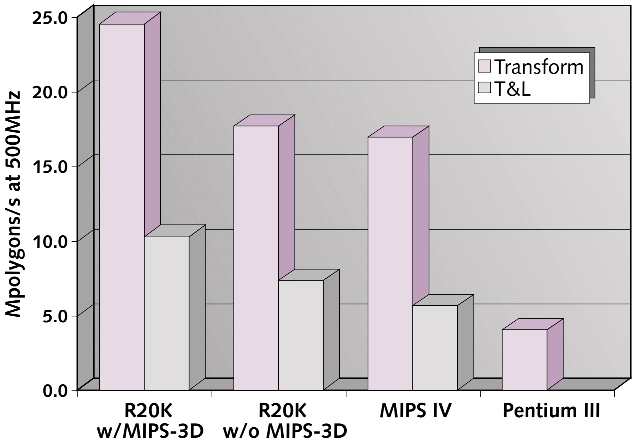


Figure 2. The MIPS-3D instructions boost 3D-transform processing by 45% over a standard MIPS IV core without those extensions at the same clock frequency; transform and lighting performance is boosted 83%. The R20K offers more than six times the 3D-transform speed of a Pentium III at the same frequency. (Source: MIPS)

The validate stage determines which instructions in each fetch block are legitimate candidates for execution, considering that branches can occur into or out of the block. The 256-entry two-bit branch history table (BHT) can predict up to two branches in each fetch block on every cycle. The best-case branch-mispredict penalty is five cycles. The instruction cache stores 16 bits (14+2 parity) of predecode information per fetch block, and branch displacement calculations are performed as branches are brought into the cache, saving time in the more performance-critical execution pipeline.

The decode phase of the pipeline occupies the next two stages. The first stage decodes instructions, and the second stage reads source operands from the register files (or bypass buses) and dispatches/issues instructions to the execution units. Up to two instructions can be dispatched on each cycle, subject to data and structural hazards. The 20K implements no register renaming, reorder buffers, or other fancy out-of-order features that reduce data-dependency-interlock stalls, but it does provide short-latency execution units and register bypass circuits to minimize the length of the stalls. Also, Peng says, the 20K uses the same trick used by many MIPS designs: it predicts exceptional conditions, such as floating-point overflow, early in the pipeline, thus eliminating all time-critical feedback circuits from the execution stages.

The primary structural hazards arise from busy execution units or from a lack of execution units of the needed type. The 20K has one dispatchable floating-point unit and

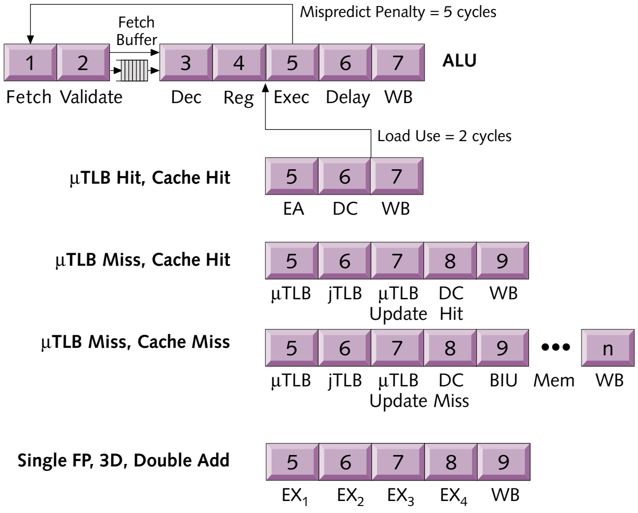
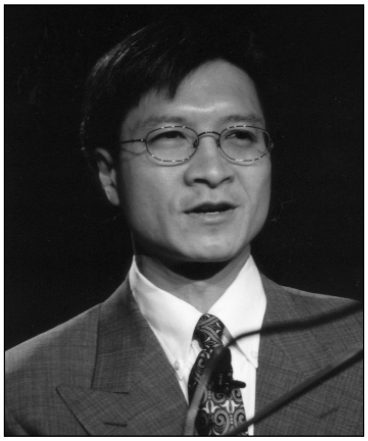


Figure 3. The 20K uses a basic seven-stage pipeline for both ALU and load/store instructions. On a load or store, a miss in the µTLB adds two stages to the basic pipeline. A cache miss adds a minimum of four cycles to the pipeline, but waits on memory usually add many more. The basic single-precision, paired-single, and double-precision-add pipeline is nine stages. Double-precision multiplies and multiply-adds require a second iteration of the multiplier array; thus, these operations are not fully pipelined and can issue at a maximum rate of only one every other cycle.

two dispatchable integer units. Integer unit A includes a full ALU, the shifter, and the load/store address-generation unit (AGU). Integer unit B contains a second full ALU, the integer multiplier/divider, and the branch unit. With this organization, the 20K can issue two instructions per cycle, as long as they are paired, as indicated in Figure 4.

