Reconfigurable Architectures for Data Analytics on Next-Generation Edge-Computing Platforms

by Darshika G. Perera,
University of Colorado, Colorado Springs, USA

In the Internet of Things (IoT) era, there will be an enormous amount of data generated from various sensors and other similar devices physically distributed throughout networks and systems, such as smart grids, smart homes, and autonomous vehicles [1], [2]. Today, most of these networks (or systems) often collect and send these data directly to the cloud infrastructure, which comprises centralized data centers for processing, analyzing, and storing the data. However, traditional cloud infrastructures face serious challenges when transmitting, processing, and analyzing this massive amount of data. As illustrated in Figure 1, these challenges include insufficient bandwidth, high latency, unsatisfactory real-time response, high power consumption, and privacy-protection issues [3]. Edge-centric computing is emerging as a complementary solution to address the aforementioned issues of the cloud infrastructure [4], thus reducing (or eliminating) the overall communication overhead and response latency of the corresponding networks and systems and enhancing the performance (i.e., processing power) and the scalability of systems [5].

Edge Computing

With edge computing, data processing and analysis can be done closer to the source of the data (i.e., at the edge of the networks, as shown in Figure 2), which, in turn, enables real-time, in situ data analytics and processing [1], [2]. Furthermore, preprocessing the data at the edge, prior to sending these data to the cloud, addresses the insufficient bandwidth issues [4]. Processing and analyzing the data at the edge can also enhance the privacy and security of the data compared to that of the cloud infrastructure because the raw data are not transmitted through unsecured networks of the cloud [4]. In addition, wireless communication modules utilized to transmit the data from the edge devices to the cloud are often power hungry, whereas the edge devices are typically energy constrained; hence, performing some data analytics

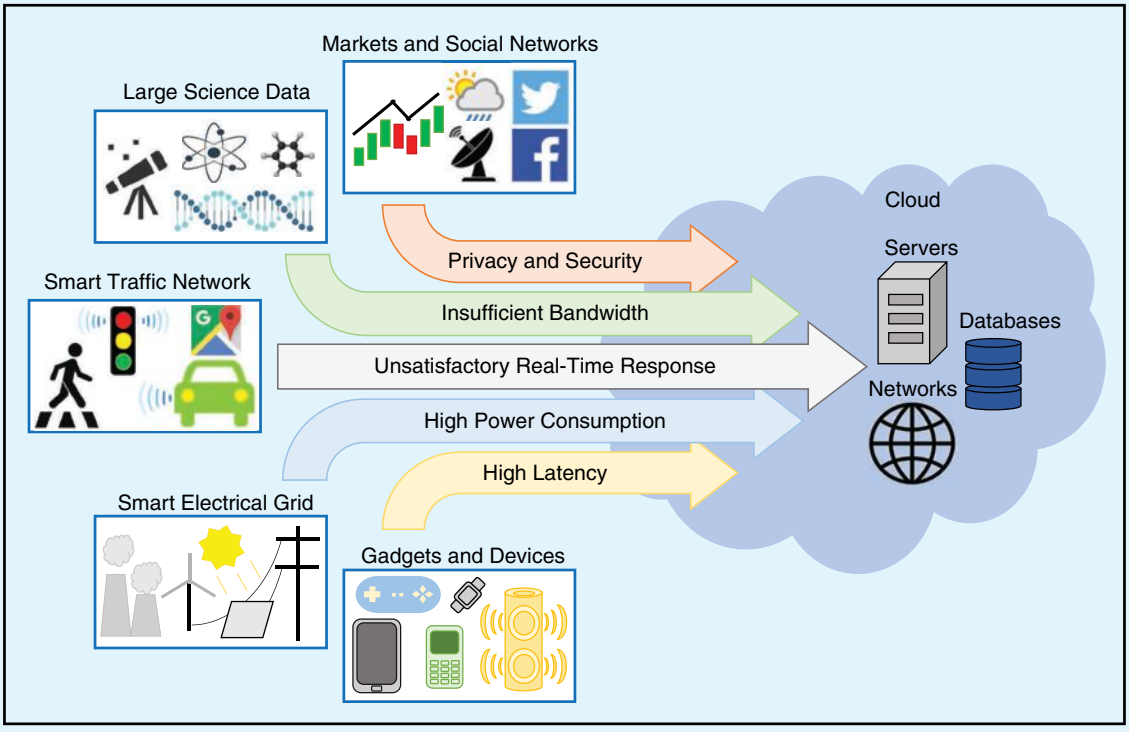

Figure 1: A conventional cloud infrastructure faces serious challenges when transmitting, processing, and analyzing massive amounts of data generated from variety of sources. These challenges include insufficient bandwidth, high latency, unsatisfactory real-time response, high power consumption, and privacy-protection issues.

