Getting Gigascale Chips

Challenges and Opportunities

Processor performance has increased by five orders of magnitude in the last three decades, made possible by following Moore's law—that is, continued technology scaling, improved transistor performance to increase frequency, additional (to avoid repetition) integration capacity to realize complex architectures, and reduced energy consumed per logic operation to keep power dissipation within limits. Advances in software technology, such as rich multimedia applications and runtime systems, exploited this performance explosion, delivering to end users higher productivity, seamless Internet connectivity, and even multimedia and entertainment.

The "technology treadmill" will continue, providing integration capacity of billions of transistors; however, several fundamental physics issues will pose barriers. In this article, we will examine these barriers, describe how they are changing the landscape, discuss ways to get around them, and predict how future advances in software technology could help continue the technology treadmill.

SHEKHAR BORKAR, INTEL

in Continuing Moore's Law

TIPS-level performance will be delivered only if engineers and developers learn to exploit emerging paradigm shifts.

Getting Gigascale Chips

Challenges and Opportunities in Continuing Moore's Law

OUR ADDICTION—WILL PERFORMANCE CONTINUE TO INCREASE?

Figure 1 shows almost five orders-of-magnitude growth in compute performance (in millions of instructions per second, or MIPS) over the last three decades; note that the performance doubled every two years. We are so used to this growth that we take for granted that performance will continue to double every two years, and we plan and predict future applications accordingly. Moore's law will still be around for at least one more decade, if not more. This is because advances in process technology, such as lithography, are on track to allow us to double transistors every two years.

Expect trillions of instructions-per-second (TIPS) performance by the end of the decade. There will be some major paradigm shifts, however, and "business as usual" will not be an option. To achieve TIPS performance, one may conclude that frequency of operation should continue to increase at the same rate and transistor integration must continue to realize even more complex systems. To continue to increase frequency of operation, the transistor performance must increase, and to reduce power consumption, the supply voltage must continue to decrease. Therefore, the threshold voltage (voltage required to turn the transistor on) must decrease.

THE DREADED SUB-THRESHOLD LEAKAGE

A transistor is not a perfect switch; it leaks when it is turned off, and this sub-threshold leakage increases exponentially as you reduce the threshold voltage. In a perfect switch, such as a light switch, no current flows through when it is turned off, and the light bulb does not glow. If the switch is bad, then even though it is turned off and the light bulb does not seem to glow, some residual current could be flowing through.

Modern transistors are analogous to these bad light switches: They leak when they are turned off. To make it worse, Moore's law allows you to double these "bad switches" every two years, exponentially increasing the leakage every two years, eventually becoming noticeable.

Figure 2 shows the increase in projected sub-threshold leakage power when you follow Moore's law by doubling logic on a chip every generation. Clearly, the leakage power will be on the order of several hundreds of watts beyond 90 nm, and will not be acceptable. Even if you integrate only 50 percent more transistors, the leakage power will still be on the order of hundreds of watts, and thus not an acceptable solution.

Figure 3 shows the sub-threshold leakage power as a percentage of total power; it is already approaching the practical limit of 50 percent. When this leakage power is

