Stochastic Dataflow Computing

Wojciech Romaszkan, Tianmu Li, Jiyue Yang, Rahul Garg, Albert Li, Di Wu, Kang Wang, Sudhakar Pamarti, Puneet Gupta, University of California, Los Angeles

ELECTRONICS RESURGENCE INITIATIVE SUMMIT & MTO Symposium 2020 | Seattle, WA | August 18-20

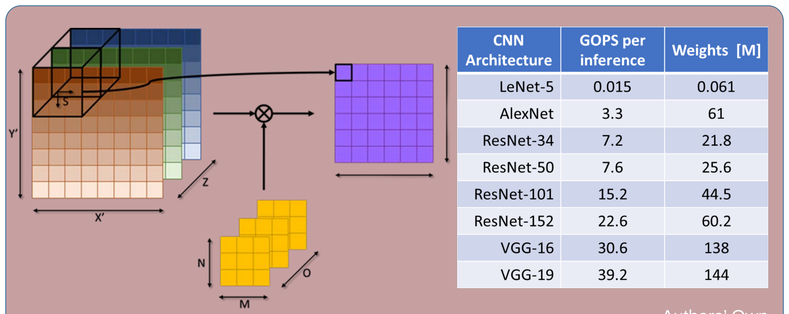
Foundation Required for Novel Compute (FRANC)

Artificial Intelligence

Background

AI in Edge Devices

- Edge AI is highly desirable due to lower latency and improved privacy and security
- Edge AI should support a variety of applications under tight, often time-varying, energy constraints
- Memory Bottleneck: Currently, large AI models incur high memory access energy and latency costs

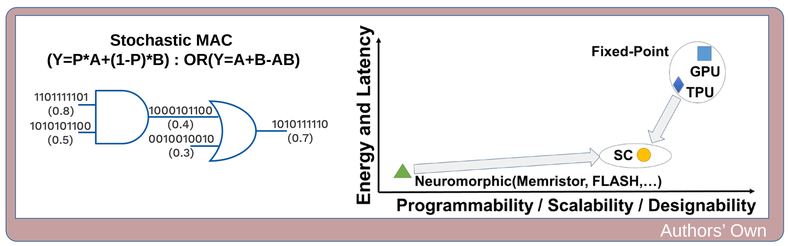


CNN Architecture	GOPS per inference	Weights [M]
LeNet-5	0.015	0.061
AlexNet	3.3	61
ResNet-34	7.2	21.8
ResNet-50	7.6	25.6
ResNet-101	15.2	44.5
ResNet-152	22.6	60.2
VGG-16	30.6	138
VGG-19	39.2	144

Authors' Own

Stochastic Computing (SC)

- Numbers represented as average of random binary streams
  - Compact multiply-and-accumulate (MAC) units enable massive parallelization
  - Longer streams improve compute accuracy
- SC is perfect for edge AI
  - Faster, smaller, and more energy efficient
  - Enables runtime adaptation of energy, latency, and accuracy



Authors' Own

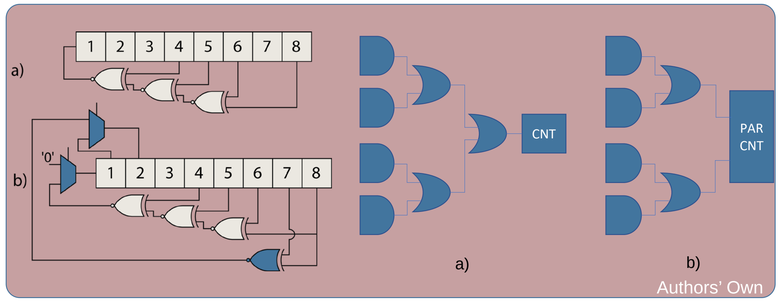
Challenges

- How to deliver high AI inference accuracy without sacrificing the parallelization benefits of SC ?
- How to maintain scalability/programmability without sacrificing energy efficiency ?
- How to scale memory bandwidth to match SC's compute parallelism? Conventional memory is not dense enough.

Approach

Improving the Accuracy of SC Inference

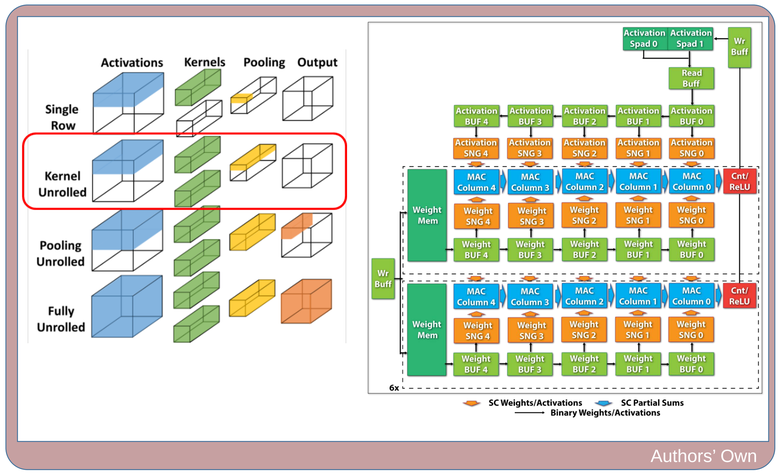
- SC-aware AI training accounts for SC induced errors completely closes the SC-fixed point accuracy gap
- Complemented by controlled stream randomness, hybrid SC-fixed point addition and other techniques



