MRAM-based Deep In-memory Architectures

Technology, Design, Modeling, and System Analysis

Kuk-Hwan Kim
Department of Electrical & Computer Engineering
University of Illinois at Urbana-Champaign, Urbana, IL 61801
Email: khk@illinois.edu

Naveen Verma
Department of Electrical Engineering
Princeton University
Princeton, NJ 08544
Email: nverma@princeton.edu

Michael Burkland
Raytheon Missiles & Defense
Tucson, AZ 85706
Email: mburkland@raytheon.com

Steven Soss
GLOBALFOUNDRIES Inc.
Malta , NY 12020
Email: steven.soss@globalfoundries.com

Naresh Shanbhag
Department of Electrical & Computer Engineering
University of Illinois at Urbana-Champaign, Urbana, IL 61801
Email: shanbhag@illinois.edu

***Abstract*— This paper describes the application of deep in-memory architectures (DIMAs) designed using MRAM devices in radio-frequency (RF) signal processing chains commonly used in range estimation systems of interest to DoD. MRAM-based DIMAs have the potential to significantly reduce system-level energy-latency costs while enhancing compute density of the platform. Thermal limits arising from the use of conventional von Neumann architectures are overcome. However, the design of MRAM-based DIMA systems requires a holistic design approach encompassing systems, architectures, circuits and devices. This paper describes a systems-to-devices methodology developed for this purpose involving organizations from the semiconductor and defense sectors and universities.**

Keywords—In-memory computing; analog computing; MRAM; deep learning; signal processing

I. INTRODUCTION

Requirements of many applications critically important to the US Department of Defense (DoD) and the commercial sector are difficult to meet using current day computational platforms due to fundamental barriers imposed by thermal (power) and energy constraints restrict mission. The von Neumann architecture and its associated Turing model of computation are misaligned with the data-centric, information-processing nature of emerging workloads. Specifically, the memory bottleneck in such architectures fundamentally limits energy efficiency and latency reduction. In addition, advances in nanoscale process technologies remain untapped by systems since such nanoscale devices require new compute models that can comprehend the unique stochastic nature of such devices.

To address the abovementioned challenges, this paper describes the use of GLOBALFOUNDRIES (GF) 22nm FDX MRAM (magnetic RAM) device [1] within deep in-memory architectures (DIMAs) [3,4] supported by Shannon-inspired statistical compute models [2] in order to optimize the signal-to-noise ratio (SNR) vs. energy trade-off inherent in systems of interest to DoD. Systems based on conventional architectures are increasingly limited by the cost of Size, Weight and Power (SW&P). In contrast, DIMA-based non-von Neumann architectures provides a path forward based on a re-alignment across applications, architectures, and technology in order to provide transformative capabilities for missions requiring embedded processing within strict thermal budgets.

II. RF SIGNAL PROCESSING CHAIN (SPC)

Fig. 1 illustrates the front end of a conventional RF signal processing chain used for generation of Range-Doppler Maps (RDMs). Demanding workloads requiring Matrix-Vector Multiplication (MVM) operations provide an opportune case study to evaluate the performance of DIMA against commercially available FPGAs used in practice. The functional blocks employed for processing by DIMA are indicated in red.

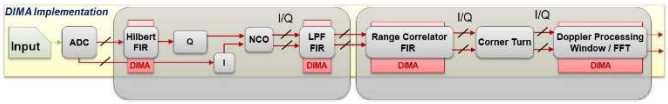

Fig. 1. A signal processing chain for generating Range-Doppler Maps.

Generation of the quadrature component of the analog signal is implemented by a Hilbert Transform of digitized RF sensor data followed by Range Correlation to produce a RDM output. The workloads required for this system place heavy processing demands on MVM operations used in FIR filtering, and FFT/IFFT transforms. Consequently, inefficiencies in von Neumann computing result in substantial amounts of waste heat, of which the mitigation solutions are expensive and detrimental to mission performance. Thus, a radical alternative to von Neumann architectures is required to meet the signal processing performance needs of near term and future DoD missions. Analysis of the energy/latency (power), as well as accuracy, between DIMA and FPGA performance for various workloads shows promising results for non von Neumann alternative to RF signal processing for these demanding DoD applications.

III. Shannon-inspired MRAM-based Deep In-memory Architectures

We employ GF's spin-torque-transfer (STT) MRAM at the 22 nm SOI node [1] as the device fabric for DIMA. MRAM is an emerging memory in the foundry ecosystem, with full-array capability demonstrated by GF including high reliability (endurance), high speed (22 ns/200 ns read/write), and low energy (0.2 pJ/17 pJ read/write), in a 1 Mb/40 Mb subarray/array while also offering non-volatility. This makes MRAM an attractive device technology for designing scaled-up DIMAs in a commercial setting while potentially offering rad-hard capabilities for DoD systems.