The execution resources for all instructions except integer multiply and divide and some floating-point double-precision instructions (MUL.d, MADD.d, RECIP1, and RSQRT1) are fully pipelined and can accept new instructions at a rate of one per cycle. These double-precision instructions require a second iteration through the multiplier and thus can accept a new instruction only on every other clock cycle. Most integer and load/store instructions (cache hits) have single-cycle latency; most floating-point operations have a latency of four cycles, except those that require a second pass through the multiplier, which take five cycles.

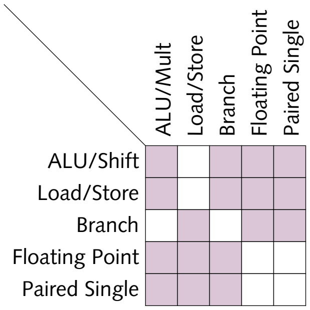


MIPS engineering director Victor Peng describes his company's new 64-bit R20K at the Forum.

MICHAEL MUSTACCHI

Computer-Style Memory System in an EC

For load and store operations, the first stage of the execution pipeline (stage #5) computes the effective address (EA) and translates it to a physical address by lookup in the microTLB (µTLB). The µTLB is a very fast fully associative data-only TLB with eight entries. A miss in the µTLB inserts two additional cycles in the load/store store pipeline: one to access the larger joint TLB (jTLB) and a second to update the µTLB. The jTLB contains 48 even/odd address pairs and includes both data and instruction translations. It is only single ported, but conflict between the load/store unit and the instruction-fetch unit is rare, because the majority of data accesses hit in the µTLB. Also, the eight-instruction fetch buffer allows the instruction-fetch unit to slip a cycle now and then without disrupting the decode pipeline. A miss in the jTLB raises an exception so that a software routine can bring in the required translation.



	ALU/Mult	Load/Store	Branch	Floating Point	Paired Single
ALU/Shift	■		■	■	■
Load/Store	■		■	■	■
Branch		■		■	■
Floating Point	■	■	■		
Paired Single	■	■	■		

Figure 4. The 20Kc pipeline can dual-issue all pairs of instructions indicated in purple. If a structural hazard blocks execution of one instruction, the processor will attempt to issue that instruction along with a new instruction from the fetch buffer on the next cycle.

Once the translated address is available, the 20K uses the physical address to index its 32K four-way set-associative data cache in one cycle, writing results back to the register file in the final (seventh) stage of the pipeline. The cache implements a 32-byte line and provides 4GB/s of read bandwidth and 16GB/s of write bandwidth at 500MHz. A 128-byte (four cache lines) fill buffer boosts write bandwidth into the cache and also enables store-load merging.

The cache-write policy is programmable to be either writeback/write-allocate or writethrough/write-no-allocate, and a store buffer is provided to collect store misses on up to four different cache lines. Data waiting in the store buffer can be forwarded directly to a pending load, thus minimizing stall time in this surprisingly common case.

The data cache is nonblocking, supporting up to four pending data-read misses. The 20K implements the MEI subset of the four-state MESI (modified, exclusive, shared, invalid) cache-coherence protocol. Generally, the MEI subset is adequate for I/O coherence in single-processor systems, but it lacks the shared state needed for efficient support of symmetric-multiprocessing (SMP) systems.

The 20K also implements cache-line locking through special instructions for both the instruction and data caches. This feature is considered a must for real-time support in embedded processors. Other special cache instructions provide data-prefetch hints, nudging lines out of the cache and streaming data through the cache. In addition, the hardware uses otherwise idle cache cycles to opportunistically refill lines, thus reducing the average cache-access latency.

Yet Another "Bus"

Although the 20Kc core leaves the issue of a memory interface entirely up to the user, MIPS had to devise a solution for the R20K processor. This problem is a huge issue for companies trying to develop a standalone processor today. Traditional multidrop bus interfaces, such as Intel's P6, AMD's Socket 7, and PowerPC's 60x, are rapidly going out of favor, because transmission-line effects prevent them from operating at sufficiently high speeds.

Point-to-point links are the answer to this problem, but at this time there are no widely accepted open standards to which a vendor can hitch its wagon. And most companies are reluctant to throw in with other companies or competitors, even though in most cases that alternative would be better than creating yet another new bus. The best alternative from a performance perspective is to implement the memory controller on the processor with a direct interface to the memory of choice. But here's the rub: there is no clear memory of choice in the embedded space. The processor must be able to accommodate a variety of memory types, a fact that brings the choice back to some type of generic interface.

