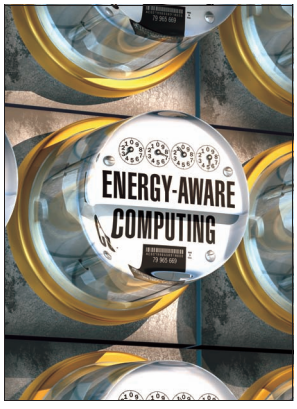
Guest Editors' Introduction

ENERGY-AWARE COMPUTING



Thomas F. Wenisch
University of Michigan

Alper Buyuktosunoglu
IBM T.J. Watson Research Center

The march of Moore's law continues to provide ever more transistors, but unfortunately Dennard scaling—the concomitant reduction of CMOS threshold and supply voltages—has come to an end. Hence, power density and peak power demands are soaring with each new silicon process generation on a trajectory that far outstrips improvements in our ability to dissipate heat. Because of these trends, energy efficiency (as opposed to area or switching speeds) is becoming the limiting factor in computing performance in platforms from smartphones to warehouse-scale computers.

Energy-efficiency constraints manifest in many ways across computing form factors. In sensors, limiting energy consumption is critical, because many sensors must operate their entire lifetimes from a single battery charge, or might harvest only a trickle of power from their environment.

In the embedded and handheld market, weight, volume, and battery life are of utmost concern. With the slow pace of battery technology improvements, we must rely on more efficient silicon to continue delivering more computing horsepower in ever-shrinking devices. Moreover, in these systems, performance is often limited by heat—no one would buy a smartphone with a large cooling fan whirring by one's ear! Hence, limiting peak power draw is a critical concern.

At the other end of the computing spectrum, warehouse-scale computers and data centers are equally limited by both efficiency and peak power concerns. The cost of a data center's power provisioning and cooling infrastructure typically grows linearly with the installed systems' peak power requirements. Many operators face the constant challenge of packing in more servers without having to build another facility—a multimillion-dollar expense. Moreover, electricity costs often account for 20 percent or more of the data center's total cost of ownership (TCO), and the worldwide environmental impact of data-center power consumption is substantial.

In this special issue, we have a collection of peer-reviewed articles that examine energy efficiency and its implications across scales, from sensors and embedded systems to data centers. The articles also span the system stack, from circuits through software. The overriding theme of these articles is to find ways to squeeze more performance out of limited power and energy budgets. Every joule that is conserved can be immediately leveraged as performance improvement by increasing clock frequency. In a world without Dennard scaling, energy efficiency is the new performance.

Energy efficiency from a circuit perspective

We begin the special issue by examining energy efficiency from a circuit perspective. Vibhu Sharma and his coauthors from Imec and KU Leuven explore circuit-level challenges in designing highly energy-efficient SRAM memories in "Ultra Low-Energy SRAM Design for Smart Ubiquitous Sensors." They describe techniques that prioritize energy efficiency over silicon area and speed for energy-sensitive applications such as smart sensor nodes. In particular, they examine the combination of local assist

Published by the IEEE Computer Society

circuits, eight-transistor SRAM cells, a low-energy sense-amplifier design, and a low-swing write scheme, and they report on their measurements of two fabricated prototypes.

Microarchitectural innovations to improve performance per watt

The next two articles examine microarchitectural innovations that improve performance per watt. In "e6500: Freescale's Low-Power, High-Performance Multithreaded Embedded Processor," David Burgess and his colleagues at Freescale detail the microarchitecture and energy-efficiency features of the e6500, a new multicore processor for embedded systems. Paramount among the challenges in designing the e6500 were meeting its power-efficiency objectives and supporting rapid transitions into low-power states.

In "DySER: Unifying Functionality and Parallelism Specialization for Energy-Efficient Computing," Venkatraman Govindaraju and his colleagues at Intel and the University of Wisconsin–Madison unify parallelism specialization and functional specialization into a single framework to provide specialized execution units for program regions. Central to their approach are three techniques—region growing, vectorized communication, and region virtualization—which enable accelerated applications that outperform conventional out-of-order, SSE (Streaming SIMD Extensions), and GPU-accelerated designs on both performance and energy.

Power and cost efficiency in data centers

We close the special issue with two articles that examine power and cost efficiency in data centers. In "Optimizing Data-Center TCO with Scale-Out Processors," Boris Grot and his coauthors explore cost-optimal design of servers chips that can increase computing density for scale-out workloads; for example, workloads like web search and data serving that rely on in-memory processing and massive parallelism across scores of servers. Central to their design for Scale-Out Processors is the notion of integrating multiple pods—stand-alone multicore servers with a carefully sized last-level cache—onto a single die, providing a cost-effective means to scale the number of cores brought to bear on scale-out workloads.

Energy efficiency (as opposed to area or switching speeds) is becoming the limiting factor in computing performance in platforms from smartphones to warehouse-scale computers.

Finally, Sherief Reda, Ryan Cochran, and Ayse K. Coskun discuss packing threads into cores, allowing now-idle cores to sleep, as a mechanism to cap peak power in servers. They describe their technique, Pack and Cap, in "Adaptive Power Capping for Servers with Multithreaded Workloads." Although many previous power-capping proposals employ dynamic frequency and voltage scaling (DVFS) to limit processor power consumption, Pack and Cap goes further by trying to pack threads on a subset of available cores, thereby allowing idle cores to enter low-power states that further reduce the minimal achievable power cap.

This special issue provides a sampling of the large activity going on in the area of energy-aware processor and system design. We hope that the articles in this special issue illuminate the breadth and importance of energy-aware computing, and help to further the conversation as energy, power, and thermal constraints become ever more important in microarchitecture and system design. MICRO

Acknowledgments

We thank Erik Altman for inviting us to guest edit this issue and providing guidance throughout the process, all the authors of submitted papers, and the reviewers who provided valuable feedback to help us select this set of articles and assist the authors in improving their work.

Thomas F. Wenisch is the Morris Wellman Faculty Development Assistant Professor of electrical engineering and computer science at the University of Michigan. His research interests include computer architecture, server and data-center energy efficiency, smartphone architecture, and multiprocessor systems. Wenisch has a PhD in electrical and computer engineering from Carnegie Mellon University. He is a member of IEEE and the ACM.

Alper Buyuktosunoglu is a research staff member at the IBM T.J. Watson Research Center. His research interests include computer architecture, especially power-aware processor microarchitecture, and multicore and multiprocessor systems. Buyuktosunoglu has a PhD in electrical and computer engineering from the University of Rochester. He is a senior member of IEEE and serves on the editorial board of *IEEE Micro*.

Direct questions and comments about this article to Thomas F. Wenisch at twenisch@umich.edu or to Alper Buyuktosunoglu at alperb@us.ibm.com.