The Linley Group

MICROPROCESSOR report

Insightful Analysis of Processor Technology

SOFT MACHINES TARGETS IPC BOTTLENECK

New CPU Approach Boosts Performance Using Virtual Cores

By Linley Gwennap (October 27, 2014)

Coming out of stealth mode at last week's Linley Processor Conference, Soft Machines disclosed a new CPU technology that greatly improves performance on single-threaded applications. The new VISC technology can convert a single software thread into multiple virtual threads, which it can then divide across multiple physical cores. This conversion happens inside the processor hardware and is thus invisible to the application and the software developer. Although this capability may seem impossible, Soft Machines has demonstrated its performance advantage using a test chip that implements a VISC design.

Without VISC, the only practical way to improve single-thread performance is to increase the parallelism (instructions per cycle, or IPC) of the CPU microarchitecture. Taken to the extreme, this approach results in massive designs such as Intel's Haswell and IBM's Power8 that deliver industry-leading performance but waste power and die area. But more-efficient designs, such as Cortex-A7, offer weak performance on single-threaded software, which still constitutes the majority of applications. VISC (which, according to the company, does not stand for virtual instruction set computing) delivers high performance on a single thread using a simpler, more efficient design.

As an added bonus, Soft Machines has created an intermediate software layer that can translate from any standard instruction set into threadable VISC instructions, as Figure 1 shows. The initial design uses this conversion layer to run ARM code, but the company claims it can create a conversion layer for x86 or other instruction sets as needed.

Soft Machines is no fly-by-night operation. The company has spent seven years and $125 million to develop and validate its technology. It currently has more than 250 employees, led by CEO Mahesh Lingareddy and president/CTO Mohammad Abdallah. Investors include AMD, GlobalFoundries, and Samsung as well as government investment funds from Abu Dhabi (Mubdala), Russia (Rusnano and RVC), and Saudi Arabia (KACST and Taqnia). Its board of directors is chaired by Global Foundries CEO Sanjay Jha and includes legendary entrepreneur Gordon Campbell.

Soft Machines hopes to license the VISC technology to other CPU-design companies, which could add it to their existing CPU cores. Because its fundamental benefit is better IPC, VISC could aid a range of applications from

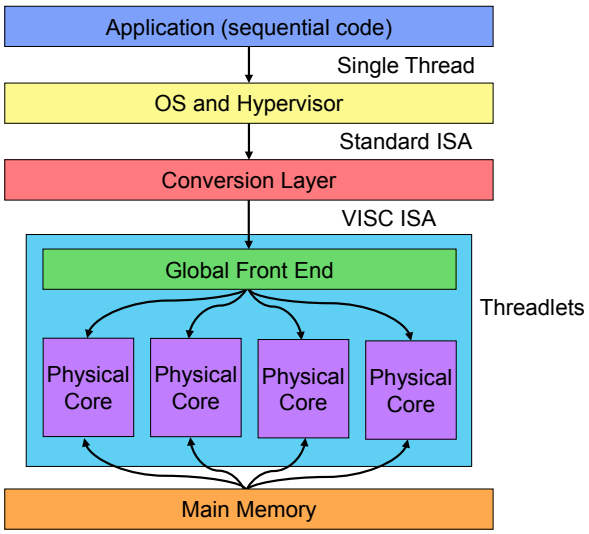

Figure 1. Soft Machines VISC technology. The conversion layer converts standard instructions (e.g., ARM) into VISC instructions. The hardware then converts a single instruction stream into multiple threadlets that can execute on multiple physical cores.

smartphones to servers. The company has applied for more than 80 patents on its technology.

Breaking the IPC Bottleneck

Since the first multiprocessor (SMP) computers appeared in the 1960s, programmers have realized that dividing their code to run on multiple CPUs improves throughput. In the past decade, multicore processor chips have driven SMP techniques into high-volume devices such as PCs, smartphones, and tablet computers.

Despite these advances, most programs continue to rely on a single instruction stream, or thread, to do most or all of the work. Breaking a program into multiple threads, ensuring that each can work independently with minimal inter-thread communication, is a challenging task. A few application types, such as graphics and packet processing, are simple to thread, but most others are not. For decades, development-tool vendors and other researchers have attempted to create a "magic compiler" that automatically creates a multithreaded program, but these tools either work only on a few easily parallelizable programs or require manual input from the programmer.

This failure puts the burden on CPU designers to improve single-thread performance. Once clock speeds hit a wall around 2005, performance per clock became the primary focus. Today's high-end CPUs, however, are well past the point of diminishing returns: adding another instruction-issue slot, another function unit, or a bigger reorder buffer burns more power while offering little IPC improvement. As a result, the IPC of high-end CPUs has increased at an annual rate of only 6% over the past decade, despite 40% annual growth in transistor count.