Since DIMA trades off its compute SNR for energy-delay benefits, methods to compensate for this SNR loss need to be developed. We employ a Shannon-inspired model of computation [2] to manage this energy-delay vs. SNR trade-off intrinsic to DIMA whereby computation is framed as a process of information transfer over a noisy circuit/device fabric. The Shannon-model of computation develops statistical error compensation (SEC) techniques to preserve the information content in data [2] and enables MRAM-based DIMA to operate in the low-SNR regime without compromising on system-level accuracy.

IV. Systems-to-Devices Exploration Methodology

The full-stack (systems-to-devices) attribute of MRAM-based DIMA requires a closed-loop systems-driven design exploration and optimization methodology that comprehends the energy-delay-robustness trade-offs at each layer of the compute stack. The disciplinary challenge of such an exploration includes managing and establishing consistency between various design metrics and constraints throughout the compute stack. To address this challenge, we developed the systems-to-devices exploration methodology shown in Fig. 2.

This *meet-in-the-middle* methodology begins with defining the functional requirements of the system (*top-down*) under consideration, e.g., the RF signal processing chain for Range-Doppler estimation; evaluating various computational kernel choices such as Hilbert transformers, numerically controlled oscillators (NCOs), range-correlators, and Discrete Fourier Transform (DFT); translating system requirements into block-level requirements in power, latency, and SNR based on the chosen kernels. Behavioral models (BeMos) of MRAM-based DIMA incorporating device variations using GF's process characterizations, circuit read-out noise, ADC non-idealities, and Shannon-inspired SEC methods to compensate for those implementation non-idealities, were developed (*bottom-up*) to explore power, latency, and SNR trade-offs of those blocks.

V. Results

Based on the exploration methodology shown in the Fig. 2, we have successfully engineered the first DIMA prototype onto GF's 22-FDX process with the embedded MRAM technology. The tight collaboration with GF necessitated by DIMA-specific requirements to the MRAM macro design and MTJ parameters lead to the first functional MRAM-DIMA IC.

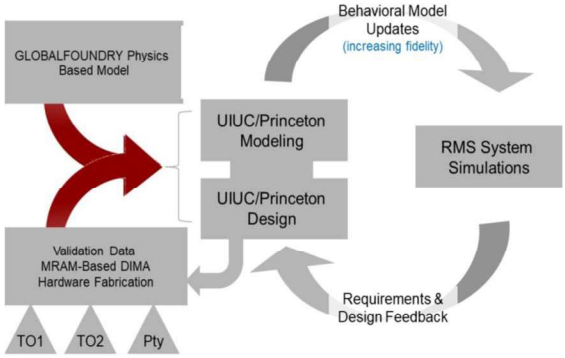

Fig. 2. Systems-to-devices exploration methodology.

The DIMA prototype is integrated in a flexible system testbed, enabling detailed characterization of the computation statistics of the prototype. Specifically: (1) in-system demonstration of a CIFAR-10 image-classification neural network was achieved, employing an SEC stochastic-training algorithm, incorporating DIMA statistics [5], to achieve accuracy at the level algorithms/systems of interest required; and (2) various mappings of a baseline radar signal-processing chain are being explored, matching block-level and system-level performance metrics based on the calibrated BeMo models.

VI. Future Directions

Future research will focus on scaling up MRAM-based DIMA, automating its synthesis via platform tools, and experimental demonstrations in various DoD and commercial applications.

Acknowledgment

This material is based on research sponsored by Air Force Research Laboratory (AFRL) and Defense Advanced Research Projects Agency (DARPA) under agreement number FA8650-18-2-7866.

References

[1] Lee, K., Chao, R., Yamane, K., Naik, V.B., Yang, H., Kwon, J., et al, "22-nm FD-SOI Embedded MRAM Technology for Low-Power Automotive-Grade-1 MCU Applications," Proceedings of the 2018 IEEE International Electron Devices Meeting (IEDM), San Francisco, California, USA (2018).

[2] Shanbhag, N.R., Verma, N., Kim. Y., Patil A.D., Varshney L.R., "Shannon-Inspired Statistical Computing for the Nanoscale Era," Proceeding of the IEEE, 107, 1, Jan 2019, pp. 90-107.

[3] Kang, M., Gonugondla, S.K., and Shanbhag, N.R., "A 19.4 nJ/Decision 364K Decisions/s Inmemory Random Forest Classifier in 6T SRAM Array," IEEE Journal of Solid-State Circuits, 53, 7, May 2018, pp. 2126-2135.

[4] Zhang, J., Wang, Z., and Verma, N., "In-memory Computation of a Machine-learning Classifier in a Sttandard 6T SRAM Array," IEEE Journal of Solid-State Circuits, 52, 4, Apr 2017, pp. 915–924.

[5] Zhang, B., Chen, L.Y., and Verma, N., "Stochastic Data-driven Hardware Resilience to Efficiently Train Inference Models for Stochastic Hardware Implementations," IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Brighton, U.K. (2019).