The Linley Group

MICROPROCESSOR REPORT

www.MPRonline.com

THE INSIDER'S GUIDE TO MICROPROCESSOR HARDWARE

DRAM+CPU Hybrid Breaks Barriers

Radical Chip Design Slashes Power Consumption, Boosts Memory Bandwidth

By Tom R. Halfhill {12/27/10-02}

Today's high-performance microprocessors are mostly memory, not logic. Of the 774 million transistors in an Intel Core i7-860 processor, for example, about 69% are SRAM transistors in the 8MB L3 cache. The balance weighs even more heavily toward memory in server processors with larger caches.

This wasn't always so. Thirty years ago, microprocessors didn't need caches, because DRAMs and even mask ROMs were fast enough to keep up with contemporary processors. Over time, processors outran external memory, prompting the integration of on-chip SRAM to cache frequently used instructions and data. Today, paradoxically, big caches are promoted as a feature, even though they inflate manufacturing costs, gulp power, and wouldn't exist if external memory was fast enough. Simply put, caches are kludges.

Now, a Texas-based startup, Venray Technology, is bucking the trend toward bigger caches—and the march toward bigger CPUs, too. Instead of building expensive six-transistor (6T) or eight-transistor (8T) SRAM cells in a logic process to accommodate the processor, Venray is moving the processor to commodity-DRAM processes, whose 1T memory cells are cheaper to manufacture and less leaky. Merging the CPU with DRAM dramatically boosts memory bandwidth, reduces memory latency, and slashes power consumption by eliminating caches and shortening the CPU-memory interface.

Venray is trying to exploit both semiconductor technology and semiconductor-industry economics. Commodity DRAM is ridiculously cheap: a 1Gb DRAM chip with one billion transistors costs about $1, whereas an Intel processor with the same number of transistors can cost $200 or much more. DRAM is more power efficient, too. Bit cells can't tolerate much current leakage without losing data, so DRAM transistors are built to leak less power than logic transistors. To prove its technology, Venray has designed a DRAM+CPU prototype, shown in Figure 1.

This isn't the first attempt to integrate microprocessors with memory. The most notable recent experiment was the Intelligent RAM (IRAM) project at the University of California at Berkeley, led by RISC pioneer David A. Patterson (see *MPR 3/9/98-04*, "New Processor Paradigm: V-IRAM"). Even earlier, Mitsubishi integrated a 32-bit microprocessor with SDRAM to make the R32R/D hybrid chip (see *MPR 5/27/96-02*, "Mitsubishi Mixes Microprocessor, Memory"). For various reasons, neither IRAM nor the M32R/D significantly altered the industry's course.

Venray hopes to succeed with a different approach that makes CPU+DRAM integration much more intimate, potentially revolutionizing microprocessor design. As is usually the case, Venray makes several tradeoffs, including slower transistor switching, less sophisticated CPUs, and perhaps a useless surplus of memory bandwidth. But whether the venture succeeds or fails, an eventual merger of microprocessors and memory seems inevitable. In time, Venray's compromises may become easier to swallow than the alternatives.

Overcoming Patterson's Walls

Venray's prime mover is Chief Technology Officer Russell Fish III, a veteran engineer variously described as brilliant and eccentric. He's most famous for his 1989 patent (U.S. 5,809,336) on the variable-speed on-chip system clock, or "Fish clock," now commonly used in microprocessors. He is also known for co-designing the ShBoom microprocessor (see *MPR 4/15/96-01*, "New Embedded CPU Goes ShBoom"), for setting a world skydiving record, for creating the first Internet sex-offender registry, and for building schools for poor children in Africa. His latest passion is the Thread-Oriented MIcroprocessor (TOMI), which melds CPUs with DRAM.

Fish's goal is to overcome three walls blocking the progress of microprocessor design, as first described by David Patterson. (Editor's note: Patterson is a recent addition to the *Microprocessor Report* editorial board; he reviewed this article in draft but did not participate in its writing.) To summarize, Patterson's walls are power consumption (engineers can design bigger CPUs than systems can afford to use); memory I/O (processors are outrunning memory latency and bandwidth); and instruction-level parallelism (which is severely limited by diminishing returns). Attacking any one of these problems often makes another worse.

