2024 IEEE 32nd Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM) | 979-8-3503-7243-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/FCCM60383.2024.00045

The BRAM is the Limit: Shattering Myths, Shaping Standards, and Building Scalable PIM Accelerators

MD Arafat Kabir*, Tendayi Kamucheka*, Nathaniel Fredricks*,
Joel Mandebi†, Jason Bakos‡, Miaoqing Huang*, and David Andrews*
*Department of Electrical Engineering and Computer Science, University of Arkansas,
‡Department of Computer Science and Engineering, University of South Carolina,
†Advanced Micro Devices, Inc. (AMD)
{makabir, tfkamuch, njfredri, mqhuang, dandrews}@uark.edu, jmandebi@amd.com, jbakos@cse.sc.edu,

***Abstract*—Many recent FPGA-based Processor-in-Memory (PIM) architectures have appeared with promises of impressive levels of parallelism but with performance that falls short of expectations due to reduced maximum clock frequencies, an inability to scale processing elements up to the maximum BRAM capacity, and minimal hardware support for large reduction operations. In this paper, we propose a "Standard" set of design objectives for PIM array-based FPGA designs.**

We then propose a PIM array-based GEMV accelerator architecture as a case study to show the proposed Standard can be realized in practice. The GEMV accelerator serves as existence proof that dispels several myths surrounding what is normally accepted as clocking and scaling FPGA performance limitations. Specifically, the proposed accelerator clocks at the maximum frequency of the BRAM and scales to 100% of the available BRAMs. Comparative analyses show execution speeds over existing PIM-based GEMV engines on FPGAs and achieving a 2.65× – 3.2× faster clock. An AMD Alveo U55 implementation achieves a system clock speed of 737 MHz, providing 64K bit-serial multiply-accumulate (MAC) units for GEMV operation.

***Index Terms*—Processing-in-Memory, Gold Standard, System Design, Block RAM, GEMV engine, Processor Array.**

I. Proposed Standard

A comparative study of existing PIM array-based accelerators [1]–[4] reveals that their system frequencies are significantly slower than the BRAM maximum frequencies of the implementation platforms. Most of these systems could not utilize all available BRAMs as PIMs. The reduced utilization, along with a decrease in clock frequency, leads to very poor use of the available internal BRAM bandwidth and diminishes the overall system compute density. Motivated by these observations, we explored two key questions: what is the highest frequency achievable by a PIM-based design on FPGAs, and whether it is possible to maximize compute density to the full BRAM capacity without compromising clock frequency. Based on our studies we propose the following standard set of design objectives for PIM array-based FPGA designs,

- The system frequency of a PIM array-based design needs to scale with the BRAM frequency of the platform, ideally running at the BRAM maximum frequency.
- The peak-performance of the PIM array needs to scale linearly with the on-chip BRAM capacity.
- The PIM array reduction network needs to be designed to balance the cycle latency and logic utilization, without affecting the system frequency.

II. Case Study: A GEMV Accelerator

A PIM array-based GEMV accelerator overlay was designed to evaluate the practicality of the proposed standard. An FSM-based controller combined with an existing PIM overlay was used to implement the GEMV tiles. The GEMV tiles were separately analyzed, which met the proposed standard goals. A scalability study reveals that the proposed PIM array can scale up to 100% BRAM capacity of almost any device of AMD's Virtex-7 and UltraScale+ FPGA families. Several iterations of implementation were required to achieve the BRAM maximum frequency as the system frequency. Device-specific optimizations were needed to avoid placement and routing issues. An implementation on Alveo U55 achieved the BRAM Fmax of 737 MHz, scaling up to 100% BRAM capacity, providing 64K MAC units. A comparative study with state-of-the-art PIM array-based GEMV accelerators shows that the proposed design runs 2.65× – 3.2× faster, outperforming all other PIM accelerators in execution time.

This material is based upon work supported by the National Science Foundation under Grant No. 1955820.

References

[1] X. Wang, V. Goyal, J. Yu, V. Bertacco, A. Boutros, E. Nurvitadhi, C. Augustine, R. R. Iyer, and R. Das, "Compute-Capable Block RAMs for Efficient Deep Learning Acceleration on FPGAs," *2021 IEEE 29th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, pp. 88–96, 2021.

[2] A. Panahi, S. Balsalama, A.-T. Ishimwe, J. M. Mbongue, and D. Andrews, "A Customizable Domain-Specific Memory-Centric FPGA Overlay for Machine Learning Applications," in *2021 31st International Conference on Field-Programmable Logic and Applications (FPL)*, Aug. 2021, pp. 24–27.

[3] A. Arora, A. Bhamburkar, A. Borda, T. Anand, R. Sehgal, B. Hanindhito, P.-E. Gaillardon, J. Kulkarni, and L. K. John, "CoMeFa: Deploying Compute-in-Memory on FPGAs for Deep Learning Acceleration," *ACM Transactions on Reconfigurable Technology and Systems*, vol. 16, no. 3, pp. 1–34, Sep. 2023.

[4] Y. Chen and M. S. Abdelfattah, "BRAMAC: Compute-in-BRAM Architectures for Multiply-Accumulate on FPGAs," in *2023 IEEE 31st Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*. Marina Del Rey, CA, USA: IEEE, May 2023, pp. 52–62.

2576-2621/24/$31.00 ©2024 IEEE
DOI 10.1109/FCCM60383.2024.00045