And so, for the R20K, MIPS invented a brand-new generic interface it calls MGB, which originally stood for "MIPS gigabyte per second" but is now just MGB. In its R20K incarnation, this interface is a 32-bit-input/64-bit-output point-to-point interface that provides a peak bandwidth of 3.6GB/s at 150MHz DDR. The protocol supports split transactions and allows for out-of-order return of data using transaction tags. Flow control is credit based to provide good efficiency without the complexity of retry mechanisms. Cache coherence is maintained by invalidate and intervention transactions. Signaling is optionally either synchronous or source-synchronous on a series-terminated 1.5V HSTL electrical interface. The 20Kc core actually implements the MGB protocol, but the details of the physical interface could be different in other implementations.

Entering a Crowded Space

Frustrated by the hopelessness of competing with Intel in the high-end microprocessor space, many architects and companies have turned their attention to the more fertile area of high-end embedded processors. The only problem with this market is that everyone else is also pursuing it. Although only a few strong competitors have products on the market today, that situation will change rapidly.

The 20K's best opportunity is probably in the digital-entertainment market, where, according to Semico, MIPS already holds a 70% market share (this market includes video games, set-top boxes, and digital TV). Approximately 80 million units were sold into this market in 1999, and the number is expected to more than triple to 250 million units by 2003, according to Dataquest. Back in 1Q99, both NEC and Toshiba licensed the R20K, presumably for use in this market.

The Nintendo 64 and the Sony PlayStation both use MIPS-architecture-based designs. However, the 20K core probably does not have a future in the next generation of these super-high-volume consoles. Nintendo says it will switch to a PowerPC for its next-generation Dolphin system (see *MPR 5/31/99-en*, "Nintendo to Battle PlayStation with PowerPC"), and Sony did its own MIPS-core design for PlayStation 2 (see *MPR 4/19/99-01*, "Sony's Emotionally Charged Chip"). The 20Kc could potentially find a home in some lower-volume game consoles—NEC, for example, supplies a graphics chip in Sega's Dreamcast—but it simply doesn't have enough horsepower to compete with next-generation game-console processors like the ones in the PlayStation 2 or X-Box (see *MPR 4/3/00-01*, "Microsoft Weighs In With X-Box").

The small size of the 20Kc core, however, as Figure 5 shows, makes it ideal for integration with 3D-rendering and video-encoding/decoding hardware for other digital-entertainment applications, such as digital TVs. In this form, the 20K could provide an alternative to highly specialized devices like Equator's MAP-CA (see *MPR 3/13/00-04*, "MAP-CA Ready for Prime Time"). The ease of programming the 20K's general-purpose architecture and the wide availability of tools for the MIPS architecture would give it an advantage over the more specialized parts.

The going will be tougher for MIPS in the networking space, an area that every microprocessor company on the planet is drooling over. In this market, MIPS has no special advantage, and the competition will be fierce. The toughest competition will probably come from processors designed from the ground up for network processing, such as Intel's IXP1200 (see *MPR 9/13/99-01*, "Intel Network Processor Targets Routers") and Agere's new NPU (see *MPR 6/12/00-02*, "Agere's Pipelined Dream Chip"). Competition may also come from general-purpose cores integrated with special-purpose network-processing hardware, such as would be possible from the Motorola/C-Port marriage (see *MPR 3/6/00-03*, "Motorola Buys C-Port: Smart Move"). Although the 20Kc is perfectly viable as a core in this same way, many of the big players in this market have already aligned with a core architecture and MIPS may have a hard time breaking in.

Furthermore, high performance is crucial for network processing. While MIPS's claim that the 20K will be the highest-performance *licensable* design is true, it will certainly not be the highest-performance *nonlicensable* design. Sources indicate that Motorola is nearing tapeout on a