Venray is mounting a frontal attack on the memory-I/O and power walls, reinforced by a flanking maneuver around instruction-level parallelism. By tightly integrating many small CPUs with DRAM, Venray makes huge leaps in memory bandwidth and latency, minimizing the need for caches and their millions of transistors. By simultaneously reducing CPU complexity, Venray eliminates millions more transistors, reducing dynamic and static power consumption. In addition, a chip with lots of small CPUs lends itself to data-level parallelism, the increasingly popular alternative to instruction-level parallelism.

Reducing power and improving memory I/O are the main goals. External memory isn't keeping up with CPUs for two reasons—one minor and one major. The minor reason is that DRAM needs lower-leakage transistors to increase retention time, so they're built smaller and have a higher voltage threshold (V_t) for switching states. They leak much less static current than a logic transistor of equal size, but their switching speed is slower.

The major reason for memory's lagging speed, however, is that external memory is *external.* There is far too much of it to integrate with the processor. To access memory, the CPU must reach through a lengthy I/O interface to chips located elsewhere in the system. The CPU must drive current through its I/O pins, traverse the board traces, and penetrate the memory chip's I/O pins. All that external wiring has resistance and capacitance that must charge and discharge for each memory transaction. Adding even more capacitance is the electrostatic-discharge protection built into external I/O interfaces. Overcoming all that capacitance requires large drive transistors and lots of power—hence the need for caches, which try to minimize off-chip I/O by storing frequently used instructions and data in fast SRAM on the CPU chip.

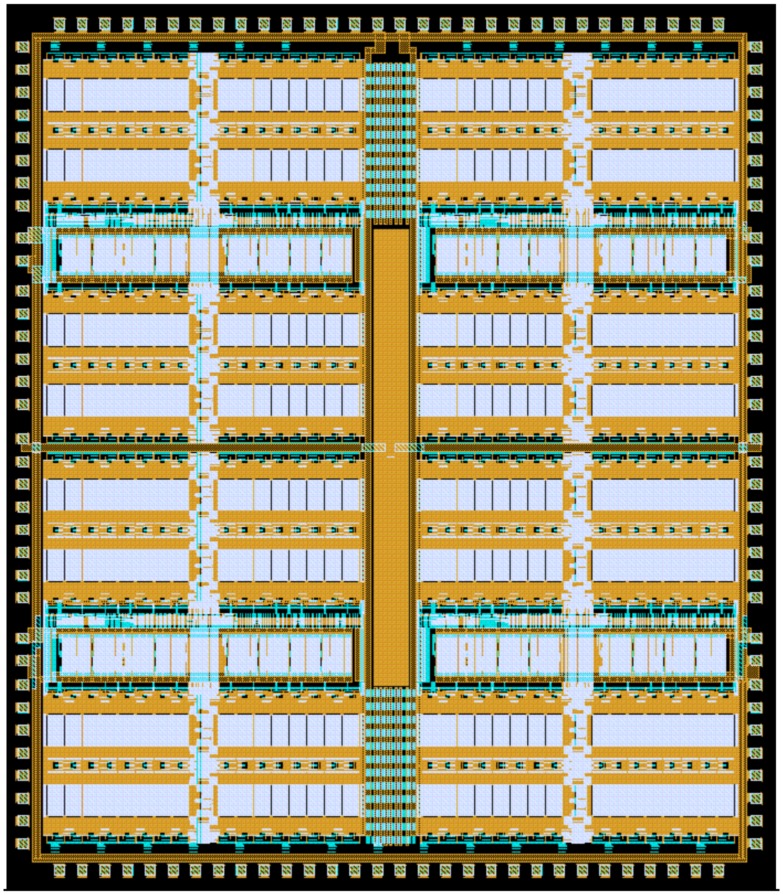

Figure 1. Venray's Aurora test chip. If this die plot looks like a DRAM, that's because it is a DRAM. It's a 64Mb memory chip designed for a commodity-DRAM process, not a logic process. Venray has added four CPU cores, which are visible in the center of each quadrant. They occupy 20% of the 5.6mm-by-6.7mm die. So far, the finished design exists only in simulation, not in silicon. (Source: Venray)

