A Cross-Layer Perspective for Energy Efficient Processing

- From beyond-CMOS Devices to Deep Learning

X. Sharon Hu

University of Notre Dame
Notre Dame, Indiana, USA
shu@nd.edu

ABSTRACT

As Moore's Law based device scaling and accompanying performance scaling trends are slowing down, there is increasing interest in new technologies and computational models for fast and more energy-efficient information processing. Meanwhile, there is growing evidence that, with respect to traditional Boolean circuits and von Neumann processors, it will be challenging for beyond-CMOS devices to compete with the CMOS technology. Nevertheless, some beyond-CMOS devices demonstrate other unique characteristics such as ambipolarity, negative differential resistance, hysteresis, and oscillatory behavior. Exploiting such unique characteristics, especially in the context of alternative circuit and architectural paradigms, has the potential to offer orders of magnitude improvement in terms of power, performance and capability.

In order to take full advantage of beyond-CMOS devices, however, it is no longer sufficient to develop algorithms, architectures and circuits independent of one another. Cross-layer efforts spanning from devices to circuits to architectures to algorithms are indispensable. This talk will examine energy-efficient neural network accelerators for embedded applications in this context. Several deep neural network accelerator designs based on cross-layer efforts spanning from alternative device technologies, circuit styles and architectures will be highlighted. A comprehensive application-level benchmarking study for the MNIST dataset will be presented. The discussions will demonstrate that cross-layer efforts indeed can lead to orders of magnitude gain towards achieving extreme scale energy-efficient processing.

CCS Concepts/ACM Classifiers

- Emerging technologies, architectures, neural networks

Author Keywords

Beyond-CMOS devices; computing with emerging technologies; neural network accelerators; cellular neural networks

BIOGRAPHY

X. Sharon Hu is a professor in the department of Computer Science and Engineering at the University of Notre Dame, USA. Her research interests include low-power system design, circuit

and architecture design with emerging technologies, hardware/software co-design and real-time embedded systems. She has published more than 300 papers in these areas. Some of her recognitions include the Best Paper Award from the Design Automation Conference and from the IEEE Symposium on Nanoscale Architectures, and the NSF CAREER award. She has participated in several large government and industry sponsored center-level projects and is a theme leader in an NSF/SRC E2CDA project. She is the General Chair of Design Automation Conference (DAC) in 2018, was the TPC chair of DAC in 2015, and was a member of TPC and Executive Committees of a number of other conferences. She also served as Associate Editor for IEEE Transactions on VLSI, ACM Transactions on Design Automation of Electronic Systems, etc. and is an Associate Editor of ACM Transactions on Cyber-Physical Systems. X. Sharon Hu is a Fellow of the IEEE.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

GLSVLSI'18, May 23-25, 2018, Chicago, IL, USA.

© 2018 Copyright is held by the owner/author(s).

ACM ISBN 978-1-4503-5724-1/18/05.

DOI: <https://doi.org/10.1145/3194554.3200204>