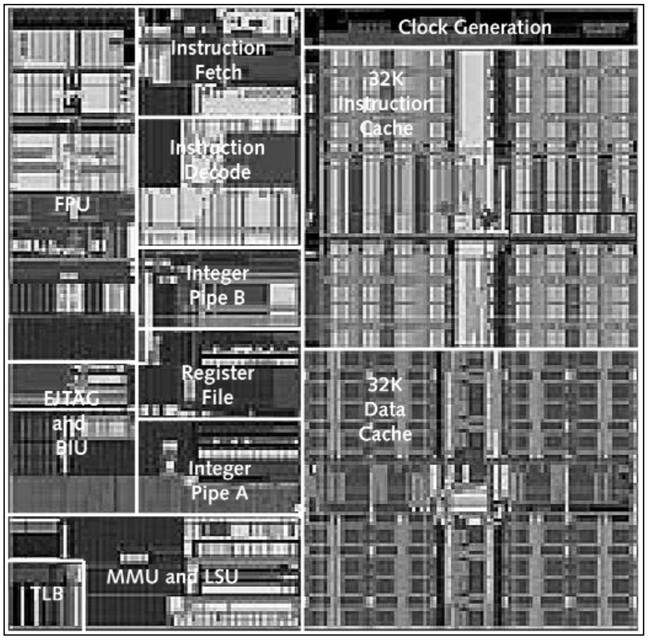


Figure 5. This layout shows the MIPS 20Kc core, which requires about 7.2 million transistors, 2 million of which are logic. The core uses 34mm^2 of silicon in a 0.18-micron six-layer-metal process and will shrink to about 24mm^2 in 0.15 micron. (Source: MIPS)

Price & Availability

MIPS expects first silicon on the R20K processor near the end of the third quarter of this year. The 20Kc core will tape out late in the fourth quarter and have first silicon on a test chip in the first quarter of next year. No pricing has been announced for either product. For more information, go to *www.mips.com*.

2GHz copper 0.13-micron (HIP7) PowerPC processor for Apple and for the networking market. Although Motorola probably won't license this core, the company itself already has a huge share of the networking market, and a PowerPC processor of such capability would be formidable competition for a 20Kc-based design from a network-processor-company wannabe.

One indication that the 20K may not have enough performance for the networking market is SiByte. SiByte is a Silicon Valley startup company, led by well-known ex-Alpha designers Dan Dobberpuhl and Jim Keller and formed to create a MIPS-based network processor. That company concluded that the 20Kc core would not have sufficient performance for its needs, so it licensed the MIPS64/MIPS-3D architecture from MIPS and designed its own core, the SB-1 (see *MPR 6/26/00-04*, "SiByte Reveals 64-bit Core for NPUs"). That core, which was described at Embedded Processor Forum by Robert Stepanian, will operate at 1GHz, issue twice as many instructions per cycle as the 20K, have twice the number of FPU/MIPS-3D and load/store execution units as the 20K, and is designed to efficiently support chip multiprocessing (CMP). Even though the SB-1 will have about the same die size as the 20Kc in a similar process, it will offer substantially more performance in networking applications (and possibly in digital-entertainment applications as well).

Targeting the IP Market

The performance difference between the 20Kc and SB-1 is probably explained by the radically different strategies the two companies are pursuing. On the one hand, SiByte, as a fabless chip company, intends to keep its core proprietary and leverage it by designing integrated parts around it and then selling those products directly to the market. MIPS's strategy, on the other hand, is to create an IP (intellectual-property) business. As such, MIPS plans to freely license the 20Kc to build momentum and volume across a much larger number of applications than any single fabless chip company like SiByte could ever hope to serve.

The price MIPS pays for the flexibility that is required to license its core to a wide range of companies and applications is, apparently, some sacrifice in performance. Whereas SiByte can tune and optimize its design for its own limited uses, MIPS must make its core more generic and easier to integrate with other devices and into different process technologies. Peng said, for example, that the 20Kc core is portable to a wide range of IC processes. MIPS designed the part using internal generic design rules so it could be easily converted to any specific design rules through a simple set of mathematical sizing and scaling transformations. Designing in this way is an approach that is bound to leave some density, frequency, and power lying on the table.

MIPS's play for the IP business has already cleared a big hurdle. Back in the first quarter of this year, TSMC announced it had licensed the 20Kc core from MIPS. TSMC, which is the largest semiconductor foundry in the world by a wide margin (see *MPR 6/5/00-01*, "TSMC Sets Sights on #1"), is a primary source of IP modules to the worldwide design community. With the 20Kc available through this source, startup and entrenched companies alike will have easy access to it. And as the highest-performance core in TSMC's IP portfolio, the 20Kc will be the obvious choice for a number of those companies.

The big question is whether the gains in licensability are worth the tradeoffs in performance. The simple answer is probably yes. Both the digital-entertainment and network-processing markets are large and growing rapidly. These markets encompass a wide and diverse set of needs and customers. Some designs will need the absolutely highest performance, and for those a special-purpose network processor or a high-performance SiByte solution may be best. But for a large number of designs, the ability to easily license an off-the-shelf core that has good performance will be a better solution. For those designs, the 20Kc looks like a great product. ◇