Patterson's aforementioned IRAM project merged the microprocessor and main memory on a single chip, but it differed in two important respects from Venray's approach. First, instead of using a commodity-DRAM process, IRAM used a conventional logic process with embedded DRAM (eDRAM). Although eDRAM has economical 1T bit cells, those cells are neither as dense nor as power efficient as conventional 1T-DRAM cells built in a commodity-memory process.

The second difference is that IRAM used a conventional MIPS-compatible CPU core (with new vector extensions) that wasn't as tightly integrated with memory as Venray's TOMI processor is. TOMI is designed specifically for DRAM integration and hooks directly into the memory arrays. It doesn't talk to DRAM through a memory controller because, in effect, the CPU becomes the controller.

The 16,384-Bit Memory Bus

Venray's strategy has multiple benefits. Because the CPU jacks straight into the DRAM's row-and-column decode logic, the internal memory interface is the width of the row, which varies with DRAM density. On Venray's Aurora prototype, memory rows are 16,384 bits wide, so the CPU's internal memory interface is the same width. Therefore, each CPU can load 16,384 bits (2KB) per row-address select (RAS) cycle.

Before translating this capacity into raw bandwidth, keep in mind that Venray designed Aurora for an ancient 110nm DRAM process—four generations behind the newest 32nm DRAM technology. Although the obsolete process demonstrates that improving memory I/O doesn't require leading-edge fabrication technology, it was not Venray's first choice. Every major DRAM manufacturer approached by the startup declined to make the test chip. (Venray says DRAM vendors can't afford to acquire its technology and won't license their own intellectual property.) The only alternative was an obscure Taiwanese company (Elite Semiconductor Memory Technology), which permitted Venray to design its prototype on a small 64Mb DRAM intended for 110nm fabrication. Since then, however, ESMT has turned toward flash memory, so Aurora will probably never reach silicon.

In this 110nm DRAM process, the RAS cycle time is 55ns, or 18.1MHz. (The CPU's core clock frequency is 500MHz.) RAS times haven't improved much in 10 years, but memory density has increased greatly, and wider rows translate into wider memory interfaces for Venray. Even so, the 16,384-bit-wide interface in this old 64Mb DRAM transfers 2KB during each RAS cycle, so memory bandwidth is an impressive 37.2GB/s per CPU. The Aurora prototype has four CPUs per chip; thus, their aggregate memory bandwidth is an astonishing 148.9GB/s.

In comparison, Intel's fastest "Gulftown" Xeon X5677 server processor has three 64-bit DDR3-1333 interfaces that provide only 32GB/s of aggregate memory bandwidth to four CPU cores. Intel builds these chips in a state-of-the-art 32nm high-*k* metal-gate logic process. The CPU clock frequency is 3.2GHz, and the DRAM interface runs at a base clock frequency of 666MHz (1.3GT/s, double data rate). Yet the Xeon chip still needs 1MB of L2 cache and 12MB of L3 cache to avoid outrunning main memory. Venray's prototype can deliver 4.6 times the bandwidth of Intel's best server processor despite using 10-year-old DRAM technology.

Only Milliwatts per CPU

Power reduction is dramatic, too. Because Venray's TOMI CPU core is so tightly coupled to memory, the 16,384-bit interface is just a row of extremely short wires. The interface doesn't need powerful drive transistors to overcome the capacitance of I/O pins and board traces, nor does the CPU need enormous SRAM caches to compensate for slow off-chip DRAM. According to Venray's simulations, dynamic power is a mere 23mW per CPU when running full bore at 500MHz. Static leakage is only 107 microwatts per core.