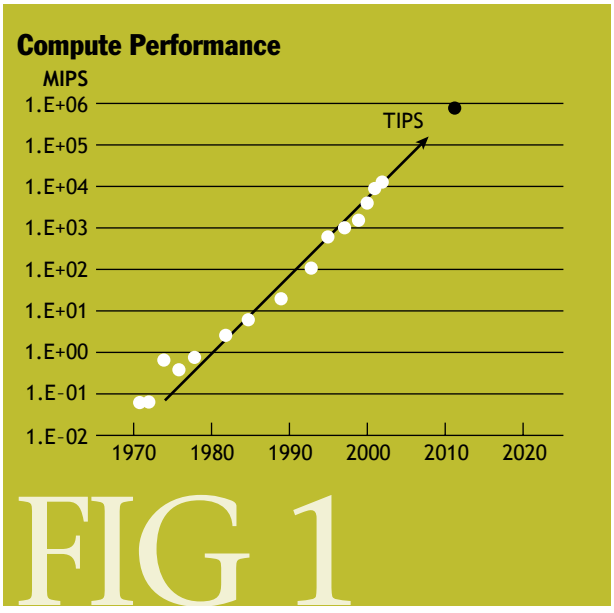


FIG 1

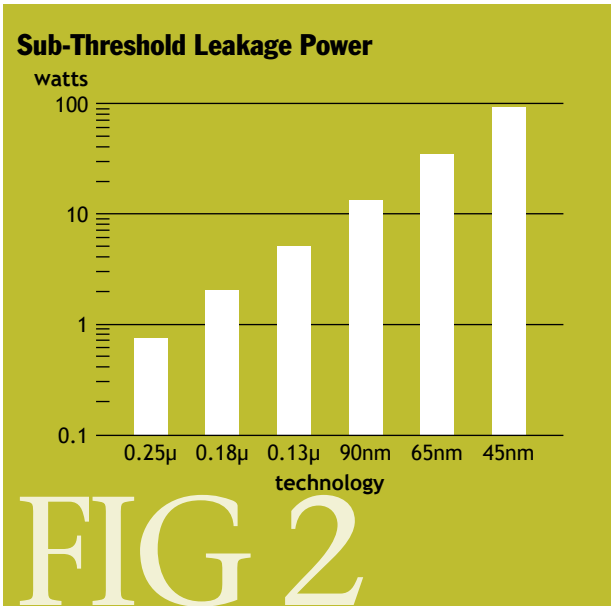
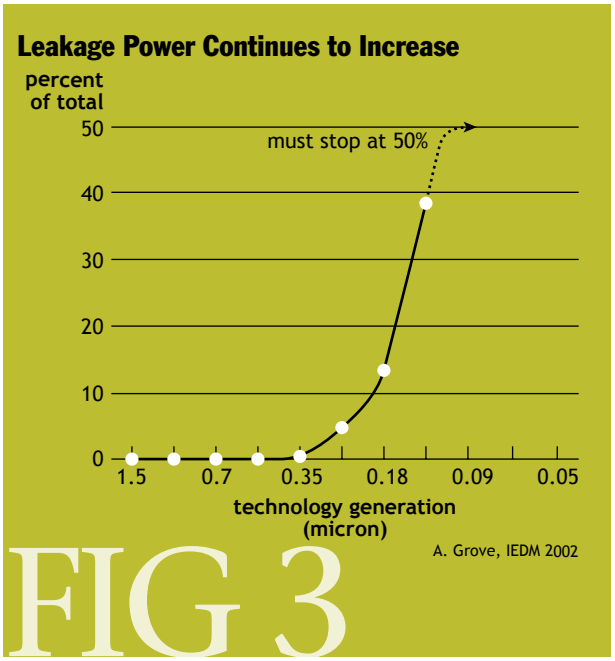
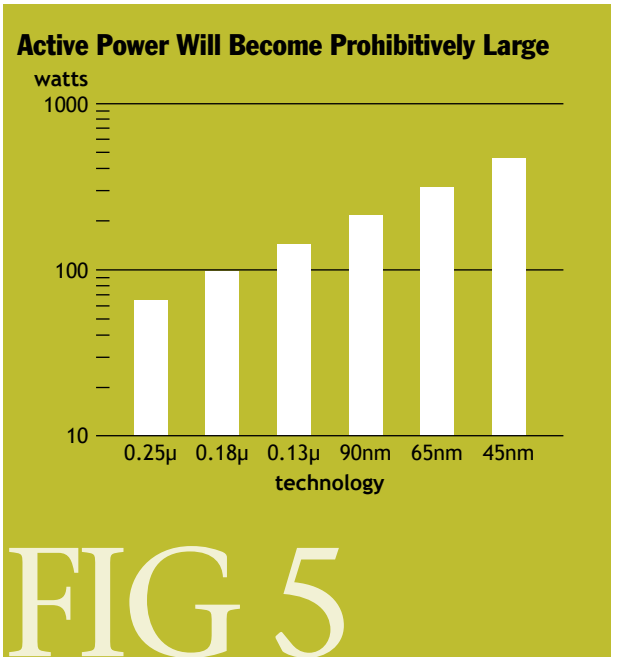
