MICROPROCESSOR REPORT

THE INSIDERS' GUIDE TO MICROPROCESSOR HARDWARE

VOLUME 8 NUMBER 15 NOVEMBER 14, 1994

PA-8000 Combines Complexity and Speed

HP Aims to Retake Performance Lead—But Not Until 1Q96

by Linley Gwennap

Long a proponent of simple, fast processors, HP has succumbed to the siren call of complexity, creating the most feature-filled RISC design yet revealed. Steve Manglesdorf, presenting at last month's Microprocessor Forum, said that the forthcoming PA-8000 will achieve high clock rates despite the burden of this feature set, a powerful combination that he claims will create the industry's fastest microprocessor. It will take quite some time, however, to validate this claim; HP does not expect system shipments for nearly 18 months.

The new chip is similar to the MIPS R10000 *(see **081403.PDF**)* in its decoupled architecture with four-instruction dispatch and aggressive out-of-order execution. It goes beyond the MIPS design by adding dual floating-point units and dual load/store pipes as well as a larger out-of-order dispatch window. The most unusual feature is the total lack of on-chip cache; large external primary caches are an HP tradition but differ sharply from the designs used by other vendors.

The PA-8000 will be the first HP chip to implement a 64-bit architecture, dubbed PA-RISC 2.0. Like other next-generation RISCs, the HP design initially will be expensive to produce, appearing first in servers and high-end workstations. A quick shrink is planned to bring cost to moderate levels while increasing performance. The first completely new processor design from HP since 1991, the PA-8000 and its derivatives will carry HP until the first fruits of its Intel alliance appear, probably around 1998.

The new processor has not yet taped out; first silicon is expected early next year. Without seeing working parts, Manglesdorf was unwilling to disclose specific clock speed or performance targets for the PA-8000. He also declined to discuss physical details such as die area and package size. He did, however, provide an extensive description of the chip's microarchitecture.

Large Primary Caches Kept Off Chip

PA-RISC processors have always used external primary caches to increase the amount of data that could be accessed in a single cycle. HP's past two designs, however, added small (1–2K) on-chip buffers to reduce the number of accesses to these caches; it had been speculated that the PA-8000, with its higher bandwidth requirements, would be forced to implement on-chip primary caches, like nearly every other high-performance CPU.

But HP has again diverged from the common wisdom by avoiding on-chip caches for its latest design. To support high clock speeds, which we expect to reach 200 MHz, the designers have switched from asynchronous SRAMs to synchronous parts. Existing PA-RISC chips use a wave pipeline to launch the address for the next cache access before the data is received from the previous access; the new synchronous SRAMs have registers on both the input and output to make this pipelining explicit. This design extends the cache latency to roughly two cycles, compared with one-and-a-half in current HP processors; with complete pipelining, however, the cache still supports one access per cycle. (By including address latency, HP calls this design a three-cycle cache.)

Manglesdorf said that availability of the required SRAMs should not be a limiting factor to the chip's clock frequency. We expect that by 1Q96, synchronous SRAMs will be available at 200 MHz (5 ns); the PA-8000 will need to operate at or near this frequency to achieve its aggressive performance goals.

In "superpipelined" designs such as the 21064 and R4400, the two-cycle cache latency significantly decreases performance in many situations. But the PA-8000's out-of-order design helps reduce the impact of this latency, particularly on the data side.

If a load instruction is followed immediately by an instruction that uses the data from the load, the latter instruction will stall for two cycles. During this time, however, other instructions, including loads and stores, can continue to execute. The entire processor will stall

only if there are not enough independent instructions in the stream to fill the gap.

HP's Manglesdorf estimates that the extra data-cache latency reduces performance by less than 5% compared with a hypothetical single-cycle cache. Such a cache could not be implemented with external SRAM, however, forcing the cache to be on-chip and much smaller, or else causing a reduction in the clock speed of the processor. The benefits of the large primary caches and the faster clock speed more than outweigh the 5% loss of efficiency.

Branch Prediction Avoids Penalties

On the instruction side, the longer latency extends the mispredicted branch penalty to five cycles. Even correctly predicted taken branches can cause a two-cycle bubble in the fetch stream. The PA-8000 implements several features to reduce the impact of these penalties.

