(SD2019-275) Mixed-Signal Acceleration Of Deep Neural Networks

Tech ID: 32038 / UC Case 2019-275-0

BACKGROUND

Deep Neural Networks (DNNs) are revolutionizing a wide range of services and applications such as language translation , transportation , intelligent search, e-commerce, and medical diagnosis. These benefits are predicated upon delivery on performance and energy efficiency from hardware platforms. With the diminishing benefits from general-purpose processors, there is an explosion of digital accelerators for DNNs. Mixed-signal acceleration is also gaining traction. Albeit low-power, mixed signal circuitry suffers from limited range of information encoding, is susceptible to noise, imposes Analog to Digital (A/D) and Digital to Analog (D/A) conversion overheads, and lacks fine-grained control mechanism. Realizing the full potential of mixed-signal technology requires a balanced design that brings mathematics, architecture, and circuits together.

TECHNOLOGY DESCRIPTION

Researchers from UC San Diego devised a patent-pending clustered 3D-stacked microarchitecture, dubbed BIHIWE, that provides the capability to integrate copious number of low-bitwidth switched-capacitor Multiply-Accumulate operations (MACC) units that enables the interleaved bit-partitioned arithmetic.

This technology is patent-pending.

CONTACT

Skip Cynar
scynar@ucsd.edu
tel: 858-822-2672.

INVENTORS

- Esmaeilzadeh, Hadi
- Ghodrati, Soroush

OTHER INFORMATION

KEYWORDS

Computer Science, Hardware Architecture, Deep Neural Networks, microarchitecture, interleaved bit-partitioned arithmetic, Acceleration, mixed-signal

CATEGORIZED AS

- **Computer**
  - Hardware
  - Software
- **Engineering**
  - Engineering

RELATED CASES

2019-275-0

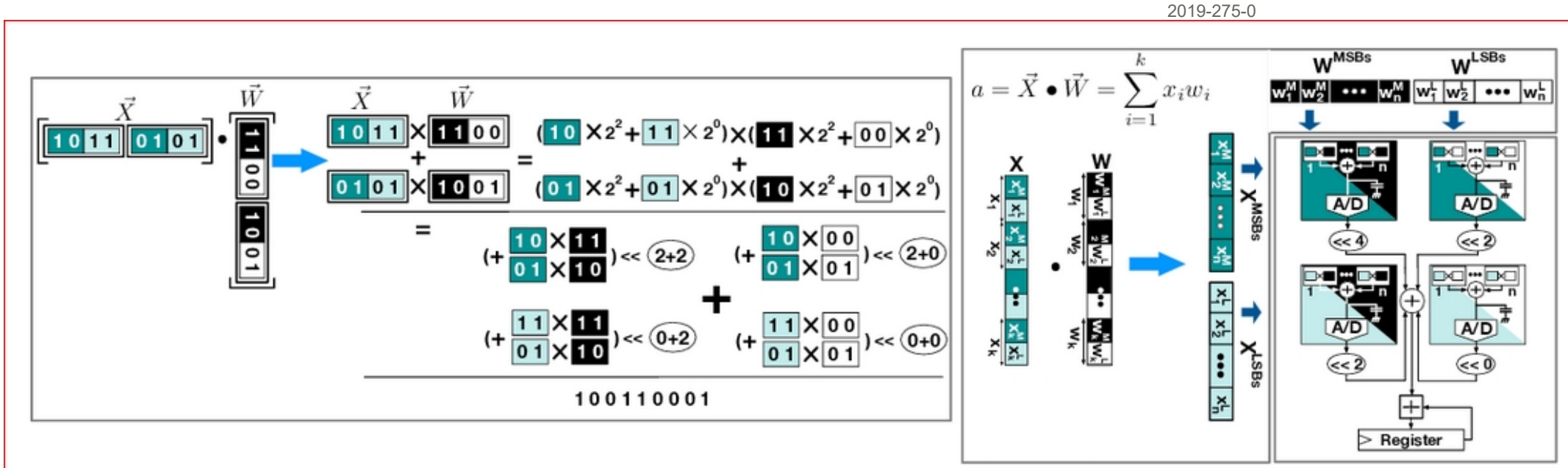

Figure 1: Wide, interleaved, and bit-partitioned mathematical formulation.

APPLICATIONS

This technology demonstrates an interleaved, and bit-partitioned arithmetic to overcome two key challenges in mixed-signal acceleration of DNNs: limited encoding range, and costly A/D conversions. This bit-partitioned arithmetic enables rearranging the highly parallel MACC operations in modern DNNs into wide low-bitwidth computations that map efficiently to low-bitwidth mixed-signal units. Further, these units operate in charge domain using switched-capacitor circuitry and reduce the rate of A/D conversion by accumulating partial results in the analog domain. The resulting microarchitecture, named BIHIWE, offers 4.5 higher performance compared to a fully-digital state-of-the-art architecture within the same power budget. These encouraging results of this invention--which combines mathematical insights with architectural innovations--can enable new avenues in DNN acceleration.

