Ultralow Power and the New Era of Not-So-VLSI

Marilyn Wolf
Georgia Institute of Technology

A CONFLUENCE OF application requirements and physics leads us to types of chips to design. A new generation of ultralow-power (ULP) chips will build upon research over the past several decades but these techniques, once considered niche, will rise to the top of the agenda for many organizations.

Microprocessors pay attention to power consumption for two reasons: thermal design power (TDP) and overall current drain. The microprocessor market is driven by high levels of functionality and integration; power and thermal are nuisances. In contrast, the motivation for ULP as a key operational characteristic comes from IoT and related applications. These types of systems are severely limited by the amount of power available. They must also meet very low cost points. IoT is driven by very low costs to create a trillion-sensor market. Since the market is very sensitive to total installed cost, not just the cost of silicon, the cost of the power supply and the cost of wiring become important factors in the design.

Low cost takes on a different spin in the modern era. Low cost was traditionally driven by visions of chip shrinks in subsequent generations. In the classic Moore's Law era, technology nodes resulted in lower cost per transistor. However, many experts and observers agree [5]–[7] that this trend no longer continues. The 28-nm node is widely believed to represent the historically lowest cost per transistor. Cost per transistor can be described by a simple formula

$$\text{cost per transistor} = (\text{cost per wafer})/(\text{number of transistors per wafer}).$$

A multitude of physical phenomena have caused the complexity of complementary metal–oxide–semiconductor (CMOS) processes to increase over the past several decades, resulting in higher wafer processing costs. Below 28 nm, the cost of a processed wafer increases enough to offset the density increases given by feature shrinks. I like to think of this phenomenon as peak silicon, similar to predictions of a period of maximum petroleum production (peak oil) or maximum television production (peak TV).

The long-standing motivation for all the fields surrounding VLSI—computer architecture, circuit design, computer-aided design (CAD)—has been to take advantage of ever-shrinking transistor costs to create ever-more-complex chips. That drive is encapsulated in the very name very large scale integration. Peak silicon changes that. While there may be reasons to build very large chips for certain applications, we can no longer rely on decreasing cost per transistor to enable a mass market for what was once an esoteric design. We must instead focus on chips that make economic sense for known, existing technologies.

ULP use cases

Internet of Things (IoT) is a marketing term that has been applied to many different use cases. Some of the initial applications are really

Digital Object Identifier 10.1109/MDAT.2016.2569433
Date of current version: 17 June 2016.

Internet-of-Thing, not Internet-of-Things. Connecting a device to the Internet and providing a smartphone app for control has its uses but is a relatively straightforward application of embedded computing. And while some of these applications require low-power devices, many add Internet connections to wall-powered devices, making power consumption less of an issue.

ULP comes to play in applications that combine distributed sensing, computing, and perhaps actuation. Adding more sensors and actuators to a large physical system gives much finer grained control over the plant. Putting sensors and actuators near their physical devices reduces delay and improves closed-loop characteristics. To achieve this fine level of granularity requires removing the sensor and actuator nodes away from the comfort of a tethered, powered environment. For example, jet engine manufacturers are moving from centralized control using full-authority digital engine controllers (FADECs) to distributed control [19]. They do so for many of the same reasons that earlier motivated car manufacturers did to introduce digital engine control: lower fuel consumption, lower emissions, and lower maintenance costs. Given the extremely harsh conditions inside a jet engine, distributed control is a powerful motivation for untethered nodes that do not rely on either external power or wired communication.

We can identify two major types of use cases for IoT systems: periodic and event driven.

Periodic sampling and actuation systems make use of more traditional signal processing and control applications. Local computation can be used to close fast loops. Cloud computation is used to fuse data from multiple sensors. Fusing happens at slower rates but provides mode control that is not possible using only local information. Vehicles make extensive use of periodic sampling and actuation, although they also make use of the event-driven model.

Event-driven systems detect and operate on discrete changes. Local computation in this case is used to detect and process events. Local processing reduces power consumption for data processing as well as communication. As with periodic systems, cloud computation fuses data from multiple sensors to provide global, long-term control of the system. One example of an event-driven system is a smart home analysis system we are developing for the long-term care of people with special needs [18]. Sensors around the house capture a variety of information on particular activities: walking through a door, using a faucet, turning on a TV. Local processing can be used over the short term, such as issuing an alert if someone needs help. Algorithms running in the cloud can process the sensor information to create a long-term view of the activity of the residents that can be used to help caregivers optimize the care of each of the residents.

