We are pleased to introduce this special issue of *IEEE Micro* featuring papers that capture the best presentations from the Hot Chips 16 conference held last summer at Stanford University.

Hot Chips provides a unique perspective on chip design. The conference collects presentations on the bleeding edge of integrated circuit design. It is unique in being a practical, technical conference focused on real chips, in contrast to the usual academic or commercial-marketing conferences. The short review cycle and the publication of slides, rather than full papers, results in timely presentations; the conference covers the latest chips from both startups and established industry leaders.

The Hot Chips 16 program provided a clear reflection of the overall semiconductor industry. The presentations illustrated the shift toward higher degrees of parallelism in microprocessors and the move toward mobile and media devices: the cell phones, televisions, and other consumer electronics devices that now consume the bulk of all computing cycles.

This year’s Hot Chips special issue is a crosssection of the conference and of the industry. The articles on Montecito, Niagra, and Horus are examples of the latest in “big-iron” microprocessors and the system-logic chips that surround them. These papers show the trend toward thread-level parallelism in high-end microprocessors. The Horus paper illustrates parallelism one level up from the processor, describing a chipset that enables 32-way symmetric multiprocessing systems (SMPs) to be built with Opteron processors.

The remaining two articles reflect the trend toward media processing. The article on the GeForce 6800 (no relation to the ancient Motorola microprocessor with the same number) reveals the state of the art in graphics processors (GPUs). In recent years, GPUs have become programmable and now have more raw processing power and memory bandwidth than most high-end general-purpose microprocessors. The 6800 is an example of this trend. The trend toward media processing brings with it a need for tighter security, to protect both privacy and intellectual property. The article by Eberle et al. describes architectural support for providing such security.

The five articles together lead to a number of interesting observations. First, several of the chips are “monsters.” Montecito, at 1.7 billion transistors and 596 mm², is a transistor and silicon monster the size of which the processor industry has not seen before. The GeForce 6800, with 6 vertex processors and 16 fragment processors (each of which is a VLIW and SIMD parallel processing engine), is a monster with an order of magnitude more floating-point muscle than any other microprocessor described...
herein. Finally, the Niagra processor, which appears to software as 32 independent Sparc processors, is a monster capable of eating enormous quantities of thread-level parallelism.

Common themes across many of the articles are power, parallelism, and memory performance. Experiencing severely diminished returns on transistors intended to squeeze the last remnants of performance from a single thread, microprocessor designers are turning their attention to explicit thread-level parallelism. Montecito embodies thread-level parallelism with two cores executing two threads each. Niagra goes more extreme with eight cores of four threads each. The GeForce 6800 is more extreme yet, offering a total of 22 processors, each of which is four-way SIMD parallel and many-way VLIW parallel. The era of single-thread instruction-parallel-only processors is clearly coming to an abrupt end.

The increase in explicit parallelism is in large part driven by power concerns. As chips push the limits of semiconductor and manufacturing technology, power-efficient designs become essential to delivering more performance. Explicitly parallel techniques offer a more efficient means of converting power into performance than do techniques that must discover the implicit—and often limited—instruction-level parallelism hidden in a single thread. Attention to power dissipation can be seen in several of the papers. Montecito, for example, uses explicit thread-level parallel techniques in its chip-multiprocessor and hardware multi-threaded architecture as well as providing closed-loop control of voltage and frequency to operate the processor as fast as possible without exceeding thermal constraints.

With the increase in performance potential from on-chip parallelism comes an increased demand for memory bandwidth. The Niagara processor satisfies this demand with a crossbar-connected, multi-banked L2 cache and four on-chip memory controllers that provide 20 GB/s of aggregate memory bandwidth. The GeForce 6800 has four 64-bit wide memory controllers that support DDR-2 or GDDR-3 memories. The 6800 also performs lossless compression on memory traffic to boost its effective memory bandwidth. The Hours chipset uses a directory and remote data cache to reduce bandwidth requirements over that required to implement cache coherence in today's snoopy SMP systems. Horus uses 3.12 GHz signaling to provide high-bandwidth between the "quads" of an SMP system.

The quality of the Hot Chips conference and of this special issue of IEEE Micro is due to the efforts of many people. First, thanks are due to all of the authors who submitted abstracts. The submitted abstracts capture the vitality of the field and provide the essential ingredient of the conference. We also thank the program committee: Forest Baskett, Allen Baum, Pradeep Dubey, Norm Jouppi, Christos Kozyrakis, John Nicholls, Tom Petersen, Chris Rowen, Mitsuo Saito, John Sell, Alan Smith, and Mateo Valero. The committee members spent many long hours soliciting submissions, reading and critiquing abstracts, participating in the selection process, and shepherding presentations. We trust that you will find the result of this process as exciting as we do.

Bill Dally is a professor of electrical engineering and computer science and chair of the computer science department at Stanford University. His research interests include streaming supercomputers, image and signal processors, and scalable network fabrics. Dally has a PhD in computer science from Caltech, an MS in electrical engineering from Stanford University, and a BS in electrical engineering from Virginia Polytechnic Institute. He is a Fellow of the IEEE and the ACM, and received the ACM Maurice Wilkes Award in 2000 and the IEEE Seymour Cray Award in 2004.

Keith Diefendorff is chief microprocessor architect at Apple Computer, where he works on next-generation high-performance microprocessors. He has been the chief architect of several high-performance processors including AltiVec/G4 at Apple and PowerPC and the 88110 at Motorola, and for two years was editor-in-chief of Microprocessor Report. Diefendorff has an MSEE from the University of Akron and holds 12 U.S. patents. He is a member of the IEEE and the IEEE Computer Society.

Direct questions and comments about this special issue to Bill Dally, Stanford University, Computer Systems Laboratory, Gates Room 301, Stanford, CA 94305; billd@csl.stanford.edu.