TOMI isn't completely cacheless. As the block diagram in Figure 2 shows, the CPU has four caches, but they're tiny. Each one is only 512 bytes—just enough to hold one-fourth of the 16,384 bits transferred per RAS cycle. These caches give the DRAM array sufficient time to precharge and load for the next RAS cycle. They also allow the TOMI processor to enter a low-power hibernation mode in which it remains operational at 12MHz while living on cached instructions and data. In this slow mode, dynamic power plunges to just 350 microwatts. Most other CPUs would be comatose at that level.

Cache sizes may vary in different implementations, depending on the DRAM's native capacity and other factors. In any case, the TOMI caches are extremely compact—not just because they hold only 512 bytes, but also because they are built like DRAM sense amps rather than costly SRAM bit cells. And SRAM is becoming costlier relative to DRAM as process geometries continue shrinking. Instead of using the once-common 6T-SRAM bit cells, Intel's recent processors use 8T cells to resist the soft errors that become more troublesome at smaller dimensions and lower voltages.

During Venray's software tests with Aurora in simulation, the cache-hit rate exceeded 99% with most programs. When a memory request does miss the cache, a complete refill takes the same amount of time as a partial refill. In effect, the entire cache is a single line.

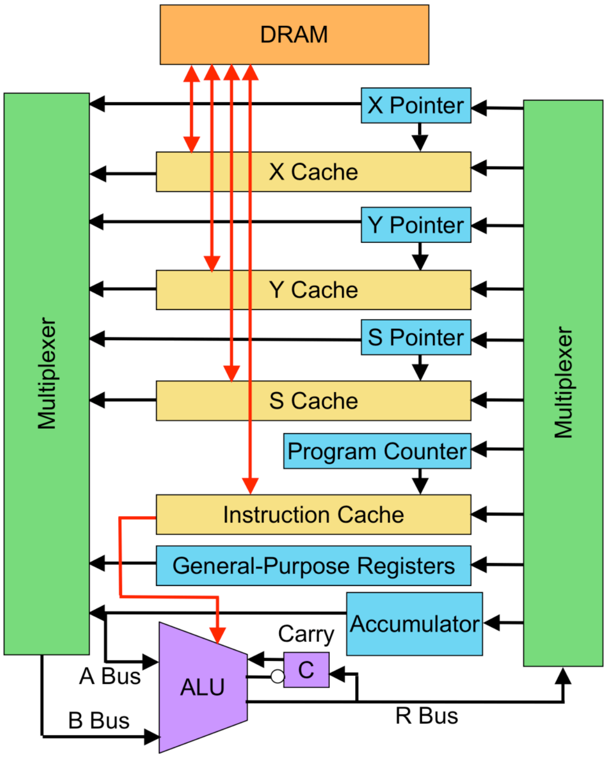

Figure 2. Thread-Oriented MIcroprocessor (TOMI) block diagram. Venray designed this 32-bit CPU core specifically for DRAM integration. Four 512-byte caches hook directly into the DRAM array, each with its own 4,096-bit memory interface. Each cache can refill in a 55ns (18.1MHz) clock cycle. Cache latency is 660 picoseconds.

Return to RISC

Venray's TOMI architecture is unusual in other respects as well. Although it's a 32-bit processor, instructions are only 8 bits long, and the entire instruction set has only about 30 basic operations. (An even simpler architecture described in Fish's 2007 patent application—which is still pending—has only seven instructions.) TOMI instructions are short because they need fewer register-address bits than modern instruction sets do. Although TOMI has 32 general-purpose registers, many instructions use immediate operands or work directly on memory as read-modify-write operations. Dual-operand instructions read one value from a source register and another from a 32-bit accumulator then destructively store the result in the source.

Of course, the drawback of short instructions and frugal register addressing is that a program often uses more instructions to do something that a single 32-bit instruction can do. On the other hand, 8-bit instructions allow the CPU to completely fill the 512-byte instruction cache in a single RAS cycle, and program code requires much less memory.

Venray designed this small CPU to remain small even when ported to fabrication processes that are newer than the gray-haired 110nm process for which Aurora was designed. Instead of spending bigger transistor budgets to inflate the processor's complexity, Venray favors adding more of these simple processors. Each CPU can execute one thread, making up for its simplicity by working with other CPUs en masse (see *MPR 3/31/08-01*, "Think Parallel").