Authors' Own

Scalable, Programable SC Architecture

- Fully digital, programmable architecture (with an instruction set)
- Massive compute parallelization:
  - MAC/mm^2: ~96k (SC) vs ~0.5k (fixed-point)
- Computation skipping based stochastic pooling to save 4X-9X energy/latency with no accuracy loss
- Run-time programmable accuracy vs energy/latency



Authors' Own

Coupling SC with Magneto-Electric RAM

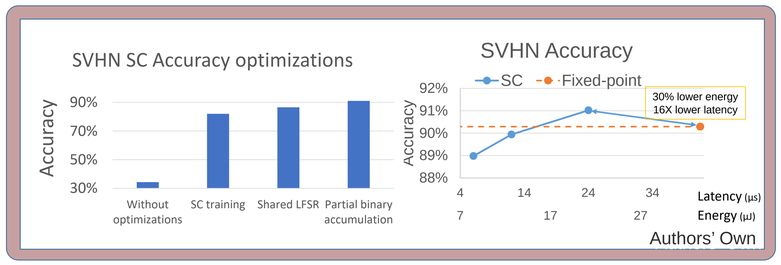
- Voltage controlled magnetic tunneling junction (VC-MTJ) based memory as cheap, high density, non-volatile, on-chip storage
  - 5x smaller and 3x less energy than other on-chip memories
- End of project goal: full integration of SC and VC-MTJ based on-chip storage

Results and Impact

Accuracy Improvements

Achieved accuracy comparable to 8-bit fixed-point

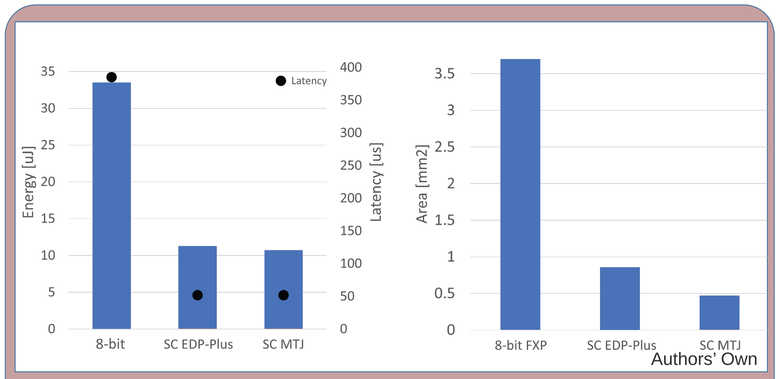
- Shown for Street View House Number (SVHN) dataset
- SC-aware training improves accuracy by up to 48%
- Arithmetic optimizations further improve accuracy by 16%



Authors' Own

Performance Improvements

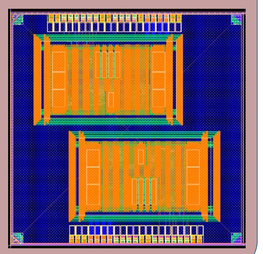
- Compared to 8b fixed-point on SVHN dataset, improvements are:
  - 2.9X (energy), 7.4X (latency), 4.3X (area), +2.2% (accuracy)
  - 120X energy-delay product (EDP) (22X iso-accuracy)
  - More expected with VC-MTJ based on-chip storage
  - Multiple performance-accuracy points supported on the same hardware
- CMOS 14nm prototypes currently in the foundry



Authors' Own

Design Variant	Accuracy [a]	GOPS	TOPS/W	Area [mm^2]
Accuracy-plus	89.0%-91.0%	150-600	15-59	0.2
EDP-plus	89.0%-91.0%	600-2400	19-76	0.4
VC-MTJ based	89.0%-91.0%	600-2400	21-81	0.2
Prior Art[b]	90.3%	188.8	8.2	1.4

[a] Based on SVHN classification, prior art assumes 8b precision.
[b] A. Biswas and A. P. Chandrakasan, "Conv-RAM: An energy-efficient SRAM with embedded convolution computation for low-power CNN-based machine learning applications," *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, San Francisco, CA, 2018, pp. 488-490, doi: 10.1109/ISSCC.2018.8310397; scaled to 14LPP, 1-bit weights. Estimated accuracy.
[c] Chen, Yu-Hsin, et al. "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks." *IEEE journal of solid-state circuits* vol 52 issue 1; scaled to 14LPP, 8-bit



Authors' Own

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA).
The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.
Distribution Statement A – Approved for Public Release, Distribution Unlimited.