What do these use cases tell us? A key metric for ULP is system energy per sample: the total system energy for a sample, including data acquisition, computation, and communication. This metric depends on system architecture and is enabled by device characteristics. System architecture partitions functionality across the network; a poor system partitioning cannot be corrected by clever node design. However, node designs must be able to deliver the required mix of functionality and operational characteristics.

One common characteristic of many IoT applications that is not so obvious from use case analysis is the need for low cost of ownership. Many of these systems will never be built if their cost is too high. Cost considerations take into account not just the components, but also installation and operation. Wireless, self-powered sensors and actuators substantially reduce the cost of system installation. Installing a wired connection in a U.S. building typically costs about $100 per drop. Eliminating power and network wires allows nodes to be placed much more cheaply.

Technology

A variety of technology factors play into the design of ULP devices. Wafer manufacturing is part of the picture, but packing also plays an important role. Which combination of technologies is the ultimate winner for ULP devices will be determined in large part by cost.

Two different device structures have been proposed to reduce leakage in the microprocessor space. FinFETs have been widely adopted for advanced nodes. The vertical structure of the finFET helps the gate to better control the channel charge. An alternative approach is FD-SOI, which uses a planar structure but with a buried insulator. Manufacturers have not yet decided to update 28-nm processes for these new devices.

Nonvolatile storage is a critical component technology for ULP nodes. Devices that frequently sleep need to be able to store state without burning power to do so. Flash is widely available on many standard processes. New devices, such as resistive RAM, may provide new capabilities and a different set of tradeoffs to system architects.

The IoT device must include not just logic but also analog plus MEMS devices for sensing and actuation, communication antennas, and power management. Interposers are one way to provide the mix of technologies required and at a reasonable cost. Interposers made of any of several substrates can provide a cheap, small, and light way to combine several chips from diverse technologies and even multiple manufacturers. Interposer-based integration is often known as 2.5-D. In the short term, it allows system houses to build cost-effective nodes. In the long term, 2.5-D may be the best long-term approach as well.

Circuits

ULP system designers must simultaneously pursue two strategies: reducing the power consumption of circuits, and gathering and managing power.

Near-threshold logic and subthreshold logic are well-established methods to reduce CMOS power consumption. As logic levels move closer to the transistor thresholds, errors start to increase. Errors are introduced by larger delays caused by lower operating points of the transistors; as the delays of critical paths exceed the clock period, the logic produces wrong answers. Kim et al. [4] gave an error model for subthreshold logic and showed that misexercising these critical paths depends not only on the current inputs to the logic but also the previous inputs. The prior inputs to the logic determine the state of the nodes; a slow path creates problems only if the new inputs require that node to change.

Adiabatic logic has been studied for several decades as an alternative to static CMOS for low-power operation [1], [3]. Adiabatic logic is inspired by the reversibility of physical phenomenon, which leads to asymptotically low energy per operation. The study of reversibility as a means to minimize the energy required for computation goes back to the earliest days of computer science [13]. Adiabatic logic has several drawbacks, notably its high overhead. It requires complementary signals. Reversible computation also requires more complex structures; Benioff, for example, designed a reversible Turing machine that is considerably more complex than the classical version [17]. Reversibility adds significant area overhead. Circuit considerations also mean that in some cases traditional CMOS provides lower total energy consumption than does adiabatic logic. However, the relatively low data rates of many IoT applications makes a natural fit with adiabatic logic. And adiabatic logic also provides reduced susceptance to side channel attacks on cryptography, as demonstrated by a University of Michigan team [3].

Paradiso and Starner surveyed the state of energy scavenging a decade ago [8]. Since then, energy scavenging has entered the mainstream in both frivolous applications such as self-lighting shoes as well as serious applications such as medical electronics. Implementing energy harvesting requires not only efficient transducers but also power supplies that can convert harvested energy into the voltage levels required for circuit operation [20].

