Making BRAMs Compute: Creating Scalable Computational Memory Fabric Overlays

MD Arafat Kabir*, Joshua Hollis*, Atiyehsadat Panahi†, Jason Bakos‡, Miaoqing Huang* and David Andrews*
Department of Computer Science and Computer Engineering
*University of Arkansas, ‡University of South Carolina, †Cadence Design Systems
{makabir, jrhollis, apanahi, mqhuang, dandrews}@uark.edu, jbakos@cse.sc.edu

***Abstract*—The increasing density of distributed BRAMs diffused throughout modern Field Programmable Gate Arrays (FPGAs) is ideal for forming processor in/near memory architectures. This breaks the traditional von Neumann memory bottleneck limiting concurrency and degrading energy efficiency. Ideally, processing density should scale linearly with BRAM capacity, and clock frequencies should be set by the read/write access times of the BRAM. In this paper, we present a PIM overlay that achieves these goals. We observe an improvement of performance by 2.25×, logic resource utilization by 2×, and accumulation delay by 17× compared to prior published work.**

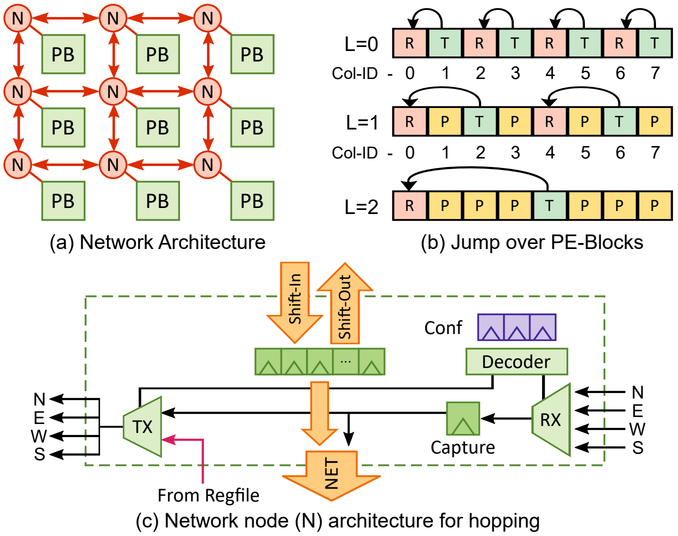
***Index Terms*—Bit-serial, Overlay, FPGA, Machine Learning, SIMD, Processor Array, Processing-in-Memory**

I. Proposed PIM Overlay Architecture

Generic bit-serial PIM architectures do not allow for fast reduction operations, such as summing product terms, between processing elements (PEs) [1]–[3]. To address this shortcoming, we introduce an operand-multiplexer (Op-Mux) module between the BRAM and the ALU that enables zero-copy reduction operations. The Network Node in Fig. 1 enables the streaming of partial products into the ALU of a destination PE for summation, without intermediate copying. As shown in Fig. 1(b), a binary hopping mechanism is implemented to accelerate the reduction/accumulation operation. A 3-bit level encoding is used along rows and columns that can handle an array containing up to 1 million PEs. Reduction operation is optimized by inserting pipeline stages that overlap and hide data transfers with ALU operations. All components of the architecture are optimized for FPGA logic fabric, and consumes minimal resources providing high scalability.

II. Analysis and Conclusion

The studies presented in [1] show significant speed-up compared to existing custom designs (1.75×) and HLS (24.51×). We adopt the architecture presented in [1] as a benchmark design for our work and comparative study. Fig. 2 summarizes our primary results normalized with respect to the benchmark design, evaluated on a Virtex-7 (xc7vx485t-2) device. Four different configurations of the proposed designs were evaluated: Full-Pipe, Single-Cycle, RF-Pipe, and Op-Pipe. As observed in Fig. 2, the **Full-Pipe design runs at the maximum BRAM frequency, 540 MHz, 2.25× faster** than the benchmark. The proposed design scales linearly with the BRAM capacity of the device, **reaching the maximum PE capacity of the device (32K).** FPGA-specific optimizations and careful design of the register control sets minimizes the placement and routing issues, improving the logic utilization. All of the configurations of the proposed design achieves ≥**2× better slice utilization** compared to the benchmark.



(a) Network Architecture

(b) Jump over PE-Blocks

(c) Network node (N) architecture for hopping

Fig. 1. Data network for fast accumulation and reduction operations

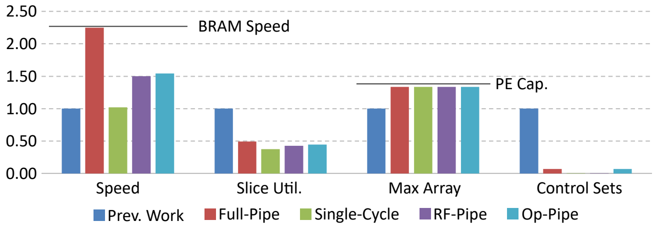


Fig. 2. Relative improvements in proposed designs w.r.t the benchmark [1]

References

[1] 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)*, 2021, pp. 24–27.

[2] A. Arora, T. Anand, A. Borda, R. Sehgal, B. Hanindhito, J. Kulkarni, and L. K. John, "CoMeFa: Compute-in-Memory Blocks for FPGAs," in *2022 IEEE 30th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, May 2022, pp. 1–9.

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