Bleeding Edge Hardware Acceleration: A literature review on the State-of-the-Art

SYSC4310A – Fall 2020

For Professor Paulo Garcia

Written By: Brannon Chan (#101045946)

Hardware acceleration is the use of specially designed circuits and architectures used to improve the performance of some given software. Within the field of Computer Engineering, continual advances in both traditional, Von Neumann and non-conventional Post-Moore computer architectures, alongside new integrated circuit technologies have provided an optimized, hardware (H/W) accelerated means in reducing the performance gap between current computer standards, and the performance required to satisfy newer software systems & applications. However, to overcome the computational limitations of the "Von-Neumann bottleneck" [1, 2, 3, 4, 5], application specific H/W accelerated solutions must be considered, over the archaic *one-size-fits-all* Von-Neumann general computation approach [6]. Cutting edge reconfigurable architectures such as the Coarse-Grained Reconfigurable Architecture and new integrated circuit technologies like Resistive Random-Access Memory can give way to implement such systems.

A popular hardware-based approach for surpassing the Von-Neumann bottleneck is to implement *Processing in Memory* (PIM) architecture [3, 5, 6, 7, 8], which brings the memory and processor closer on the silicon fabric; allowing for massive parallelizable potential [5], largely increasing throughput by forgoing the need for expensive memory access operations and minimizing the power expenditure normally required for a fetch-decode-execute cycle in a traditional architecture [8].

The *Resistive Random-Access Memory* (RRAM) is an emerging memristor-based device that operates by storing data as a resistance [3, 4, 5] and is very similar in operation to other existing non-volatile CMOS memory technologies such as DRAM that exhibit the potential for parallelism. However, unlike DRAM, RRAM can make much better use of the parallelism provided by PIM, since it is not constrained to long and power-intensive chip to chip data transfers [5]; meaning that RRAM could be a relatively high density, high throughput, low power, scalable storage technology. Due to the analog computational nature of such PIM devices, memristor accelerators could be favourable for Neural Network acceleration & Machine Learning applications [8], PIM enabled CPUs (mMPU) [5], Gate-level RRAM memory arrays conducive for parallel processing [4] or constructed in a 'RRAM crossbar array' to act as a PIM enabled storage device [3]. It is worth noting that PIM capability can also be achieved without memristive devices, as seen in Castañeda et al. [6], a fully digital PIM processor layout was implemented on a 28nm standard cell-based CMOS.

An interesting use of a resistive memory crossbar at the reconfigurable computer architecture level is the *Field-Programmable Crossbar Array* (FPCA), proposed by Zidan et al. [13], which divides sections of a resistive crossbar into 'memory cores' that could be dynamically reconfigured to act as storage, or a digital/analog processing unit; all of which lies on the same crossbar fabric. By extension, the FCPA and other similar crossbar-based architectures [1, 2] could effectively hardware accelerate a Binary Neural Network [13], logical bitwise operations, facilitate in-memory arithmetic operations [1, 13] and compute SIMD instructions in a massively parallel manner [1].

Aside from circuit level solutions for exceeding the Von-Neumann bottleneck, relatively novel reconfigurable computing Post-Moore architectures such as the *Course-Grained Reconfigurable Architecture* (CGRA) [9, 10] and the *Field-Programmable Gate Array* (FPGA) have recently regained interest due to the architecture's ability to retain most of the performance offered by costly ASIC implementations while providing some lesser degree of the reconfigurability usually seen in general purpose processor [10].

Of these architectures, the FPGA is the current industry standard for reconfigurable performance hardware, seeing use in various commercial products, industrial applications, and possesses extensive software support/toolchains in addition to a mature research background [9]. Some current state of the art research utilizes FPGAs to implement resource efficient ASIC designs, to act as a H/W accelerator for both High Performance Computing (HPC) [16] and other low power, high performance applications [11]. For example, Fan et al. uses an FPGA for use in an optimized H/W & algorithm co-design of a 3D CNN for Human Action Recognition, being 13 times faster than previous FPGA implementations [11]. Other FPGA accelerated applications include GPU-FPGA coupling [16], algorithm & architecture co-optimization for GCN inference [17], dynamic Network flow-specific microburst detection [18], workload reduction on low power IoT devices [19], stall-free H/W design for LTSM networks [20] and providing H/W acceleration for the ABSW algorithm [21]. Thanks to robust toolchains, extensive knowledge and software support that exists for FPGAs, the corresponding development costs are very attractive for both academic researchers and tech corporations alike.