Image sensors provide one interesting case of energy scavenging. We can use the same image sensor to scavenge energy from light and to capture images. While scavenging does not provide enough energy to run at 30 frames/s, the sensor can run on the order of 1 frame/s, which is fast enough for many applications. For example, Hanson and Sylvester [9] designed a 128×128 pixel sensor that operates over power supply voltages of 0.45–0.7 V with energy consumption of 140 nJ per frame at 8.5 frames/s. Tang et al. [10] describe an energy-harvesting CMOS image sensor with an energy-harvesting mode that can harvest 80 nA at 350 lux and 9.7 μA at 3500 lux and consumes 10 μW at 10 frames/s; one frame can be processed using 200 ms of energy harvesting.

Architectures

Node architectures are the interface between IoT systems and VLSI. ULP device architectures are, like systems-on-chips, heterogeneous. But the mix of technologies used for ULP is decidedly different from the SoC case.

Analog/digital codesign is an intriguing approach for sensing and actuating nodes. The traditional limitations of analog computation are its

lack of flexibility and its limited dynamic range. Flexibility is not an issue when algorithms are well defined. And limited dynamic range can be countered by tighter control loops. However, to make best use of analog processing, we need to develop improved techniques to codesign analog and digital components. Imaging is one area in which certain types of processing are well suited to analog computation. Liu et al. [11] designed a CMOS image sensor that performs motion tracking using 1-D Gaussian filters to smooth the image to generate row and column events. Dubois et al. [12] designed a CMOS image sensor for high-speed gradient analysis using an analog multiplier; the analog circuit output could be used to implement the Sobel operator and Laplacian.

How much CPU does an ULP device need? A back-to-the-future approach for onboard computation has some appeal, particularly for devices that push signal processing onto analog units. Early microprocessors were very small: the 6502 required 3510 transistors, and the Z80 was built from 8500 transistors. These microprocessors did not provide sophisticated memory management units, floating-point computation, or other features to which system designers have become accustomed. But keep in mind that they were used to build what were at the time state-of-the-art systems.

Nonvolatile state is critical for the operation of ULP nodes. The device must be able to restart itself with knowledge of its previous state. Both reading and writing must have reasonable energy costs. Depending on the characteristics of the device, state may need to be copied from the nonvolatile devices to traditional CMOS volatile storage.

Security must be baked into many systems and IoT is a poster child for the need for modern security-driven architectures. However, security must be delivered at appropriate costs in both energy and area. AES is widely used for encryption. Hamalainen et al. [16] developed an AES implementation with low area and energy consumption. SIMON is a block cypher algorithm designed by the National Security Agency (NSA) for low-area, lightweight applications. Aysu et al. [15] built a very area-efficient implementation of SIMON. The adiabatic encryption unit mentioned above also offers intriguing results for power consumption.

Nodes must be able to communicate with other nodes and with the clouds. Much of this communication will be conducted indirectly to minimize energy costs. A hub structure links ULP nodes to hubs that provide more computational and communication capabilities. Hubs can enable communication between local nodes or to and from the cloud. Bluetooth low energy (BLE) is a very carefully engineered standard for low-power communication. Its widespread adoption is a measure of the success with which it solves key problems in low-rate, low-power communication.

Ultralow power VLSI systems are exploding from niche applications to the mainstream thanks to the popularity of the IoT as an application paradigm. ULP devices represent a shift in focus from highly integrated devices to right-sized devices. Building a new generation of ULP devices will require the development of some new design technologies as well as the exploitation of many mature techniques. ■

Acknowledgments

This work was supported by the National Science Foundation under Grant 1513404.

References

[1] W. C. Athas, L. J. Svensson, J. G. Koller, N. Tzartzanis, and E. Ying-Chin Chou, "Low-power digital systems based on adiabatic-switching principles," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 2, no. 4, pp. 398–407, Dec. 1994.

[2] Y. Moon and D.-K. Jeong, "An efficient charge recovery logic circuit," *IEEE J. Solid-State Circuits*, vol. 31, no. 4, pp. 514–522, Apr. 1996.

[3] S. Lu, Z. Zhang, and M. Papaefthymiou, "1.32 GHz high-throughput charge-recovery AES core with resistance to DPA attacks," in *Proc. Symp. VLSI Circuits*, Kyoto, Japan, 2015, pp. C246–C247.

[4] S. H. Kim, S. Mukhopodhyay, and M. Wolf, "System-level energy optimization for error-tolerant image compression," *IEEE Embedded Syst. Lett.*, vol. 2, no. 3, pp. 81–84, Sep. 2010.