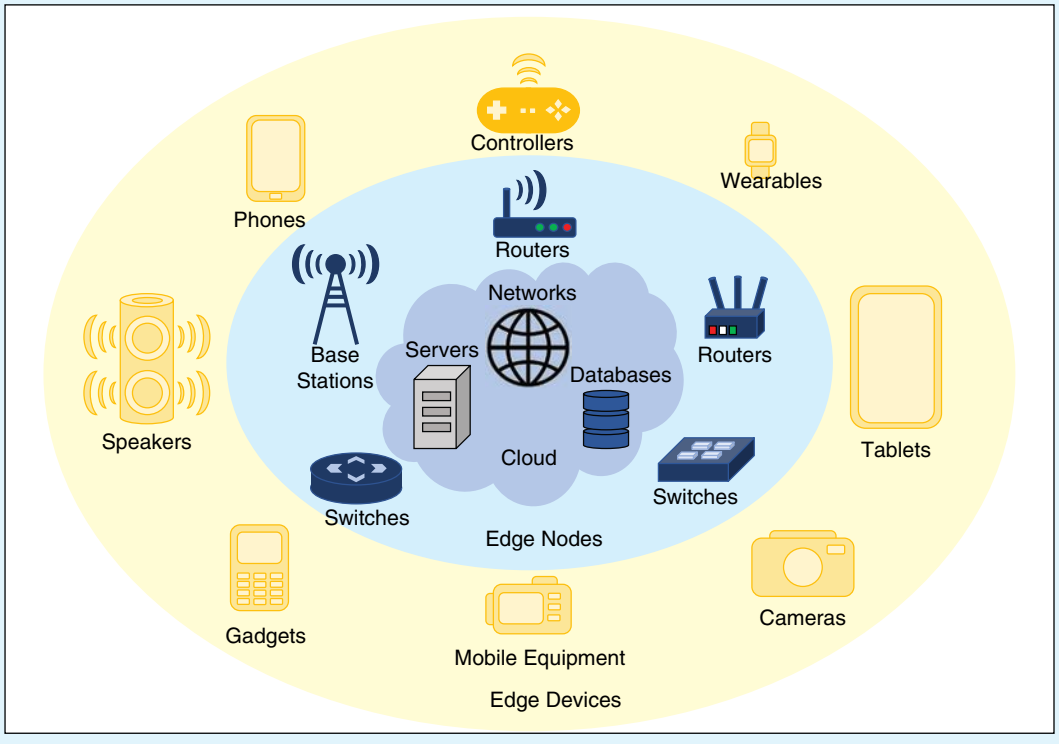

Figure 2: The concept of edge computing. The cloud infrastructure often consists of large data centers, server farms, and data storage. The cloud infrastructure can be used as the final stage for processing, analyzing, and storing big data. The edge nodes typically consist of routers, base stations, and switches. The edge nodes can be used as the intermediate stage for certain types of data analysis/mining prior to sending the data to the cloud. The edge devices usually consist of wearables, smartphones, cameras, and many other smart devices. The edge devices can be used as the initial stage for real-time data processing, including data preprocessing, data filtering, optimization, on-the-fly decision making, and so on. The edge devices comprise various sensors for real-time data collections. The edge devices can be used to connect to the cloud, either via edge nodes or directly.

and processing at the edge would be more energy efficient [1].

Real-Time Data Analytics

In many scenarios and often in experimental setups, data scientists and data analysts process enormous amounts of data without considering the real-time constraints. In these cases, the data are typically being processed offline (not in real time) for days, or even months, depending on the processing power of the computing platforms utilized. However, for many real-world scenarios, the data need to be processed in situ with real-time constraints, and the results must be available within strict time limits, for the subsequent analysis or actions to reap the actual benefits [6]. Real-time and in situ data analysis and processing are imperative in many edge-computing applications, including smart energy grid management and autonomous vehicles. For instance, autonomous vehicles typically produce gigabytes of data per second, and these data need to be processed in real time for these vehicles to make correct, split-second decisions on the fly [1]. In this case, the high (round-trip) latency associated with sending the data to the cloud and processing and receiving a response can lead to catastrophic scenarios [1].