Some of these processors could be designed for special purposes, such as floating-point math (omitted from the TOMI architecture), graphics, or media processing. In any case, the chips will probably have one processor core per DRAM memory bank. Aurora is designed for a 64Mb DRAM with four 16Mb memory banks, so it has four CPUs. If a larger 1Gb DRAM has eight 128Mb banks, it will probably have eight CPUs. A 64-bit local bus connects the CPUs together (a potential bottleneck, if the CPUs must share lots of data).

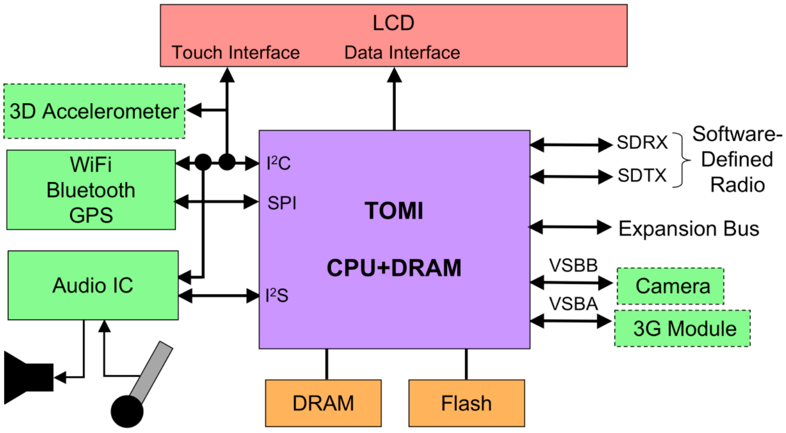

Figure 3. Block diagram of Venray's "Shirtbook" tablet. Although this conceptual system design is built around a DRAM chip with integrated TOMI processors, it still requires external memory and additional chips to perform functions normally integrated in an SoC. Only systems with lesser memory requirements could get by on internal memory alone.

As with CPUs built in logic processes, CPUs built in DRAM processes will shrink with each process generation, assuming similar gate counts. Whereas the four TOMI CPUs in Aurora occupy 20% of the 64Mb memory chip, Venray estimates that eight CPUs would occupy only about 7% of a 1Gb DRAM.

Porting the CPU to a different DRAM process requires five steps: characterizing Venray's digital and analog cell library in the new process; scaling the CPU's caches to match the pitch of the DRAM's sense amps; simulating the CPU, caches, and clocking circuits to establish performance and power requirements; placing the CPU block and interprocessor bus block; and, finally, full-chip simulation and verification. These steps aren't too different from porting any CPU to a logic process, but they do require some special tools and cell libraries that Venray is developing.

Now for the Downsides

Any solution to a problem that stumps the world's best engineers must have some drawbacks and tradeoffs, and Venray's solution is no exception. There's no such thing as a free lunch.

One drawback is that integrating CPUs with DRAM doesn't completely eliminate the need for external memory. The Aurora prototype has only 8MB of internal memory, which is sufficient for many embedded applications but not enough for a PC or server. Of course, it's only one chip, and it's built in turn-of-the-century technology. Using more chips of greater density adds much more memory—and more CPUs, too. Using today's 1Gb DRAMs, a single chip would have 128MB of internal memory and eight CPUs.

As Figure 3 shows, however, even Venray's conceptual design for a tablet computer relies on external DRAM and flash memory, requiring the addition of external memory interfaces to the hybrid DRAM. In effect, the integrated memory would function as an L2 cache for the slower external memory. Indeed, the 1T-DRAM bit cells would be more economical than the 6T or 8T SRAMs typically found in CPU caches. Nevertheless, the conceptual system design shows that Venray isn't agitating for a complete overthrow of hierarchical memory systems.

Another shortcoming is that the tablet needs additional integration or additional chips to perform functions normally done with an SoC. Adding peripheral functions and I/O interfaces to the TOMI chip, as shown in the figure, requires building more logic in the less-efficient memory process. Some application functions, such as cryptography, must either be offloaded to separate chips or executed by the relatively slow CPUs.