Regarding newer reconfigurable architectures, the CGRA can functionally be considered the successor of the FPGA, since the CGRA is essentially a further generalized and less reconfigurable version of an FPGA, allowing for larger, specialized functional blocks to be created running at a faster clock frequency, and requires significantly less time for reconfiguration of synthesized or compiled designs [10]. Current state of the art research seeks to expand the range of purpose-built designs, for instance: a tiny variable floating-point CGRA for use in near-edge IoT devices [12], a traditional CGRA with DMA for critical section acceleration [14], and a CGRA tuned for predicting real-time processes [15] to name a few. In comparison to FPGAs, CGRAs also see use in HPC, embedded, and other applications needing more 'compute per unit area' [10], however existing documentation and research on CGRAs is sparce.

To conclude, post-Moore reconfigurable architectures like CGRA and novel integrated circuit technologies such as RRAM will aid in surpassing the Von-Neumann bottleneck. Based on the research above, we show that memristive devices can provide PIM capability and perform back-compatible conventional memory operations at a fraction of the time and power required from standard CMOS technology, making resistive memory like RRAM very attractive for parallelizable, high throughput, low latency & power applications. In addition, reconfigurable architectures such as FPGAs and CGRAs provide a relatively low-cost alternative to ASICs for H/W acceleration, offering either fine gate-level or coarse-grained processing element (PE) mesh reconfigurability for synthesized H/W optimized designs. Though research for resistive memory and reconfigurable architectures is growing, there certainly is a vast amount of research that must be done to bring these technologies to the forefront of H/W acceleration. Regardless, one should pay attention as research continues to develop, because of the shear potential for performance inherent to these technologies.

References:

[1] X. Wang, M. A. Zidan and W. D. Lu, "A Crossbar-Based In-Memory Computing Architecture," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 67, no. 12, pp. 4224-4232, Dec. 2020, doi: 10.1109/TCSI.2020.3000468.

[2] B. Chen, F. Cai, J. Zhou, W. Ma, P. Sheridan and W. D. Lu, "Efficient in-memory computing architecture based on crossbar arrays," 2015 IEEE International Electron Devices Meeting (IEDM), Washington, DC, 2015, pp. 17.5.1-17.5.4, doi: 10.1109/IEDM.2015.7409720.

[3] F. Wang, G. Luo, G. Sun, J. Zhang, P. Huang and J. Kang, "Parallel Stateful Logic in RRAM: Theoretical Analysis and Arithmetic Design," 2019 IEEE 30th International Conference on Application-specific Systems, Architectures and Processors (ASAP), New York, NY, USA, 2019, pp. 157-164, doi: 10.1109/ASAP.2019.000-8.

[4] J. Reuben and S. Pechmann, "A Parallel-friendly Majority Gate to Accelerate In-memory Computation," 2020 IEEE 31st International Conference on Application-specific Systems, Architectures and Processors (ASAP), Manchester, United Kingdom, 2020, pp. 93-100, doi: 10.1109/ASAP49362.2020.00025.

[5] S. Kvatinsky, "Real Processing-in-Memory with Memristive Memory Processing Unit (mMPU)," 2019 IEEE 30th International Conference on Application-specific Systems, Architectures and Processors (ASAP), New York, NY, USA, 2019, pp. 142-148, doi: 10.1109/ASAP.2019.00-10.

[6] O. Castañeda, M. Bobbett, A. Gallyas-Sanhueza and C. Studer, "PPAC: A Versatile In-Memory Accelerator for Matrix-Vector-Product-Like Operations," 2019 IEEE 30th International Conference on Application-specific Systems, Architectures and Processors (ASAP), New York, NY, USA, 2019, pp. 149-156, doi: 10.1109/ASAP.2019.000-9.

[7] R. Fife, I. Udoh and P. Garcia, "Coherency overhead of Processing-in-Memory in the presence of shared data," 2020 IEEE International Conference on Industrial Technology (ICIT), Buenos Aires, Argentina, 2020, pp. 237-242, doi: 10.1109/ICIT45562.2020.9067234.