Issues of Processor-Based Systems

Most of the existing algorithms and techniques for data mining/analytics are typically processor-based, software-only designs [7], [8]. They are designed for general-purpose computers (such as desktops and servers) and consist of CPUs; thus, they are often incapable of analyzing and processing enormous amounts of data efficiently and effectively. As a result, these software algorithms and techniques might be incapable of being executed directly on next-generation edge-computing platforms in their current form. These processor-based designs have high execution overhead because each instruction has to be fetched from the memory, decoded, and then executed [9]. Furthermore, due to this instruction fetch-decode-execute cycle overhead and because of general-purpose circuits, the power consumption of a processor-based design is much higher than special-purpose (or customized) hardware [9]. A survey done in [10] demonstrated that pure processor-based, software-only computing platforms, including multiprocessor, multicore, general-purpose GPUs, are simply not sufficient to handle this enormous amount of data.

As stated in [11] and [12], at the edge, the processing power needed to analyze and process such an enormous amount of data will soon exceed the growth rate of Moore's law. As a result, edge-computing frameworks and solutions, which currently solely consist of CPUs, will be inadequate to meet the required processing power. Although they could potentially meet the processing power requirements, GPUs are extremely power hungry and have limited energy efficiency [13].

Need for Novel Solutions

All of the aforementioned facts illustrate that the existing algorithms and techniques used for data mining/analytics and the conventional computing platforms utilized for current edge-computing platforms will not suffice to process and analyze this ever-increasing data as well as to handle the associated complex computational problems efficiently and effectively.

Consequently, novel, unique, and innovative architectures, methodologies, and techniques are required to propel edge computing from its infancy to support and accelerate real-time, in situ data mining/analytics on next-generation edge-computing platforms that are heterogeneous in nature, considering the constraints associated with the computing platforms and requirements of the edge applications.

Special-Purpose Architectures

In this case, apart from algorithmic development, it is imperative to incorporate some special-purpose (or customized) hardware (and/or software) architectures and techniques into next-generation edge-computing platforms. Unlike general-purpose, process-based designs, customized hardware architectures are optimized for specific applications and avoid the instruction fetch-decode-execute cycle overhead. Furthermore, special-purpose hardware provides superior speed performance and consumes less power than equivalent software running on a general-purpose processor [9]. As operating frequency is directly proportional to dynamic power consumption, high-performance processors often consume high power [9]. Also, customized circuits are usually function specific and occupy less hardware space on a chip compared to the circuitry of a general-purpose processor. As a result, they are more suitable for resource-constrained edge-computing platforms.

Reconfigurable Architectures

To provide the flexibility required in an ever-changing application environment, it is also crucial to incorporate reconfigurable (adaptive) hardware (and/or software) architectures into next-generation edge-computing platforms. Reconfigurable hardware has advantages similar to those of special-purpose hardware: providing efficient and customized circuits and avoiding the instruction fetch-decode-execute cycle overhead as in a processor, thus leading to low power and high performance [9]. Apart from having advantages similar to those of customized hardware, reconfigurable computing systems, such as field-programmable gate arrays (FPGAs), have added advantages, including reusing the on-chip hardware circuitry to perform variety of tasks and reducing the time to market and design time due to their flexibility and programmability postfabrication. Our analysis [9] demonstrated that FPGA-based reconfigurable hardware provides numerous advantages, including flexibility, durability, upgradeability, compact circuits, reduced time to market, and relatively low cost, which are important to support real-time, in situ data analytics/mining on next-generation edge-computing platforms. In this case, multiple applications and tasks can be executed on a single FPGA by dynamically reconfiguring the hardware on chip from one application/task to another as needed.

Real-time and in situ data analysis and processing are imperative in many edge-computing applications, including smart energy grid management and autonomous vehicles.