Another tradeoff is CPU clock frequency. Venray saves power by building the CPUs in a commodity-DRAM process with higher-threshold transistors, but those transistors can't match the switching speed of transistors built in a logic process of the same geometry. Venray hopes to compensate by keeping the CPU small and efficient and by taking advantage of the leap in memory performance. In Aurora, the TOMI CPU runs at only 500MHz because it's limited by the old 110nm process; in a more modern process, clock speeds exceeding 1.0GHz should be feasible. In any event, compromising on clock frequency shouldn't seriously disrupt the industry, because users have already accepted lower frequencies to control power consumption.

Coping With CPU Deflation

A bigger tradeoff is CPU-design flexibility. Here's where Venray seriously bucks long-term trends. Commodity-DRAM processes typically have only three metal layers instead of the half-dozen or more such layers commonly found in today's logic processes. With so little metal available for wiring, place-and-route tools have fewer options, so any CPU core implemented in a DRAM process must be small and simple. For Venray, that's a virtue; for most CPU architects, it's a showstopper.

Venray's TOMI CPU has only 18,600 logic gates—fewer than half as many as ARM's 15-year-old ARM7 TDMI, which was considered a very small processor even in 1995. The TOMI architecture is a throwback to the 1980s, and not just because it has tiny caches. It also has a tiny 8-bit instruction set that makes ARM's 16- and 32-bit RISC instructions seem luxurious.

Although TOMI is a true general-purpose processor capable of doing anything that any other processor can do, reverting to such frugal CPUs would be a disruptive change for the computer industry. CPU architects accustomed to rich transistor budgets may find their skills less valuable in this new age of austerity. Maybe they'll find employment as programmers, who would have to port today's high-level languages and application programming interfaces (APIs) to architectures with fewer instructions than a Zilog Z80.

It's not impossible. Venray has already ported the Gnu C tool chain (GCC), an MP3 audio decoder, an H.264 video decoder, GIF/JPEG still-image decompressors, and FFT routines for signal processing. Using an early version of the ported GCC compiler and an instruction-level simulator (not cycle accurate), Venray benchmarked a 500MHz TOMI against a 400MHz ARM11. (The ARM11 was in a Texas Instruments OMAP2420 application processor.) When decoding an MP3 audio file or decompressing a 3.1-megapixel JPEG image, the ARM11 was 2.3 times faster. In another test, the ARM11 was 1.5 times faster at alpha-blending two images together.

To compensate for TOMI's simplicity, programmers can use more processors. As currently envisioned, a DRAM chip will have one CPU for each memory bank. Unfortunately, some programs can't usefully exploit multiple processors, or they require extensive refactoring to do so. And adding more than one CPU per memory bank is difficult, because the bit pitch of the DRAM cells limits the room available for wiring. This limitation is in addition to the restriction of having only three metal layers for routing.

Adapting Other CPUs Isn't Easy

Venray's TOMI architecture isn't integral to the concept of merging CPUs with DRAM. Venray optimized TOMI to fit in the limited space available and to exploit Aurora's 16,384-bit-wide memory interface. Any CPU architecture reducible to a similar size should work as well. In theory, the CPU core could be an ARM, MIPS, Power, SPARC, or x86 processor.

Indeed, some of those familiar architectures are old enough to have early implementations with fewer than 20,000 gates. Even today, crafting a stripped-down design while sacrificing little or no software compatibility may be possible, but it wouldn't be easy and would be an abrupt shift into reverse gear for an industry accustomed to doing more with more, not more with less. The physical restrictions imposed by DRAM will deter designers from applying Venray's technology to standard CPU architectures.

Upgrading DRAM processes to add more metal would ease the routing problem. Perhaps engineers can find a work-around for the bit-pitch limitation, too. These solutions, however, would inflate costs and prevent DRAM manufacturers from stamping out the chips on the same cookie-cutter production lines as their commodity memory chips. Memory is much cheaper than logic not just because triple-metal fabrication is simpler, but mainly because the DRAM industry's business model is based on razor-thin profit margins and rapid fab amortization.