Thus, the implications of Soft Machines' VISC technology are tremendous. By breaking a single software thread into multiple hardware threads, the technology can combine multiple CPU cores into a single virtual core that delivers greater performance than any single CPU. This accomplishment could break the IPC bottleneck, accelerating performance scaling well beyond single-digit percentages to 2x or greater.

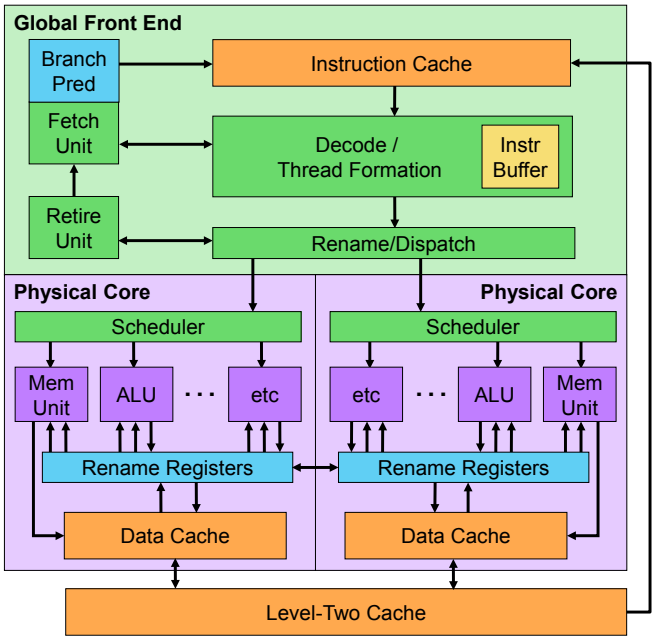

Figure 2. Conceptual diagram of a VISC processor. The global front end fetches a single instruction stream and divides it into threadlets that are dispatched to the multiple physical cores, which schedule and execute the instructions.

Pulling Apart the Threads

How can Soft Machines accomplish in hardware a task that software vendors have failed at? The hardware sees only a binary instruction stream, so it has less visibility than the compiler regarding data and code structures. The hardware has certain advantages, however, since it can access pointer values and other run-time information. At the conference, the company disclosed some description of its hardware design, although it did not provide complete details. The following provides a high-level view.

A VISC processor comprises a single instruction cache, a global front end, and two or more physical cores that each has its own instruction scheduling, register files, and data cache. The processor may also have a level-two (L2) cache that backs both the global instruction cache and the individual data caches. The initial test chip includes two physical cores and 1MB of L2 cache, as Figure 2 shows, but future designs are likely to include four or more cores.

The front end fetches instructions from the I-cache and places them into an instruction buffer. From this buffer, it attempts to form sequences of related instructions. For example, instructions with register dependencies can be grouped. These sequences are fairly short and are quite different from a conventional software thread; they can be better thought of as threadlets. The front end also performs global register renaming to avoid false dependencies. It can check pointers so as to group instructions that refer to the same memory location. The process of forming threadlets adds three cycles to the test chip's pipeline.

After creating the threadlets, the front end dispatches them to the physical cores. When it sends a threadlet to a core, it also allocates in that core the rename registers that are needed to execute the instructions. The goal is to minimize cases in which an instruction in one core requires a register allocated in another core, as this situation requires multiple cycles to resolve. Figure 2 shows a link between the two register files that handles this situation; the link becomes more complex in a processor with more than two cores.

Getting to the Core

Each core buffers the instructions received from the front end and can reorder them to avoid stalls. The buffer can

contain instructions from various threadlets, and these instructions can intermix freely. The core need not switch threads, because each threadlet has its own set of rename registers. When instructions are ready to execute, the core fetches data from its register file and completes the operations using its function units.

As in any out-of-order (OOO) design, speculative results are held pending until the completion of all previous instructions. In the VISC approach, the retirement unit must track instructions across all physical cores, since the original instruction stream could have been divided among multiple cores. Similarly, a branch misprediction detected in one core can affect instructions executing in other cores.

Thus, a VISC processor can be viewed as having global hardware for branch prediction as well as for fetching, grouping, tracking, and retiring instructions, but local hardware for scheduling and executing instructions and accessing memory. Compared with a high-IPC processor such as Haswell, the front end is of similar complexity, but the scheduler in each physical core is much simpler, as it manages only a few function units (versus eight in Haswell). The data cache also has fewer ports and can cycle faster. Because the execution resources of both cores can apply to a single thread, however, even a two-core VISC processor can deliver total ALU operations or memory operations per cycle that match or exceed those of Haswell.

