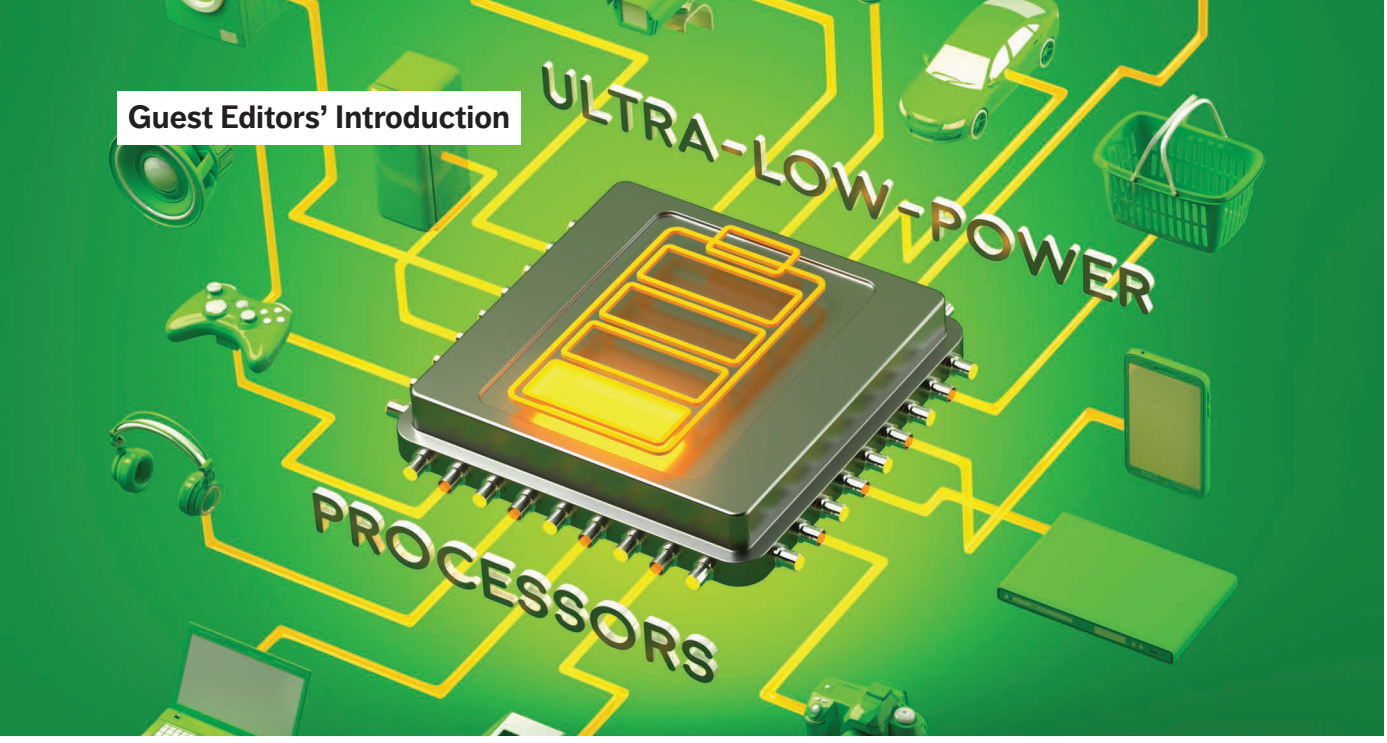


Guest Editors' Introduction

# Ultra-Low-Power Processors

**David Brooks**
*Harvard University*

**John Sartori**
*University of Minnesota*

Society's increasing use of connected sensing and wearable computing has created robust demand for ultra-low-power (ULP) edge computing devices and associated system-on-chip (SoC) architectures. In fact, the ubiquity of ULP processing has already made such embedded devices the highest-volume processor part in production, with an even greater dominance expected in the near future. The Internet of Everything calls for an embedded processor in every object, necessitating billions or trillions of processors. At the same time, the explosion of data generated from these devices, in conjunction with the traditional model of using cloud-based services to process the data, will place tremendous demands on limited wireless spectrum and energy-hungry wireless networks. Smart, ULP edge devices are the only viable option that can meet these demands.