Although one might think that DRAM manufacturers would welcome a higher-value business model that supports richer margins, such a fundamental change would be as disruptive as requiring CPU architects to design 18,000-gate processors. Venray's inability to sell its technology to DRAM manufacturers or even convince them to license their intellectual property for the purpose of making a test chip speaks volumes. DRAM analyst Jim Handy of Objective Analysis notes that memory vendors live a world apart from microprocessor vendors and are unwilling to gamble on a radical new technology that integrates logic, no matter how promising it looks. Even Handy, who closely tracks the DRAM industry, was unfamiliar with the Taiwanese company that Venray was forced to employ for a 10-year-old memory process.

For all these reasons, if Venray does succeed in selling its technology, the buyer will probably have to adopt the TOMI architecture or design an equally simple CPU—then start porting software. These obstacles will likely restrict the technology to narrow embedded applications, barring a revolutionary overthrow of the industry's ruling CPU architectures.

For More Information

Venray Technology has no plans to sell chips or broadly license its technology. The company seeks a buyer to acquire all rights and blaze its own path to market. The Aurora test chip—a 64Mb DRAM with four Thread-Oriented MIcroprocessor (TOMI) cores—is a finished design but exists only in simulation. For more information about Venray and Aurora, access www.venraytechnology.com.

For more information about the Berkeley Intelligent RAM (IRAM) project, point your browser to http://iram.cs.berkeley.edu.

Rebalancing the Bottlenecks

Venray is rebelling against historical trends in many ways, but it is following the most important trend: greater integration. In the most recent example of this evolution, AMD and Intel are merging GPUs with CPUs (see MPR 12/6/10-01, "AMD's Fusion Finally Arrives"). Ironically, however, Venray has replaced one bottleneck with another. Whereas today's muscular processors are often limited by memory, Aurora has an embarrassing surplus of memory bandwidth mismatched with puny processors. Better CPUs might rebalance the design if Venray can work around the limitations of commodity DRAM.

Until fabrication technology improves enough to combine a reasonably fast CPU with a competitively sized DRAM, or to combine a reasonably large DRAM with a competitively fast CPU, this integration makes sense only for some specialized applications. Today, microcontroller-based systems that require no external memory are common. Many other embedded systems need only one memory chip. The latter designs are candidates for CPU+DRAM integration, but only if the integrated chip can duplicate the performance and functions of the two-chip solution. Previous CPU+DRAM experiments have failed either because they cripple the CPU (by reducing performance) or because they cripple the DRAM (by reducing capacity).

One alternative to chip-level integration is package-level integration—either multichip modules that bond the chips side by side or vertical (3D) chip stacking. The most recent example of the tandem approach is Intel's Atom E600C ("Stellarton"), which packages an Atom-based SoC with an Altera FPGA (see MPR 12/13/10-01, "Intel Debuts ASIC Alternative"). The latest adopter of 3D stacking is Xilinx, which plans to sample some stacked Virtex-7 FPGAs in 3Q11 (see MPR 12/27/10-01, "3D Packaging Gains Momentum").

By using microbump contacts and an interposer layer, a processor chip could communicate with a DRAM chip through a 16,384-bit bus just like Venray's. (The Xilinx technology supports 20,000 connections.) This bus could easily operate at the DRAM's modest cycle time, even through the interposer. The bus would dissipate more power than Venray's DRAM+CPU design but would still consume less than 1W. Someday, if through-silicon vias become cost effective, the DRAM could sit directly atop the processor die and cut power even further. This two-die approach would allow manufacturers to build each chip in a suitable fabrication process while still overcoming the bandwidth wall.

In any case, some kind of revolution seems inevitable. If no other detour around Patterson's three walls is found, the industry may have to accept a compromise solution, regardless of the drawbacks. External memory is the last major subsystem to resist integration with CPUs. Ultimately, the laws of physics are working against memory's continued exile from the CPU kingdom. Someday, either memory will move to logic or logic will move to memory, no matter what happens to Venray. ♦