FPGA-Based Edge Computing

Several studies illustrate that FPGA-based systems are being considered as viable solutions for future edge-computing platforms and devices [14], [15]. There are several key advantages to adopting FPGAs for edge computing over other computing systems. For instance, FPGAs provide superior performance because

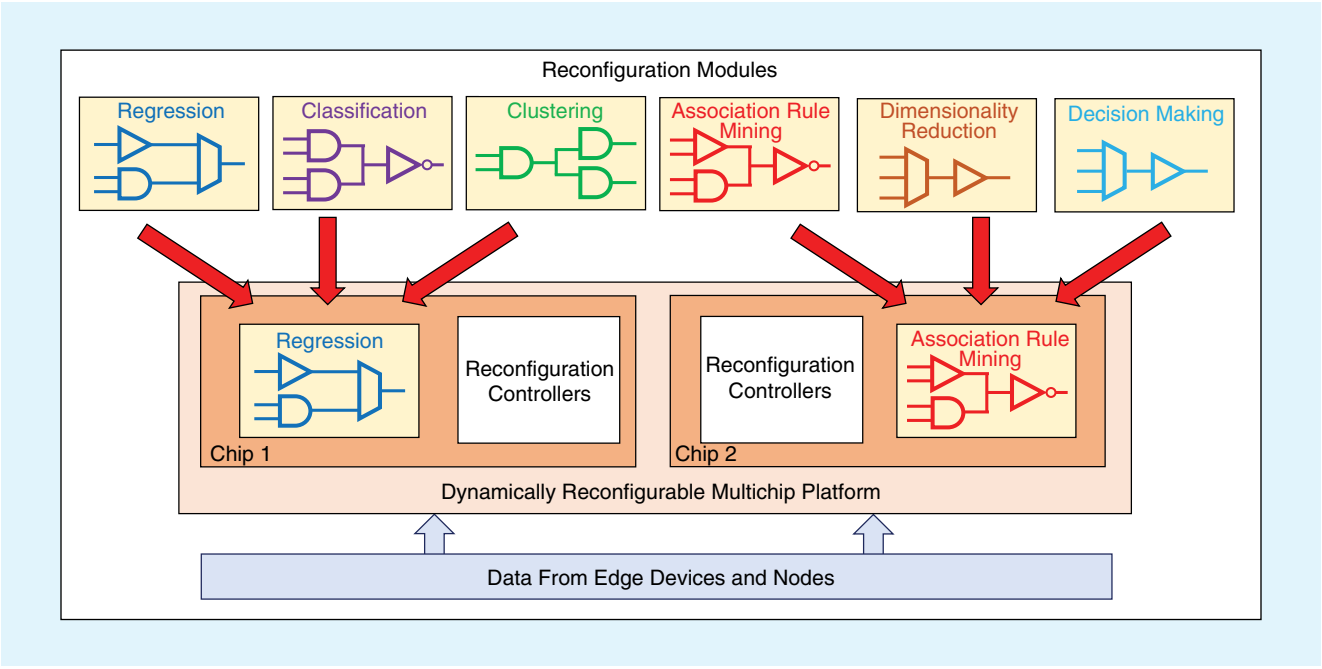

Figure 3: An overview of the next-generation edge-computing platform. As shown, this platform will comprise multiple FPGAs. Also, architectures and techniques will be created to dynamically reconfigure the whole platform globally, and to dynamically reconfigure some systems (e.g., individual chips) locally, to provide the necessary self-adaptive, self-healing, and self-correcting traits as well as the postdesign and postdeployment optimizations and upgrades required for edge applications. Based on the requirements of the edge application, different data analytic/mining algorithms (shown in the reconfiguration modules) will be dynamically swapped in and out of the multichip platform to perform a variety of applications/tasks.

they leverage spatial and temporal parallelism and fine- and coarse-grain parallelism in computations and in massive scale [16]–[18], which, in turn, would accelerate high-concurrency, high-dependency algorithms and applications at the edge [12]. In addition, FPGAs typically provide a steady throughput to the application's workload size (unlike GPUs) and also with a reduced latency, which is imperative for integrating and addressing requests from various IoT sensors (or devices) in the network [19]. Furthermore, FPGAs consume significantly less power than that of CPUs and GPUs to execute the same edge applications, which is crucial for small-footprint edge devices/platforms [20]. Most importantly, FPGAs provide hardware flexibility (i.e., reconfigurable postfabrication); hence, with FPGA-based edge computing, we can dynamically upgrade processing power, enable scalability, modify and optimize the features and resources as needed, and adapt to any algorithmic characteristics, which is important for edge workloads and applications [21]–[23]. The aforementioned facts illustrate that FPGA-based systems comprise numerous traits and advantages that are important for realizing next-generation edge-computing platforms.

As operating frequency is directly proportional to dynamic power consumption, high-performance processors often consume high power.