One big area of expansion for ULP processors is the Internet of Things (IoT). Most projections forecast the number of connected IoT devices to grow exponentially, easily reaching over 100 billion within the next decade. Even assuming a conservative ULP power budget of a few milliwatts per device, the total energy consumption of all these connected devices will be over 10 trillion kWh per year. That's more energy than over half of the countries in the world consume in a year. Given the sheer number of devices that will be connected (IoT devices will outnumber humans by more than an


Published by the IEEE Computer Society

order of magnitude), even trying to change or charge all their batteries will become an infeasible task, necessitating more research on energy harvesting to create energy-neutral devices that can fend for themselves by collecting their own energy. Likewise, more research will be needed on novel ways to reduce power dramatically—by an order of magnitude or more, enabling ULP devices to be integrated in more places and in higher quantities. In conjunction with this research on ultra-low power and energy, the fact that all these devices will be connected to the Internet demands more research on energy-efficient security measures. In a world where IoT devices have access to all of our data—personal, health-critical, financial, and all the rest—the attack surface for potential information security leaks becomes larger than ever. With all this critical information entrusted to devices that can barely scrape together enough power to boot up, much less implement a host of security protocols, ensuring information security at ultra-low power and energy levels will be critical.

The articles in this special issue highlight some of the critical research and explore some of the potentially viable solutions that will help to advance the state of the art toward a more power- and energy-efficient future for ULP processing.

## Beyond CMOS

The continuation of CMOS device scaling is of utmost importance to computer architects, and in recent years the perception has been that CMOS scaling has slowed. In "CMOS Scaling Trends and Beyond," Mark T. Bohr and Ian A. Young dispel this notion by showing that through the hard work and ingenuity of device R&D, several new transistor design innovations have been brought to bear on the problem over the past several generations of CMOS technology at Intel. The article also highlights several "beyond-CMOS" technologies that have the potential to complement CMOS by outperforming it in certain niche applications. An example highlighted in the article is Tunnel FETs that can drastically improve the energy-delay product over conventional CMOS. Such devices would be especially attractive for the ULP processors that are the focus of this special issue.

## Implementing a Low-Power Neural Network on Chip

Implementing a neural network in a ULP chip is a challenging feat. Neural networks are one of those applications that require intensive computation that is typically delegated to massively parallel GPGPUs. Nevertheless, in "Low-Power Convolutional Neural Network Processor for a Face-Recognition System," Kyeongryeol Bong and colleagues took on this challenge and fabricated a face-recognition chip based on convolutional neural networks (CNNs) that boasts power consumption of less than a milliwatt for a computation rate of 1 fps. They achieved this low power consumption by splitting the task of face recognition into two stages—face detection, which is performed by a low-power ASIC, and face verification, which is performed by a highly accurate CNN. In their chip, the ULP face-detection circuitry acts as an energy-conscientious gatekeeper for the higher-powered CNN logic, such that the CNN is called on only when needed (that is, when a face has been detected and extracted from an input image). The chip also dynamically adapts its power characteristics using dynamic voltage and frequency scaling based on the number of faces detected to keep power consumption low even under heavy load conditions. The result is a low-power chip that performs a task that's integral to many ULP applications.

## Edge–Cloud Computing

Machine learning is a key component of many IoT systems that must make decisions based on the data they gather in the wild. However, the computationally intense nature of machine learning makes it unsuitable for execution on ULP processors. Typically, massively parallel GPGPUs are used for such computations; however, powering and carrying around a GPU is out of the question for most ULP systems, which are constrained to small form factors, low cost, and ultra-low power and energy budgets. "Flying IoT: Toward Low-Power Vision in the Sky" by Hasan Genc and colleagues explores a computing paradigm in which data are collected at the edge by ULP processors but are processed by high-performance computing resources in the cloud. While this approach enables edge systems

to function within their restrictive constraints, it obviously introduces a communication bottleneck. The article explores the proposed tag-team edge–cloud computing paradigm in a stress-test scenario—a drone that requires real-time results for computations performed in the cloud. The authors investigate how to design ULP systems that can meet real-time deadlines while simultaneously meeting requirements for low power, small form factor, and low cost by harnessing cloud computing intelligently.