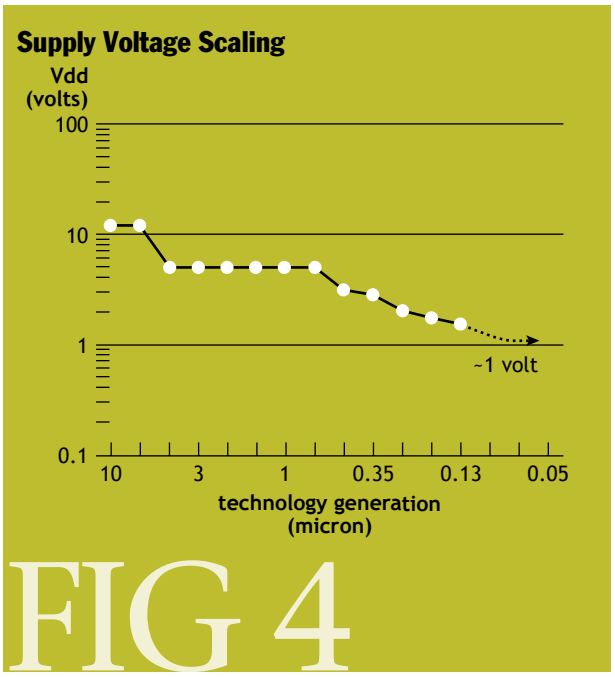
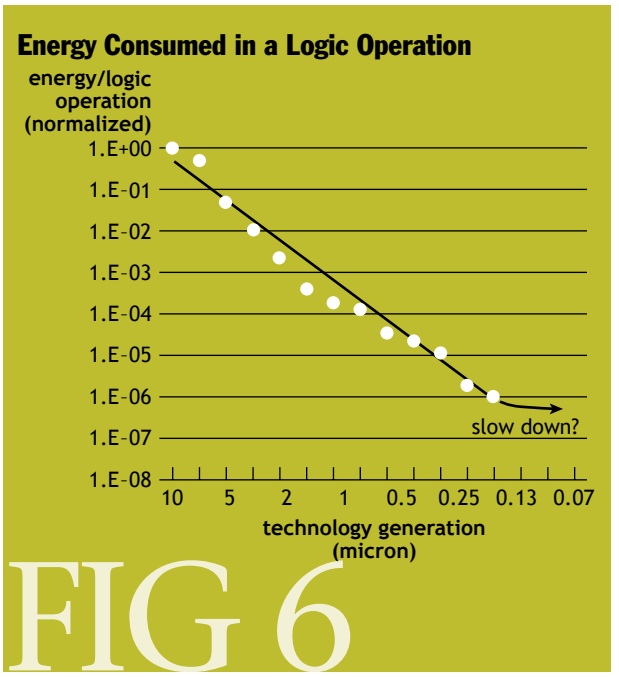


FIG 2









about 50 percent of total power, further supply-voltage scaling does not make sense.