[8] S. Mittal, “A Survey of ReRAM-Based Architectures for Processing-In-Memory and Neural Networks,” Machine Learning and Knowledge Extraction, vol. 1, no. 1, pp. 75–114, Apr. 2018 [Online]. Available: http://dx.doi.org/10.3390/make1010005

[9] A. Podobas, K. Sano and S. Matsuoka, "A Template-based Framework for Exploring Coarse-Grained Reconfigurable Architectures," 2020 IEEE 31st International Conference on Application-specific Systems, Architectures and Processors (ASAP), Manchester, United Kingdom, 2020, pp. 1-8, doi: 10.1109/ASAP49362.2020.00010.

[10] A. Podobas, K. Sano and S. Matsuoka, "A Survey on Coarse-Grained Reconfigurable Architectures From a Performance Perspective," in IEEE Access, vol. 8, pp. 146719-146743, 2020, doi: 10.1109/ACCESS.2020.3012084.

[11] H. Fan et al., "F-E3D: FPGA-based Acceleration of an Efficient 3D Convolutional Neural Network for Human Action Recognition," 2019 IEEE 30th International Conference on Application-specific Systems, Architectures and Processors (ASAP), New York, NY, USA, 2019, pp. 1-8, doi: 10.1109/ASAP.2019.00-44.

[12] R. Prasad, S. Das, K. Martin, G. Tagliavini, P. Coussy, L. Benini, et al., "TRANSPIRE: An energy-efficient TRANSprecision floating-point Programmable archItectuRE", Proc. Des. Automat. Test Eur. Conf., pp. 1-7, 2020

[13] M. A. Zidan, Y. Jeong, J. H. Shin, C. Du, Z. Zhang and W. D. Lu, "Field-Programmable Crossbar Array (FPCA) for Reconfigurable Computing," in IEEE Transactions on Multi-Scale Computing Systems, vol. 4, no. 4, pp. 698-710, 1 Oct.-Dec. 2018, doi: 10.1109/TMSCS.2017.2721160.

[14] D. L. Wolf, L. J. Jung, T. Ruschke, C. Li and C. Hochberger, "AMIDAR project: Lessons learned in 15 years of researching adaptive processors", Proc. 13th Int. Symp. Reconfigurable Commun. Syst. Chip (ReCoSoC), pp. 1-8, Jul. 2018.

[15] H. Siqueira and M. Kreutz, "A coarse-grained reconfigurable architecture for a PRET machine", Proc. VIII Brazilian Symp. Comput. Syst. Eng. (SBESC), pp. 237-242, Nov. 2018.

[16] R. Kobayashi et al., "Accelerating Radiative Transfer Simulation with GPU-FPGA Cooperative Computation," 2020 IEEE 31st International Conference on Application-specific Systems, Architectures and Processors (ASAP), Manchester, United Kingdom, 2020, pp. 9-16, doi: 10.1109/ASAP49362.2020.00011.

[17] B. Zhang, H. Zeng and V. Prasanna, "Hardware Acceleration of Large Scale GCN Inference," 2020 IEEE 31st International Conference on Application-specific Systems, Architectures and Processors (ASAP), Manchester, United Kingdom, 2020, pp. 61-68, doi: 10.1109/ASAP49362.2020.00019.

[18] S. Yoshida, Y. Ukon, S. Ohteru, H. Uzawa, N. Ikeda and K. Nitta, "FPGA-Based Network Microburst Analysis System with Flow Specification and Efficient Packet Capturing," 2020 IEEE 31st International Conference on Application-specific Systems, Architectures and Processors (ASAP), Manchester, United Kingdom, 2020, pp. 29-32, doi: 10.1109/ASAP49362.2020.00014.

[19] S. Kang, J. Moon and S. Jun, "FPGA-Accelerated Time Series Mining on Low-Power IoT Devices," 2020 IEEE 31st International Conference on Application-specific Systems, Architectures and Processors (ASAP), Manchester, United Kingdom, 2020, pp. 33-36, doi: 10.1109/ASAP49362.2020.00015.

[20] Z. Que et al., "Efficient Weight Reuse for Large LSTMs," 2019 IEEE 30th International Conference on Application-specific Systems, Architectures and Processors (ASAP), New York, NY, USA, 2019, pp. 17-24, doi: 10.1109/ASAP.2019.00-42.

[21] Y. Liao, Y. Li, N. Chen and Y. Lu, "Adaptively Banded Smith-Waterman Algorithm for Long Reads and Its Hardware Accelerator," 2018 IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP), Milan, 2018, pp. 1-9, doi: 10.1109/ASAP.2018.8445105.