To avoid mispredicted branches, the processor uses a 256 × 3-bit branch history table (BHT). HP does not use the two-bit Smith algorithm shared by most other next-generation processors, instead opting to store the results of the last three iterations of each branch. The prediction is then based on a majority vote of the three bits. This algorithm offers a similar level of hysteresis and accuracy as the two-bit algorithm, but it is easier to update the BHT, as the processor must simply shift in the new result rather than perform a read-modify-write operation.

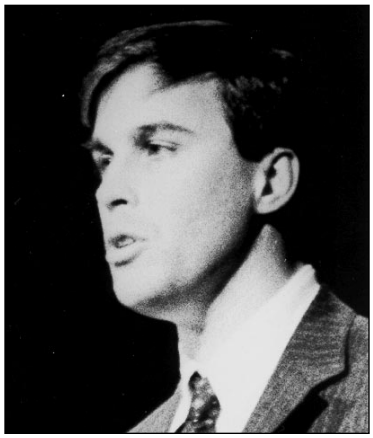
We expect the PA-8000 to achieve about 80% prediction accuracy on SPECint92, although accuracy will be lower on most real applications. The company believes that feedback-directed compilation techniques can achieve even better accuracy in some cases. The PA-8000 implements new branch instructions with a predict bit (similar to PowerPC) that allows the compiler to indicate the expected branch direction.

Unlike PowerPC, however, the PA-8000 allows the predict bit to override the dynamic branch prediction. Each TLB entry includes a bit that enables static branch prediction, disabling the BHT on a page-by-page basis. Thus, recompiled programs can use static prediction, if it is more accurate, while older programs (or libraries) can default to dynamic prediction.

To avoid the taken-branch penalty, the PA-8000 includes a branch target address cache (BTAC) similar to that in the PowerPC 620 *(see **081402.PDF**)*. This fully associative 32-entry structure is accessed along with the instruction cache but responds in a single cycle. If a branch instruction is predicted taken and is found in the BTAC, the predicted target address is issued to the instruction cache immediately, allowing zero-cycle branching in spite of the extended latency of the external instruction cache.

To improve the hit rate, the processor tries to keep only predicted-taken branches in the BTAC. If a branch hits in the BTAC but is predicted to be not taken, that entry is deleted. If a branch is predicted taken but doesn't hit in the BTAC, its target address is added to the BTAC; in this situation, a two-cycle bubble is created, because the fetch stream cannot be redirected until the branch is received from the cache and decoded. If there are enough decoded instructions already queued, some or all of this latency may be hidden.

Although these techniques reduce the impact of the instruction cache latency, they will be most effective on code with small numbers of predictable branches. HP's BHT is much smaller than those in other next-generation RISC chips, and its BTAC is one-eighth the size of the 620's. Furthermore, both the 620 and R10000 have much shorter misprediction penalties than the PA-8000, allowing them to better handle branches that are difficult to predict. In general, commercial applications have more branches with less predictability than technical code; HP says it will rely on its feedback-directed compilers to improve commercial performance, counteracting the PA-8000's relative shortcomings in branch handling.



CLARENCE TOWERS

"I expect that, when the PA-8000 begins shipping in systems, it will outperform all competitive RISC systems at that time."
Steve Manglesdorf, HP

Flat Caches Better for Large Programs

The advantage of the off-chip cache design is that the primary caches can be 1M or more, much larger than the 32K on-chip caches used by the R10000 and the 620. Both these competitors require at least six cycles to access their second-level caches, three times the latency of the PA-8000's primary caches, which are approximately the same size.

Because the competitors have single-cycle primary caches, programs that hit in their on-chip caches at least 80% of the time will have roughly the same average latency as in a PA-8000 system. Programs with small working sets, including most benchmarks, will perform well with a 32K on-chip cache. Many real applications, however, have less than an 80% hit rate with such small caches; these programs will do better on the HP design.

The flat cache hierarchy of the PA-8000 also eliminates the overhead associated with a two-level cache design. When an access misses the small primary on-chip cache of most processors, the processor must refill that cache from the L2 cache. This refill typically blocks the

on-chip cache for four cycles. In the PA-8000, the primary cache miss rate is much lower, greatly reducing the number of cycles that refills block the cache.

Manglesdorf pointed out that removing the cache from the die also frees space for additional features. The 620's on-chip caches, for example, occupy 90 mm^2, nearly one-third of the die. He did not reveal the die area of the PA-8000; it may be smaller than other next-generation chips due to the lack of on-chip cache.