SUPPLY VOLTAGE AND ACTIVE POWER

Active power consumption (that is, total power minus the leakage) of a chip has a quadratic relationship with the supply voltage (V):

$$\text{Active Power} = CV^2f$$

Therefore, scaling down supply voltage is beneficial to reduce the active power. Yet, along with supply-voltage scaling, threshold voltage also must scale, which in turn exponentially increases the leakage power. That is why supply-voltage scaling will have to slow down, or even stop, when the leakage power is about 50 percent of the total power, as shown in figure 4. Notice that the supply voltage remained 5V until 0.7 micron generation, and has

Getting Gigascale Chips

Challenges and Opportunities in Continuing Moore's Law

been decreasing approximately 30 percent per generation since then. But will it in the future?

Even if you assume that the supply voltage will scale aggressively at about 15 percent per generation, and the frequency of operation will continue to increase by about 40 percent per generation, figure 5 shows predicted active power if we blindly follow Moore's law, doubling the logic on a chip every generation. Clearly, even the active power will be prohibitively large. To put this into perspective, today's high-performance Pentium 4 processors clock at 3.2 GHz and consume about 75 watts at 1.4V. The Itanium 2 processor is well above 100 watts and the power supply currents to these processors reach 50 to 100 amperes, comparable to the currents a car battery provides to the starter.

Therefore, although the transistor performance will continue to increase, albeit at a slower rate than in the past, the energy consumed per logic operation (and power) will not go down as much as it did in the past, as shown in figure 6.

WILL INTERCONNECTS (WIRES ON THE CHIP) BE THE LIMITERS?

Metal interconnects on the die were always considered to be limiters to further performance in the future. Metal interconnects (i.e., wires) shrink in size as geometries scale down, thereby increasing the resistance (R). They also get closer to each other, since the space scales down as you shrink, thereby increasing the capacitance (C). Therefore, the product RC delay does not scale well. Typically, more levels of metals are added to make up for it.

If power and energy truly turn out to be the limiters, then the size of the logic on a die will continue to shrink, reducing the lengths of the interconnections, R and C, and the RC delay. Subsequently, interconnections will not be the limiters. Therefore, our job is to remove power and energy as the limiters so that we can start thinking about limitations of the interconnections.

THE CHALLENGES...

In the near future multi-billion transistor integration capacity will be available, but it will be unusable because of power and energy consumption and limited transistor performance. So, how do you deliver TIPS performance?

Performance at any cost will not be an option in the future as in the past; system architectures will have to emphasize performance delivered in a given power envelope, with complexity limited by energy efficiency. Several options exist, most requiring major paradigm shifts in architectures, systems, and application software. The underlying theme is to exploit the integration capacity and deliver value performance, with even higher integration of potentially slower transistors.

IMPROVE POWER AND ENERGY EFFICIENCY

The first step is to realize inefficiencies of our present-day micro-architectures and supporting software. Each generation of micro-architecture—from scalar to super-scalar, out-of-order and speculative, and deep pipelines—exploited inherent instruction-level parallelism, but

incurred about a 20 to 30 percent loss in energy efficiency with each generation. Because of our quest for higher and higher peak performance, designers employed architectures and design methods to improve peak performance at the expense of energy and power efficiency. We have also pushed higher frequency as a means to get higher performance. Although energy-inefficient, these architectures provided an invaluable benefit of hiding the instruction-level parallelism from the programmer—software did not have to explicitly know about it. Thus, it delivered higher performance with backward software compatibility. Now we need to reclaim this energy inefficiency.

Intel's Centrino processor is a good example of this strategy. This processor clocks at 1.6 to 1.7 GHz, moves away from frequency alone to provide performance, exploits latest advances in architecture to improve instructions executed per clock, and provides about the same integer performance as a Pentium 4 clocked at 2.4 GHz. Yet it consumes only 22 watts of power, compared with 60 to 70 watts by a Pentium 4.