ADVANTAGES

BIHIWE comes with a mixed-signal building block that performs wide bitpartitioned vector dot-product. BIHIWE then organizes these building blocks in a clustered hierarchical design to efficiently make use of its copious number of parallel low-bitwidth mixed-signal MACC units. The clustered design is crucial as mixed-signal paradigm enables integrating a larger number of parallel operators than the digital counterpart.

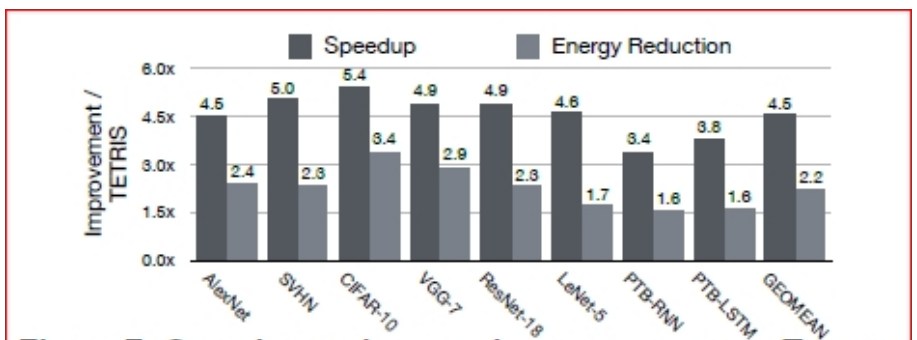

Figure 7: Speedup and energy improvement over TETRIS.

STATE OF DEVELOPMENT

Evaluating the carefully balanced design of BIHIWE with eight DNN benchmarks shows that BIHIWE delivers 4.5X over the leading purely digital 3D-stackedDNNaccelerator, TETRIS, with virtually no loss (< 1%) in classification accuracy. BIHIWE offers 31X higher Performance-per-Watt compared to Titan Xp GPU with 8-bit execution while running 1.7 faster. With these benefits, this technology paves the way for a new shift in DNNs acceleration.

INTELLECTUAL PROPERTY INFO

Patent-pending technology. UC San Diego is seeking licensees for commercial development.

RELATED MATERIALS

- Soroush Ghodrati, Hardik Sharma, Sean Kinzer, Amir Yazdanbakhsh, Kambiz Samadi, Nam Sung Kim, Doug Burger, Hadi Esmaeilzadeh Mixed-Signal Charge-Domain Acceleration of Deep Neural Networks through Interleaved Bit-Partitioned Arithmetic. Submitted to Cornell pre-print archive on 27 Jun 2019 - 06/27/2019

PATENT STATUS

Patent Pending

OTHER INFORMATION

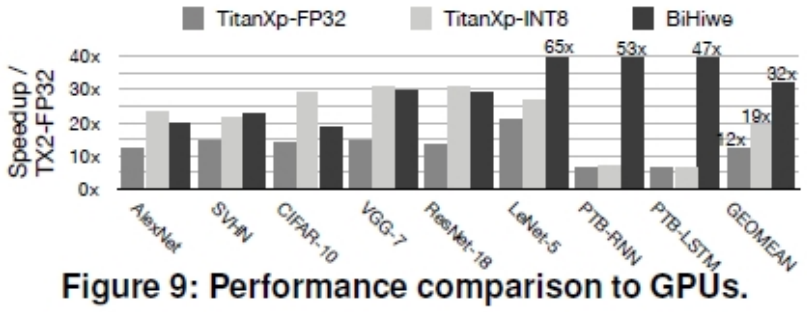

Figure 9: Performance comparison to GPUs.

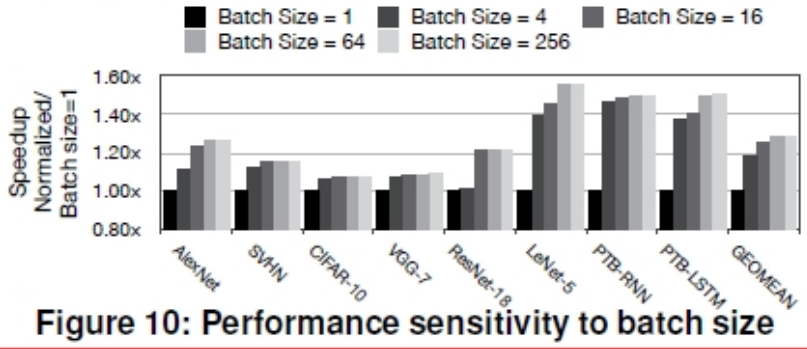

Figure 10: Performance sensitivity to batch size

University of California, San Diego
Office of Innovation and Commercialization
9500 Gilman Drive, MC 0910, ,
La Jolla,CA 92093-0910

Tel: 858.534.5815
innovation@ucsd.edu
https://innovation.ucsd.edu
Fax: 858.534.7345