As a system vendor, HP can trade off the cost savings of a smaller CPU die against the cost of building fast external caches, which will be more expensive than the slower L2 caches typically used by other processors. These caches will be particularly expensive when the PA-8000 is running at its maximum speed, where the SRAMs carry a significant price premium. This premium is smaller at lower speeds, however, allowing HP to hit a range of price/performance points by varying the speed of the CPU and its primary caches.

The primary caches are direct mapped to avoid the timing and pinout problems of external set-associative caches. Unlike its predecessor, the PA-7200 *(see **080302.PDF**)*, the new design includes no on-chip buffer to avoid cache thrashing. If the processor encounters a string of alternating memory references that map to the same cache line, it will tend to group accesses to each location, issuing several accesses to one location while the other is being fetched from memory.

Large Window for Out-of-Order Execution

Four instructions are received from the instruction cache per cycle. As in the PA-7200, each instruction includes five predecode bits, speeding the decode process. The decoded instructions are then placed in the instruction queue (I-Q), as Figure 1 shows. This queue is conceptually similar to the K5's reorder buffer *(see **081401.PDF**)* but much larger. All active instructions are held in the I-Q until they are retired. Each entry in the I-Q also contains a location to store the result of its instruction, implementing register renaming as well as enabling speculative and out-of-order execution.

The I-Q is physically implemented as two separate structures: a compute queue and a load/store queue. Each of these subqueues has 28 entries, yielding a total of 56 instructions that can be active at any given time; logically, the two structures function as a single queue. Loads and stores require that additional information (such as the memory address) be stored; by separating them from simpler computational instructions, the total storage requirement is reduced.

Each cycle, the processor issues up to two instructions to the address units and an additional two instructions to the computation units (FPUs and integer ALUs). If there are more than two executable instructions of a given type, the oldest instructions receive priority and are issued to the function units. As in the R10000, there are no reservation stations in the function units; instructions are not issued until they are ready to be executed.

Once instructions are issued, they must fetch their operands. Because of the register renaming that occurs in the I-Q, the desired operands may be found in either the register file or in the I-Q. For each register request, the I-Q must perform an associative lookup to find the desired register value. In some cases, a logical register may appear more than once in the I-Q; in this situation, the I-Q must deliver the value that was generated most recently before the instruction that is being executed. Thus, simply acquiring the operands is a complex process that must occur in less than a single clock cycle.

Once instructions are executed by the function units, their results are written to the rename registers and made available to other pending instructions. Instructions are retired once all preceding instructions have successfully completed, maintaining an in-order programming model with precise exceptions. Up to four instructions can be retired per cycle. If an exception or mispredicted branch occurs, all subsequent instructions in the I-Q can be invalidated in a single cycle.

Fast Floating-Point Units

Unlike other next-generation chips, the PA-8000 provides two of everything, avoiding almost all resource conflicts in the issue process. The dual address units

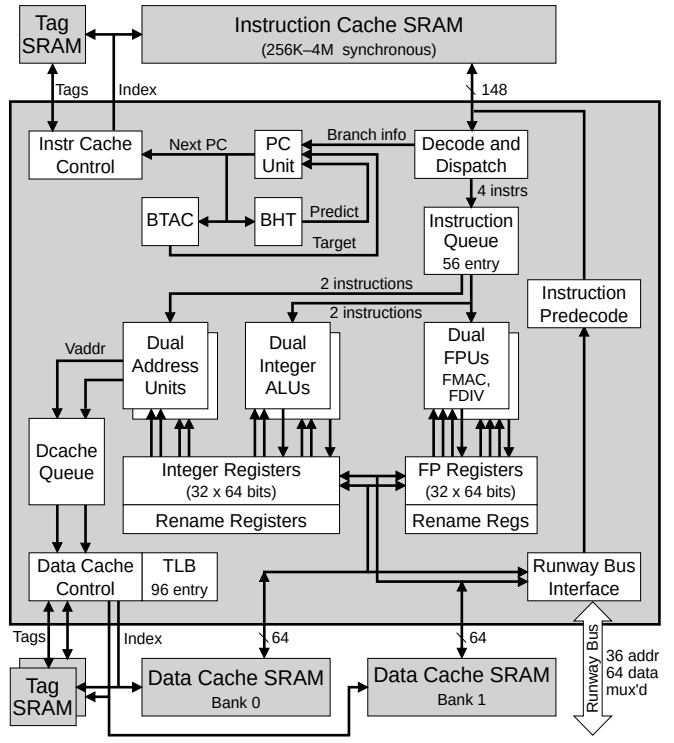


Figure 1. The PA-8000 uses a decoupled microarchitecture with dual load/store pipelines and large primary caches that are external to the processor.

each generate one physical address per cycle; these addresses are then routed to the dual-ported memory system. The dual integer ALUs are truly duplicated; each has a complete shift-merge unit. As in previous PA-RISC chips, integer multiplication is handled in the FPUs.