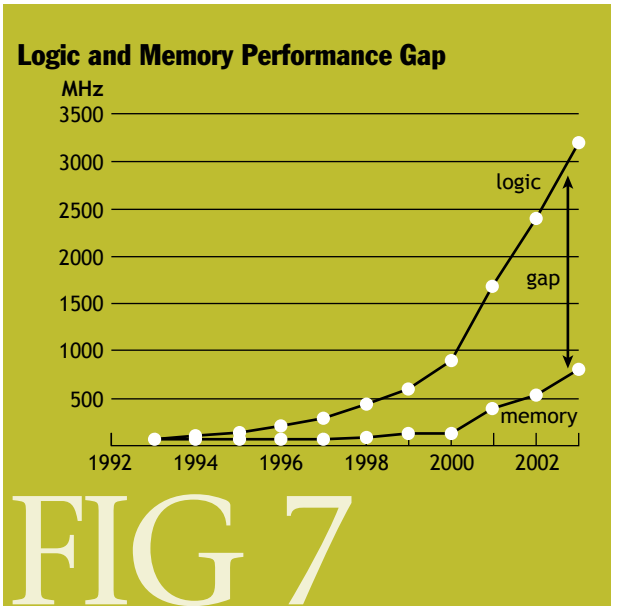
MULTITHREADING: A BOLD STEP IN THE RIGHT DIRECTION

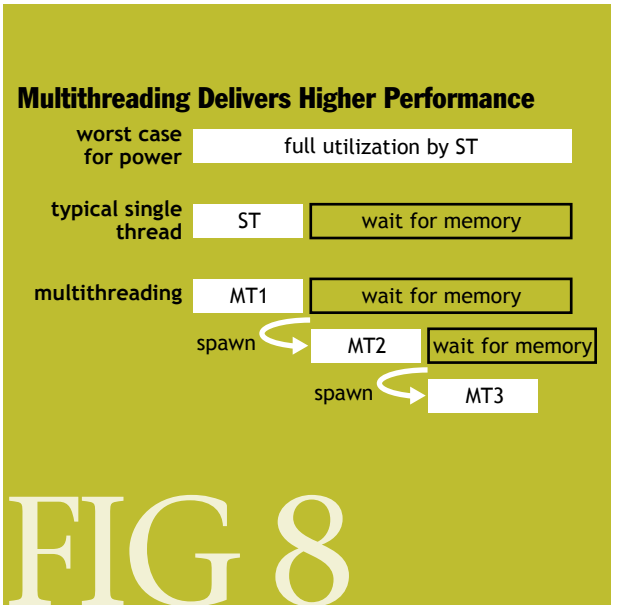
The next step is to go beyond instruction-level parallelism to thread-level and processor-level parallelism. In a typical system, the thermal management and power delivery system is designed for the worst-case application thread, which is mostly cache-bound and keeps all the hardware units busy. In reality, however, most of the application threads have frequent cache misses and wait for data from the memory. Memory subsystems are much slower, as shown in figure 7, and the processors have to wait several clock cycles idling before execution resumes. This leaves a large gap in thermal and power capabilities between what a realistic application needs and what the worst-case application demands.

If the software is written in terms of multiple threads, then this gap can be narrowed, resulting in better use of the hardware. In the event of a cache miss, when the processor is waiting for the memory to supply data, another thread can be spawned to keep the hardware units busy, thereby improving the overall performance in the same thermal and power envelope, as shown in figure 8.

MULTIPROCESSOR: THE ULTIMATE IN PARALLELISM?

Another energy-efficient micro-architecture solution is to use multiple processors on a single chip with a large shared cache. The principle behind this notion is Pollack's rule, which states that the increase in performance is roughly proportional to the square root of the increase





in complexity. In other words, if you double the logic in a processor, then it delivers only 40 percent more performance—as evidenced by today's leading processors. Multiprocessing, on the other hand, has the potential to provide near-linear performance improvement. Two smaller processors, instead of a large monolithic one, can potentially provide 70 to 80 percent more performance. Compare this to only 40 percent from the large monolithic processor. Multiprocessors have other benefits as well:

- Each processor can be individually turned on or off, thereby saving power.