Envisioning Next-Generation Edge-Computing Platforms

We envision that next-generation, heterogeneous edge-computing platforms will ultimately consist of multiple computing systems, especially including multiple FPGAs that provide the required reconfigurability, and also GPUs, application-specified integrated circuits, and CPUs. These multiple systems will be distributed throughout the platform with distributed and shared memory systems and connected by a network of interconnections. Furthermore, we envision that novel and innovative techniques and methodologies will be introduced and incorporated to make next-generation edge-computing platforms smart and autonomous enough to seamlessly and independently process and analyze data in real time, with minimal or no human intervention. As illustrated in Figure 3, the architectures and techniques will be created to dynamically reconfigure the whole platform globally, and to dynamically reconfigure some systems locally, to provide the necessary self-adaptive, self-healing, and self-correcting traits required for edge applications. The global and local reconfigurations (as in Figure 3) will be performed dynamically without interrupting the system operations. The interconnection network will also be reconfigurable. We also envision that machine and deep learning techniques will be created and incorporated to further enhance the smart, autonomous, and adaptive features of the next-generation edge-computing platforms. ■

References

[1] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, Oct. 2016. doi: 10.1109/JIOT.2016.2579198.

[2] J. Pan and J. McElhannon, "Future edge cloud and edge computing for internet of things

applications," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 439–449, Feb. 2018. doi: 10.1109/JIOT.2017.2767608.

[3] B. Varghese, N. Wang, S. Barbhuiya, P. Kilpatrick, and D.S. Nikolopoulos, "Challenges and opportunities in edge computing," in *Proc. IEEE Int. Conf. Smart Cloud*, Nov. 2016, pp. 20–26. doi: 10.1109/SmartCloud.2016.18.

[4] T. Yaofeng, D. Zhenjiang, and Y. Hongzhang, "Key technologies and applications of edge computing," *ZTE Commun*., vol. 15, no. 2, pp. 26–34, Apr. 2017.

[5] M. Satyanarayanan, "The emergence of edge computing," *Computer*, vol. 50, no. 1, pp. 30–39, 2017. doi: 10.1109/MC.2017.9.

[6] "Real-time IoT stream processing and large-scale data analytics for smart city applications." CityPulse. www.ict-citypulse.eu

[7] A. A. Liu, Y. T. Su, W. Z. Nie, and M. Kankanhalli, "Hierarchical clustering multi-task learning for joint human action grouping and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 102–104, Jan. 2017. doi: 10.1109/TPAMI.2016.2537337.

[8] K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, and W. Xiang, "Big data-driven optimization for mobile networks toward 5G," *IEEE Netw.*, vol. 30, no. 1, pp. 44–51, Jan./Feb. 2016. doi: 10.1109/MNET.2016.7389830.

[9] D.G. Perera and K.F. Li, "Analysis of single-chip hardware support for mobile and embedded applications," in *Proc. IEEE Pacific Rim Int. Conf. Commun., Comput., Signal Process., (PacRim'13)*, Aug. 2013, pp. 366–376. doi: 10.1109/PACRIM.2013.6625505.

[10] D. Singh and C. Reddy, "A survey on platforms for big data analytics," *J. Big Data*, vol. 2, no. 1, p. 8, 2014. doi: 10.1186/s40537-014-0008-6.

[11] F. Peper, "The end of Moore's law: Opportunities in natural computing?" *Next Gener. Comput.*, vol. 35, no. 3, pp. 259–269, July 2017. doi: 10.1007/s00354-017-0020-4.

[12] S. Biookaghazadeh, M. Zhao, and F. Ren, "Are FPGAs suitable for edge computing?" in *Proc. Workshop on Hot Topics Edge Comput. (HotEdge'18)*, July 2018.

[13] M. P. Puig, L. De Giusti, M. Naiouf, and A. De Giusti, "GPU performance and power consumption analysis: A DCT based denoising application," in *Proc. 23rd Congreso Argentino de Ciencias de la Computación*, Oct. 2017, pp. 185–195.

[14] A. Rodríguez, J. Valverde, J. Portilla, A. Otero, T. Riesgo, and E. de la Torre, "FPGA-based high-performance embedded systems for adaptive edge computing in cyber-physical systems: The ARTICo3 framework," *Sensor J.*, vol. 18, no. 6, p. 1877, June 2018. doi: 10.3390/s18061877.

