Effectively tackling the upcoming "zettabytes" data explosion requires a huge quantum leap in our computing power and energy efficiency. However, with the Moore's law dwindling quickly, the physical limits of CMOS technology make it almost intractable to achieve high energy efficiency if the traditional "deterministic and precise" computing model still dominates. Worse, the upcoming data explosion mostly comprises statistics gleaned from uncertain, imperfect real-world environment. As such, the traditional computing means of first-principle modeling or explicit statistical modeling will very likely be ineffective to achieve flexibility, autonomy, and human interaction. The bottom line is clear: given where we are headed, the fundamental principle of modern computing has to be reexamined, and transformative changes in the foundation of modern computing must be made.

This dissertation presents a novel stochastic-based computing methodology. It efficiently realizes the algorithmic computing through the proposed concept of Probabilistic Domain Transform (PDT). The essence of PDT approach is to encode the input signal as the probability density function, perform stochastic computing operations on the signal in the probabilistic domain, and decode the output signal by estimating the probability density function of the resulting random samples. The proposed architecture possesses many notable advantages. Specifically, it uses much simplified circuit units to conduct complex operations, which leads to highly area and energy efficient designs suitable for parallel processing. Moreover, it is highly fault-tolerant because the information to be processed is encoded with a large ensemble of random samples. As such, the local perturbations of its computing accuracy will be dissipated globally, thus becoming inconsequential to the final overall results. Finally, the proposed probabilistic-based computing can facilitate building scalable precision systems, which provides an elegant way to trade-off between computing accuracy and computing performance/hardware efficiency for many real-world applications.

To validate the effectiveness of the proposed PDT methodology, two important signal processing applications, discrete convolution and 2-D FIR filtering, are first implemented and benchmarked against other deterministic-based circuit implementations. Furthermore, a large-scale Convolutional Neural Network (CNN), a fundamental algorithmic building block in many computer vision and artificial intelligence applications that follow the deep learning principle, is also implemented with FPGA based on a novel stochastic-based and scalable hardware architecture and circuit design. The key idea is to implement all key components of a deep learning CNN, including multi-dimensional convolution, activation, and pooling layers, completely in the probabilistic computing domain. The proposed architecture not only achieves the advantages of stochastic-based computation, but can also solve several challenges in conventional CNN, such as complexity, parallelism, and memory storage.
Approved for distribution by Mingjie Lin, Committee Chair, on June 20, 2016.

The public is welcome to attend.