Getting Gigascale Chips

Challenges and Opportunities in Continuing Moore's Law

- Each processor can be run at its own optimized supply voltage and frequency.
- It's easier to load balance among processors to distribute heat across the die.
- They can potentially produce lower die temperatures, improving reliability and leakage.

Multithreading and multiprocessor schemes are not difficult to implement in hardware, but the biggest challenge is software. The programming models are different from traditional single-threaded models. They need different programming paradigms and careful software engineering practices, yet have the potential to deliver performance beyond today's single-threaded model. Therefore, a major paradigm shift in software is needed to shift from today's era of instruction-level parallelism to thread- and processor-level parallelism in order to deliver TIPS performance.

LARGE MEMORIES: HIGHER PERFORMANCE AT LOWER POWER

On-die memories, such as larger caches, can also provide higher performance at much lower power and energy. In the past, we always allocated the available transistor budget to complex logic, with minimal budget for on-die caches, resulting in cache-starved processors. The cache

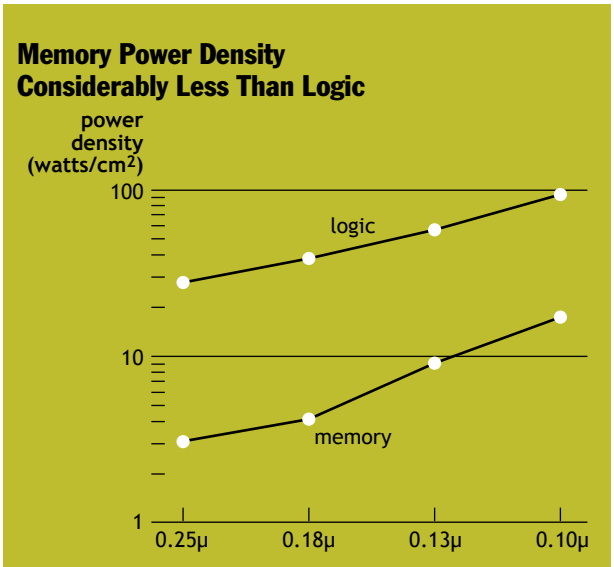


FIG 9

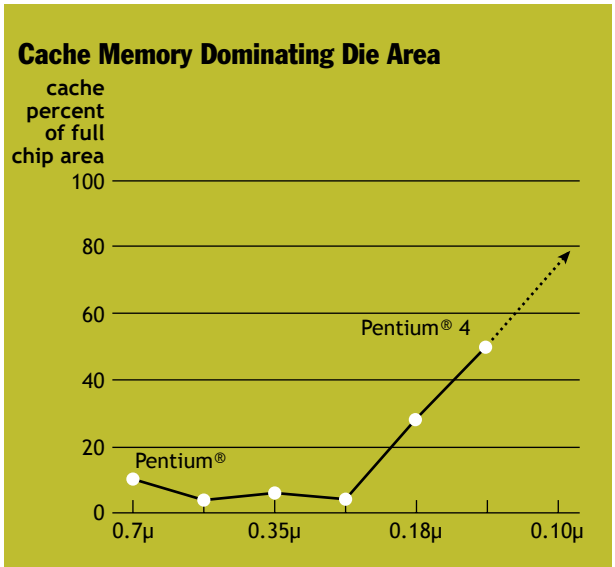
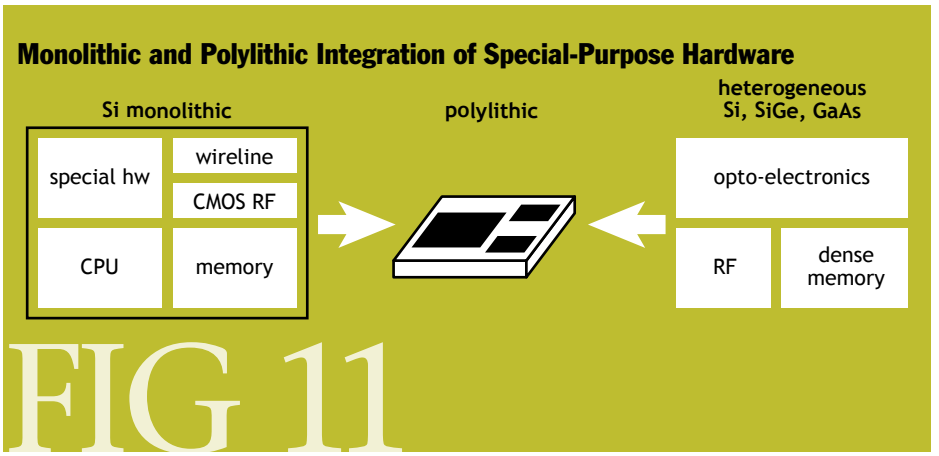