VISC relies on a unique internal instruction set to do its magic, but true to its name, Soft Machines provides a conversion layer that converts from a standard instruction set to VISC. This approach is similar to how Transmeta processors executed x86 instructions (see *MPR 1/24/00,* "Transmeta Unveils Crusoe"), but VISC uses a completely different internal architecture. The company says the conversion overhead is less than 5%. In addition to its ARM-to-VISC translator, Soft Machines has prototyped an x86 translator and says it can develop other translators on customer request.

Better Load Balancing

The previous example shows how the global front end can break down a single thread, but it can also handle multiple software threads at the same time. In this way, a single VISC processor can emulate a traditional multicore design. But unlike traditional designs, it can more easily perform load balancing, matching processing power to the task at hand.

Figure 3 shows an example with two active threads: one heavy (high performance) and one light. In a processor with several identical cores, each thread would run on one core, wasting cycles for the light thread and limiting performance for the heavy thread. In a VISC design with two physical cores, the heavy thread is split into two hardware threads, one running on each core. Note that the second core automatically shares its resources between the heavy thread and the light thread, because a VISC core can mix instructions from multiple threadlets.

For More Information

A copy of the Soft Machines presentation from the Linley Processor Conference is available for free download at *www.linleygroup.com/events/proceedings.php?num=29* (registration required). For additional information on Soft Machines, access the company's web site at *www.softmachines.com*.

Soft Machines calls this approach "virtual cores." From a software viewpoint, the heavy thread runs on a single virtual core that comprises more than one physical core, while the light thread runs on a virtual core that uses only a portion of a physical core. This allocation of cores is invisible to software, except that the heavy thread runs faster than it would otherwise. According to the company, its front-end hardware will recognize which threads need more performance and allocate the virtual core resources appropriately. The operating system need not know the number of physical cores in the processor to assign threads or balance the load.

Demonstrated Performance

Soft Machines has designed and fabricated a test chip that implements its VISC architecture using two physical cores. It refused to discuss the number of function units or other basic microarchitecture capabilities of these cores but characterized them as A15-class CPUs. We interpret this statement to mean that each core can execute three to four operations per cycle with moderate instruction reordering.

The company also withheld the test chip's clock speed. Because it uses a pipeline with only 10 stages (including some extra stages for VISC scheduling), the CPU cannot match the high clock speeds of leading-edge x86 and ARM processors. We estimate the chip runs at several hundred megahertz. Even at this low speed, it completes some programs in less time than a low-end Haswell processor.

Despite its relatively simple design, the chip achieves spectacular performance. On the single-thread SPEC2006

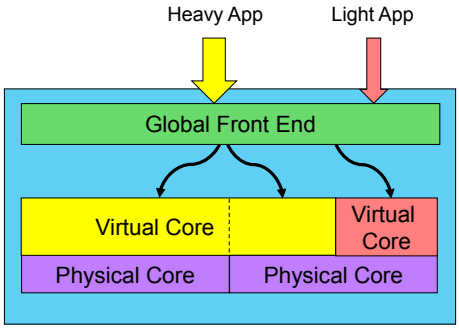

Figure 3. Soft Machines virtual cores. A single virtual core can comprise multiple physical cores or only a portion of a physical core. In this way, the front-end hardware aligns the performance of the virtual core with the needs of each thread.

test suite, the company reports an average IPC of 2.1, counting ARM instructions rather than VISC instructions. This IPC compares with 0.71 for Cortex-A15 and 1.39 for Haswell. (For consistency, Soft Machines measured the IPC on all three processors using GCC rather than Intel's favorite compiler, ICC.) Thus, the VISC chip achieved three times the IPC of ARM's highest-end CPU shipping today and 50% better IPC than Intel's fastest mainstream CPU.

Figure 4 shows more-detailed performance results comparing the VISC test chip against Cortex-A15 (measured in a Samsung Exynos processor). The figure shows the results of 62 different benchmarks, including components of SPEC2000, SPEC2006, EEMBC's digital entertainment benchmark (DenBench), and the Kraken JavaScript benchmark. While individual results range from 1.5x to 7x, the average gain across these tests ranges between 3x and 4x. Soft Machines points out that although the test chip runs Linux and other applications, the software contains workarounds for certain hardware bugs, so it will likely obtain better results with future tuning and bug fixes.

Although these results are impressive, they require some caveats to put them into perspective. A shorter CPU pipeline reduces branch penalties and other pipeline hazards, thereby improving IPC compared with a longer pipeline. In addition, a low CPU speed reduces the effective latency of caches and main memory (measured in CPU cycles), again improving IPC relative to a CPU with a faster clock. The latter effect might explain why the test chip appears to perform better on SPEC2006 than on SPEC2000, which has a smaller memory footprint.