## Visual IoT

Visual computing at the edge has clear applications to future IoT devices, with applications ranging from security and surveillance to augmented reality devices. Visual data collected through today's high-resolution cameras tend to demand quite high bandwidth, and the number of such cameras is exploding as low-cost image sensors become common. This means that sending all of the visual data in the cloud for computing is impractical, and edge-based solutions are a growing necessity. In "Visual IoT: Ultra-Low-Power Processing Architectures and Implications," Vui Seng Chua and colleagues describe a mixed-mode approach to such systems, ranging from static image feature detection to dynamic video analytic applications. Neural networks are now an essential component to any type of visual computing application, and the article describes challenges and opportunities with neural network hardware accelerator designs for application in the visual IoT realm.

## Time-Based Stochastic Computing

Stochastic computing is a potentially promising technology for ULP systems because it allows extreme reductions in system hardware for certain functions. For example, a multiplier, which can be synthesized as thousands of gates in a traditional digital circuit, can be implemented with a single logic gate in a stochastic computing circuit. This is an exciting prospect for applications that are amenable to stochastic processing, such as real-time image or video processing, since they can be supported with hardware that has orders of magnitude smaller area and power requirements than traditional hardware architectures for the applications. However, one of the main drawbacks of existing approaches for stochastic computing in the context of ULP processing is that they reduce power and area but increase energy due to the data encoding used, which represents values as a probabilistic bitstream. This potentially makes stochastic computing infeasible for the vast majority of ULP systems, which are severely energy constrained (such as energy harvesting or battery-powered systems). "An Overview of Time-Based Computing with Stochastic Constructs" by M. Hassan Najafi and colleagues provides an overview of a new time-based encoding that uses pulse-width modulation to harness stochastic computing's strengths—namely, ultra-low power and area—for ULP computing while allowing stochastic computing circuits to reach ultra-low energy targets as well.

The articles in this special issue highlight some of the critical research and explore some of the potentially viable solutions that will help to advance the state-of-art toward a more power- and energy-efficient future for ULP processing.

## Ultra-Low-Power Security Constructs

IoT systems will be successful when they become a pervasive element in our society. For this to happen, they will become embedded into our daily lives in areas where security and privacy issues are paramount. For example, if life-saving medical equipment or self-driving cars are susceptible to hacking attacks, practical deployments will be slow due to safety and regulatory concerns. IoT systems are susceptible to multiple attack vectors due both to their placement in potentially hostile environments and their network connectivity requirements. Due to cost reasons, it is also not practical to deploy significant hardware resources to maintain security and privacy. In "Hardware Designs for Security in Ultra-Low-Power IoT Systems: An Overview and Survey," Kaiyuan Yang and colleagues explore a range of low-power and low-cost hardware building blocks that can provide the underpinnings for security and privacy at the higher levels. Examples of such blocks include physically unclonable functions (PUFs) that rely on device properties to provide a unique signature that provides an authentication code for a given system. The article outlines a taxonomy of designs that can be used to develop PUFs and describes several practical hardware implementations of PUFs that have been realized in silicon.

We appreciate all the authors who submitted papers to this issue, and we thank the anonymous reviewers for their efforts. We hope readers will enjoy this special issue of *IEEE Micro*.

**David Brooks** is the Haley Family Professor of Computer Science at Harvard University. Contact him at dbrooks@eecs.harvard.edu.

**John Sartori** is an assistant professor at the University of Minnesota. Contact him at jsartori@umn.edu.