FIG 10



memory has power density an order of magnitude lower than that of logic (see figure 9). Its leakage power can be controlled and an on-chip cache can generally provide higher performance.

Memories have lower power density because you typically access only a small portion of the large memory every clock cycle, reducing overall activity in the memory. Since memory is a regular structure, access patterns can be predicted and leakage control techniques are easy to implement. Large caches make a lot of sense in a power-constrained scenario, which is why you now see a larger and larger portion of the die area allocated for cache memories, as shown in figure 10.

SPECIAL-PURPOSE HARDWARE IMPROVES POWER AND ENERGY EFFICIENCY

In the gigascale integration era, we must go beyond traditional general-purpose compute performance and focus on the system-level end-user performance. Applications tend to have several special-purpose tasks, such as MPEG encode and decode, TCP/IP network protocol processing, and so forth, which are processed today on general-purpose processors, but not energy efficiently. A special-purpose hardware unit, customized for a specific task, is more area-, power-, and energy-efficient by almost an order of magnitude or more. This is because the hardware is optimized to do a predetermined set of tasks and is therefore compact and more efficient. For example, your car and lawn mower both have internal combustion engines, and in principle you could install your car engine in a lawn mower and it would do the same job. Would it be efficient? Hardly!

Even today you see several instruction-set architectures implementing single-instruction stream multiple-data stream (SIMD) integer and floating-point operations for multimedia applications, but we need to take this a few steps further—using a general-purpose processor with several special-purpose hardware units that are optimized and integrated for specific tasks, as shown in figure 11. With the availability of unlimited transistor integration capacity, which you could not otherwise use because of power and energy factors, it would make a lot of sense to use this capacity to implement these functions in hardware and provide value performance at lower power and energy. These special-purpose hardware units will provide orders of magnitude of equivalent general-purpose performance.

SUMMARY

Gigascale transistor integration capacity will be available in the future, but its use could be limited by transistor performance, energy, and power dissipation. Performance at any cost will not be an option, yet we must stay on the technology treadmill to deliver TIPS end-user performance. There are several emerging paradigms, such as the shift from instruction-level parallelism to thread- and processor-level parallelism, large on-die caches, and the integration of special-purpose hardware. Together, all these paradigms have the potential to deliver the expected TIPS-level performance, provided the application and system software take appropriate steps to exploit them. Q

LOVE IT, HATE IT? LET US KNOW:

feedback@acmqueue.com or www.acmqueue.com/forums

SHEKHAR BORKAR received B.S and M.S. degrees in physics in 1977 and 1979, and a master's in electrical engineering in 1981 from the University of Notre Dame. He joined Intel, where he worked on the design of the 8051 family of microcontrollers, high-speed communication links for the iWarp multicomputer, and Intel supercomputers. Borkar is an Intel fellow and director of circuit research in the Intel Labs, researching low-power, high-performance circuits and high-speed signaling. He is also an adjunct faculty member of the Oregon Graduate Institute and teaches Digital CMOS VLSI design.