[5] J. Hruska, "Nvidia deeply unhappy with TSMC, claims 20 nm essentially worthless," *ExtremeTech.com*, Mar. 23, 2012. [Online]. Available: http://www.extremetech.com/computing/123529-nvidia-deeply-unhappy-with-tsmc-claims-22nm-essentially-worthless

[6] Z. Or-Bach, "28 nm: The last node of Moore's Law," *EE Times*, Mar. 19, 2014. [Online]. Available: http://www.eetimes.com/author.asp?doc_id = 1321536

[7] "Beyond Moore's Law," *The Economist*, May 26 2015. [Online]. Available: http://www.economist.com/news/science-and-technology/21652051-even-after-moores-law-ends-chip-costs-could-still-halve-every-few-years-beyond

[8] J. A. Paradiso and T. Starner, "Energy scavenging for mobile and wireless electronics," *IEEE Perv. Comput.*, vol. 4, no. 1, pp. 18–27, Jan.–Mar. 2005.

[9] S. Hanson and D. Sylvester, "A 0.45 μm 0.7 V sub-microwatt CMOS image sensor for ultra-low power applications," in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2009, pp. 176–177.

[10] F. Tang, Y. Cao, and A. Bermak, "An ultra-low power current-mode CMOS image sensor with energy harvesting capability," in *Proc. ESSCIRC*, Sep. 2010, pp. 126–129.

[11] X. Liu, M. Zhang, and J. Van der Spiegel, "A low power multi-mode CMOS image sensor with integrated on-chip motion detection," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2013, pp. 2416–2419.

[12] J. Dubois, D. Ginhac, and M. Paindavoine, "A single-chip 10000 frames/s CMOS sensor with in-situ 2D programmable image processing," in *Proc. IEEE Int. Workshop Comput. Architect. Mach. Percept. Sens.*, 2006, pp. 124–129.

[13] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM J. Res. Develop.*, vol. 5, no. 3, pp. 183–191, 1961.

[14] J. Zhang, S. Iyer, X. Zheng, P. Schaumont, and Y. Yang, "Hardware-software co-design for heterogeneous multiprocessor sensor nodes," in *Proc. IEEE Global Commun. Conf.*, Austin, TX, USA, 2014, pp. 20–25.

[15] A. Aysu, E. Gulcan, and P. Schaumont, "SIMON says: Break area records of block ciphers on FPGAs," *IEEE Embedded Syst. Lett.*, vol. 6, no. 2, pp. 37–40, Jun. 2014.

[16] P. Hamalainen, T. Alho, M. Hannikainen, and T. D. Hamalainen, "Design and implementation of low-area and low-power AES encryption hardware core," in *Proc. 9th EUROMICRO Conf. Digital Syst. Design*, Dubrovnik, 2006, pp. 577–583.

[17] P. Benioff, "Quantum mechanical models of Turing machines that dissipate no energy," *Phys. Rev. Lett.*, vol. 48, no. 23, pp. 1581–1585, Jun. 7, 1982.

[18] M. Wolf, M. van der Schaar, H. Kim, and J. Xu, "Caring analytics for adults with special needs," *IEEE Design Test*, vol. 32, no. 5, pp. 35–44, Oct. 2015.

[19] M. Pakmehr et al., "Distributed architectures integrated with high-temperature electronics for engine monitoring and control," in *Proc. 47th AIAA/ASME/SAE/ASEE Joint Propulsion Conf.*, San Diego, CA, USA, Jul. 31–Aug. 3, 2011. [Online]. Available: http://dx.doi.org/10.2514/6.2011-6148

[20] K. Z. Ahmed, M. Kar, and S. Mukhopadhyay, "(Invited paper) energy delivery for self-powered IoT devices," in *Proc. 21st Asia South Pacific Design Autom. Conf.*, Macau, 2016, pp. 302–307.

Marilyn Wolf is Rhesa S. "Ray" Farmer Distinguished Chair in Embedded Computing Systems and Georgia Research Alliance Eminent Scholar at the Georgia Institute of Technology, Atlanta, GA, USA. Her research interests include cyber–physical systems, embedded computing, embedded video and computer vision, and VLSI systems. Wolf has a PhD in electrical engineering from Stanford University, Stanford, CA, USA (1984). She is a Fellow of the IEEE and the Association for Computing Machinery (ACM) and an IEEE Computer Society Golden Core member.

■ Direct questions and comments about this article to Marilyn Wolf, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA; wolf@ece.gatech.edu.