Doubling Performance Is Realistic

The test results are difficult to interpret, owing to both the lack of information about the test chip's core CPU and the effects of the short pipeline. A better way to quantify the advantage of using VISC is the performance increase relative to a single core. By coupling the resources of two cores to execute a single thread, the maximum performance increase (compared with a single core of similar design) is 2x. In reality, the front end cannot fully utilize the resources of the second core, particularly for code that is difficult to break apart. For example, code that contains many branches or many register dependencies will be harder to thread and thus will use less of the second core.

Soft Machines has run many tests and simulations of its VISC technology. It estimates the second core improves performance by an average of 50–60% across a variety of benchmarks. This factor implies that the test chip would achieve about 1.3 IPC when using a single core—considerably higher than Cortex-A15. This higher IPC includes the effects of the lower clock speed and shorter pipeline.

Although the test chip has only two physical cores, Soft Machines has run simulations on a four-core design. As one might expect, the performance gains diminish for the additional cores: the third core adds 20–30% to single-thread performance, and the fourth adds only 10–20%. In total, the four-core design delivers about twice the performance of a single core. The unused resources in the extra cores, however, can be devoted to additional threads. For example, a four-core design could run two threads at close to their maximum performance.

Performance-critical applications will benefit from VISC, but the technology can also apply to low-power designs. As the test chip demonstrates, a VISC design can operate at a relatively low clock speed while achieving the same absolute performance as a traditional design operating at a higher clock speed. Thus, it should use less power, particularly if the voltage is reduced as well. Soft Machines, however, declined to reveal the power consumption of its test chip.

Other details also remain undisclosed, including die area. Details of how the processor handles privileged operations, inter-thread synchronization, traps, and interrupts could all affect how well it runs certain applications. Performance could vary widely across different workloads. The company provides additional information to potential customers under NDA and plans to make further disclosures over time.

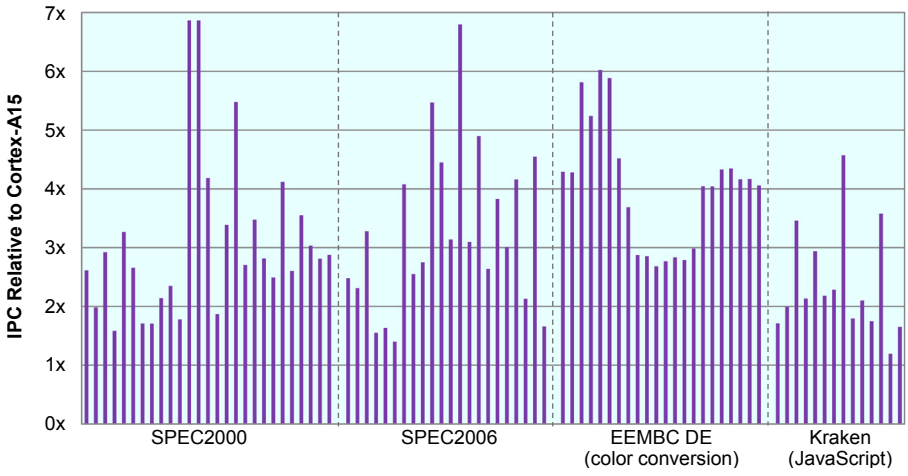

Figure 4. VISC performance results. This chart shows measured instructions per cycle (IPC) for Soft Machines' test chip normalized against Cortex-A15. The company did not disclose the clock speed of the test chip. (Source: Soft Machines)

A Rising Tide

Soft Machines has identified a critical (perhaps the most critical) problem in CPU design today: the minimal improvement in single-thread performance over the past decade. Despite the obvious need for multi-threaded software, most applications—even performance-intensive ones—continue to

rely on a single thread. The software tools for creating multithreaded applications remain primitive.

With the announcement and demonstration of its VISC technology, Soft Machines has taken a big step toward solving this problem. By shifting the burden to hardware, VISC aims to deliver the benefits of multithreading to all applications. The initial performance results are excellent: a 50–100% gain in single-thread IPC represents a decade of progress at the industry's current sluggish pace. As with any radical new technology, however, we remain skeptical, particularly given the startup's limited disclosure. Potential customers must fully assess the technology to determine how it will perform in an actual product across a range of workloads.

Assuming the technology works as advertised, it will change the way all processors are designed. CPU designers will stop trying to improve IPC by adding more hardware; in fact, complex high-IPC designs like Haswell could disappear in favor of smaller, simpler ones. Replacing large cores with clusters of simpler cores will improve performance and power efficiency, benefiting almost every type of processor. To deliver on this promise, Soft Machines must fully validate VISC and license it to leading processor vendors. ♦