The dual FPUs implement a multiply-accumulate ($A \times B + C \rightarrow C$) architecture. The new multiply-accumulate (FPMAC) instruction has a latency of just three cycles. As Figure 2 shows, there is no intermediate rounding step between the multiply and the add, saving one cycle and improving the accuracy of the operation. If only FPMAC instructions are used, the PA-8000 can sustain four flops per cycle, twice the rate of competitive designs.

Individual multiply or add instructions use the same pipeline, yielding a three-cycle latency. Single- and double-precision calculations have the same latency. The PA-8000 also supports the existing FPMPYADD instruction, which calculates $A \times B \rightarrow C$; $D + E \rightarrow E$. Using both FPUs in parallel, this instruction also has a latency of three cycles.

Both FPUs contain a divide/square-root unit as well. Although only one instruction per cycle can be issued to each FPU due to register port limitations, these long-latency operations can execute in parallel with multiply-accumulate operations.

Two Memory Accesses Per Cycle

Like the R10000, the PA-8000 queues load and store requests with their addresses. Loads can execute speculatively and out of order; store data is sent to the cache only when the store instruction is retired. Up to two loads or stores can be executed in a single cycle.

The primary data cache is dual ported, as Figure 1 shows. The cache tags are fully duplicated to allow independent accesses. The cache data is arranged in two interleaved banks to avoid the cost of duplication; thus, two accesses can be paired only if they are to different banks. Fortunately, the out-of-order design facilitates pairing: two accesses can always be made as long as there are two instructions anywhere in the load/store queue that use different banks.

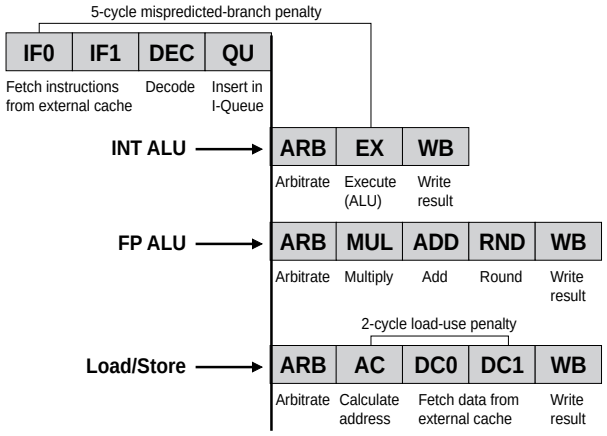


Figure 2. The PA-8000 has separate pipelines for the individual function units. The instruction-fetch sequence is extended by the two-cycle access to the external primary cache.

The 96-entry main TLB is fully associative and also dual ported, translating two addresses per cycle. (The instruction unit contains a four-entry micro-TLB.) As Figure 2 shows, the latency of the cache access creates a two-cycle load-use penalty. With its pipelined design, the cache can sustain two accesses per cycle.

System Interface Same As PA-7200

Rather than jump to a new system bus, the PA-8000 sticks with the Runway bus pioneered by its predecessor, the PA-7200. This bus supports split transactions and glueless multiprocessing, offering a sustainable throughput of 768 Mbytes/s, similar to that of other next-generation processors.

Instead of moving to a wide 128-bit bus, however, Runway maintains high bandwidth by operating a 64-bit multiplexed bus at 120 MHz, much faster than competing solutions. This clock rate will make it challenging to design multiprocessor systems, but HP is known for its ability to design at high frequencies. Since few other vendors will use the PA-8000, the difficulty of design is not a major issue for HP.

A more challenging design task will be to move data between the synchronous cache and the processor at speeds approaching 200 MHz. To improve the electrical environment, HP will use flip-chip technology *(see **071304.PDF**)* to mount the die directly to the ceramic carrier, eliminating the discontinuities associated with bond wires. The cache signals use GTL interfaces to reduce the size of the signal swings. The SRAMs will use BGA packages, allowing them to be mounted closer to the CPU and reducing the trace length. These techniques should allow the caches to operate at high frequencies. Future versions may use a multichip module (MCM) for the processor and cache subsystem.