[15] Z. Zhe, J. Zhang, J. Zhao, J. Cao, D. Zhao, G. Jia, and Q. Meng, "A hardware and software task-scheduling framework based on CPU+FPGA heterogeneous architecture in edge computing," *IEEE Access,* vol. 7, pp. 148,975–148,988, Sept. 2019. doi: 10.1109/ACCESS.2019.2943179.

[16] S. N. Shahrouzi and D. G. Perera, "Optimized hardware accelerators for data mining applications on embedded platform: Case study principal component analysis," *Microprocess. Microsyst.*, vol. 65, pp. 79–96, Mar. 2019. doi: 10.1016/j.micpro.2019.01.001.

[17] A. K. Madsen and D. G. Perera, "Efficient embedded architectures for model predictive controller for battery cell management in electric vehicles," *EURASIP J. Embedded Syst.,* vol. 2018, July 2018, Art. no. 2.

[18] D.G. Perera and K.F. Li, "Parallel computation of similarity measures using an FPGA-based processor array," in *Proc. 22nd IEEE Int. Conf. Adv. Informat. Netw. Appl., (AINA'08)*, Okinawa, Japan, Mar. 2008, pp. 955–962. doi: 10.1109/AINA.2008.97.

[19] J. Cong, Z. Fang, M. Lo, H. Wang, J. Xu, and S. Zhang, "Understanding performance differences of FPGAs and GPUs," in *Proc. Field Programmable Gate Array (FPGA'18)*, Feb. 2018, pp. 93–96. doi: 10.1145/3174243.3174970.

[20] M. Qasaimeh, K. Denolfy, J. Loy, K. Vissersy, J. Zambreno, and P. H. Jones, "Comparing energy efficiency of CPU, GPU and FPGA implementations for vision kernels," in *Proc. IEEE Int. Conf. Embedded Softw. Syst. (ICESS)*, 2019, pp. 1–8. doi: 10.1109/ICESS.2019.8782524.

[21] S. N. Shahrouzi and D. G. Perera, "Dynamic partial reconfigurable hardware architecture for principal component analysis on mobile and embedded devices," *EURASIP J. Embedded Syst.,* vol. 2017, Feb. 2017, Art. no. 25. doi: 10.1186/s13639-017-0074-x.

[22] A. Alkamil and D. G. Perera, "Towards dynamic and partial reconfigurable hardware architectures for cryptographic algorithms on embedded devices," *IEEE Access*, vol. 8, pp. 221,720–221,742, Dec. 2020. doi: 10.1109/ACCESS.2020.3043750.

[23] D.G. Perera and K.F. Li, "FPGA-based reconfigurable hardware for compute intensive data mining applications," in *Proc. 6th IEEE Int. Conf. P2P, Parallel, Grid, Cloud, Internet Comput., (3PGCIC'11)*, Oct. 2011, pp. 100–108. doi: 10.1109/3PGCIC.2011.25.

About the Author

Darshika G. Perera (darshika.perera@uccs.edu) is a fellow Canadian. She received her Ph.D. degree in electrical and computer engineering from the University of Victoria, Canada, and her M.Sc. and B.Sc. degrees in electrical engineering from the Royal Institute of Technology, Sweden, and the University of Peradeniya, Sri Lanka, respectively. She is currently an assistant professor in the Department of Electrical and Computer Engineering, University of Colorado, Colorado Springs, Colorado (UCCS), USA, and also an adjunct assistant professor in the Department of Electrical and Computer Engineering, University of Victoria, Canada. Prior to joining UCCS, she worked as the senior engineer and group leader of embedded systems at CMC Microsystems, Canada. Her research interests include reconfigurable computing, mobile and embedded systems, data mining, and digital systems. She received the Teacher of the Year—Tenure Track Award from the Engineering and Applied Science Collage, UCCS, in April 2019. She also received a best paper award at the IEEE 3PGCIC'11 conference in 2011 and a best poster paper award at the ACM HEART'18 symposium in 2018. She serves on organizing and program committees for several IEEE/ACM conferences and workshops and as a reviewer for several IEEE, Springer, and Elsevier journals. Currently, she serves on the IEEE VLSI Systems and Applications Technical Committee of the IEEE Circuits and Systems Society, and also serves as an associate editor of the Elsevier *Microelectronics Journal*. She is a Senior Member of IEEE; the IEEE Circuits and Systems, Very Large-Scale Integration, and Computer Societies; and IEEE Women in Engineering.