The processor will be built in a 0.5-micron four-layer-metal CMOS process and operate at 3.3 V. This process, called CMOS-14C, is a 10% gate shrink of HP's CMOS-14 process *(see **080504.PDF**)*, which is used for the PA-7200. To accommodate the complex design, the die area is likely to be around 250 mm^2, somewhat less than other next-generation processors due to the lack of on-chip cache.

The package size will almost certainly set a record: with one 128-bit interface and three 64-bit buses—plus pins for the external cache tags, control signals, power, and ground—the chip will require about 700 pins. The flip-chip process allows the pads for these signals to be spread across the surface of the die, reducing die size by avoiding the need for a huge pad ring.

PA-RISC 2.0 Moves to 64 Bits

The PA-8000 implements PA-RISC 2.0, the first major revision of the HP architecture since it debuted in 1986. The new architecture is a full 64-bit design, making HP the last of the major RISCs to make this leap. Applications will have access to a flat 64-bit virtual address space. For compatibility with older code, PA 2.0 retains the 32-bit "space ID" registers of earlier versions; these may be concatenated with the linear address to create a 96-bit address space. Let's see the software to fill that!

Floating-point enhancements include the FPMAC instruction mentioned earlier. Current HP processors can have a bottleneck on floating-point compares, all of which write their result to a single bit (the C-bit) in the FP status register. The new architecture incorporates multiple C-bits to avoid this problem.

Most of the other enhancements are relatively minor; these will be disclosed in more detail in the future, according to the company. The PA-8000 does include two enhancements pioneered in the PA-7100LC: it supports multimedia data types and can operate in either big- or little-endian modes.

Why Trade Complexity for Clock Speed?

Historically, CPU designs have been forced to trade off clock speed for complexity *(see **0703ED.PDF**)*. This tradeoff has generally favored the Speed Demons, and HP has consistently been in this camp. With the PA-8000 design, however, it appears that HP has taken a sharp left turn into the Brainiac domain.

HP's Manglesdorf, however, argues that the apparent complexity of the PA-8000 is not a limiting factor to clock speed. Neither the decoupled design nor the plethora of function units reduces the clock speed. Where complex or lengthy tasks are required, HP has added extra pipeline stages, relying on accurate branch prediction and out-of-order execution to reduce the impact of these extended latencies.

This argument is also made by the designers of the R10000, who feel that 200 MHz will not be a problem for their chip, which uses an IC process similar to HP's. If the PA-8000 can achieve a similar clock speed, it would indicate that the complexity of these decoupled designs does not adversely affect clock speed.

For More Information

HP does not sell its processors on the merchant market. For more information about the PA-8000, contact HP at 408.447.4747; fax 408.447.7983.

For the PA-8000 to achieve its goal of outperforming all other microprocessors in its class, it will need to deliver 350–400 SPECint92. To achieve this performance, the processor will have to deliver between 1.75 and 2.0 SPECint92/MHz at 200 MHz; by comparison, the R10000, 620, and UltraSparc are all rated at 1.5 to 1.7 SPECint92/MHz.

The PA-8000's dispatch rate and set of function units are similar to those of these other processors, but none of these competitors has dual load/store pipes or such a large window for out-of-order execution. In addition, the HP chip's large primary caches should give it a performance advantage. The chip's only drawback will be the longer latencies for branches and cache accesses, but it will likely be more efficient (when measured in SPECint92/MHz) than its competitors. The large caches and dual FP MAC units should generate leadership floating-point performance.

On integer performance, HP's toughest competition will come from Digital's 21164 *(see **081201.PDF**)*, the king of the Speed Demons. A 300-MHz version of this part, rated at 330 SPECint92, is due about one year sooner than the PA-8000. By the time the HP processor ships, Digital is likely to have parts approaching 400 SPECint92. It will be difficult for the HP design to achieve an efficiency great enough to outperform a chip running at twice the clock rate of the PA-8000.

Due to its 1M primary caches, the PA-8000 will excel on real applications, which typically have poorer locality than the SPECint92 benchmarks. The 21164, in contrast, has tiny 8K primaries along with a 96K secondary cache on chip. On the forthcoming SPECint95 suite, which will use larger benchmarks than the current suite, the Digital design may not fare as well. HP is typically conservative in its public statements, but we must wait to see if the PA-8000 meets its design goals and makes Manglesdorf's aggressive prediction come true. ♦

