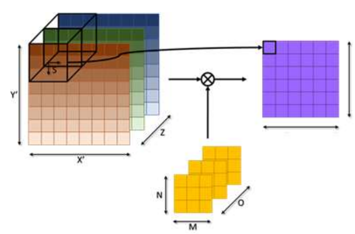
Spintronic Stochastic Dataflow Computing

Sudhakar Pamarti[1], Puneet Gupta[1], and Kang–L. Wang[1]   [1]Department of Electrical and Computer Engineering, University of California, Los Angeles

Materials & Integration: Framework for Novel Compute (FRANC)

Challenges in Modern Computing/Learning System

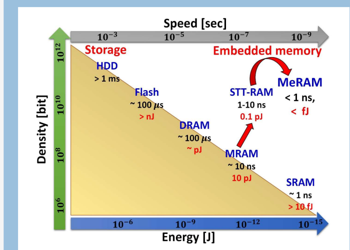
- Very large number of multiply-and-accumulate (MAC) and memory access operations needed
- Memory challenge: how can we move the memory close to the compute and reduce communication costs?
- Compute challenge: how can we parallelize to reduce latency and memory access costs?
- Generality challenge: how can the hardware be morphable to handle a variety of power and latency requirements?

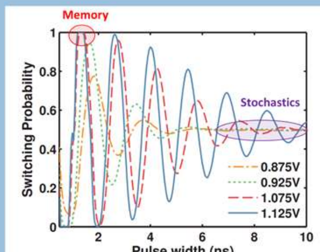


CNN Architecture	GOPS per inference	Weights [M]
LeNet-5	0.015	0.061
AlexNet	3.3	61
ResNet-152	22.6	60.2
VGG-19	39.2	144

Approach: Spintronic Technology + Stochastic Computing

Voltage-Controlled Spintronics





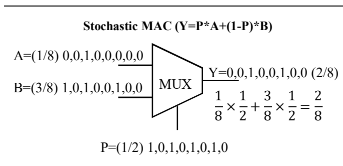
High-Performance Memory[1]

- Switching time < 1ns
- Switching voltage < 1V, current < 10uA, energy < 5fJ
- Nonvolatile, endurance > 10^{15}
- Area < $20F^2$

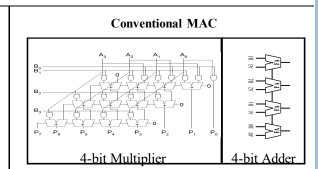
Magnetic Stochastics[2]

- True uncorrelated random signals from thermal noise
- 10x more efficient vs. state-of-the-art

Stochastic Computing



4-bit Stochastic MAC: 0.05pJ/op, ~4µm²

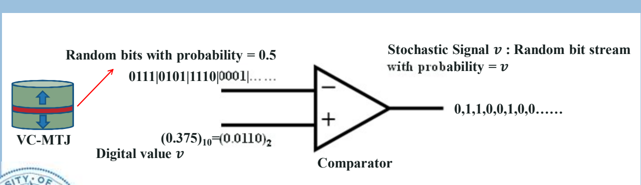


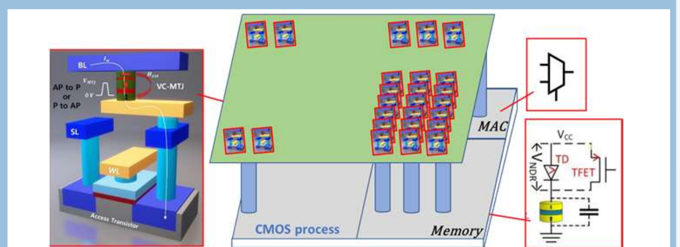
4-bit fixed-point MAC: 0.08pJ/op, ~200µm²

- Numbers represented by the frequency of occurrence of "1"s in a random, binary, bit-stream
- Ultra-compact arithmetic – 50x smaller than fixed-point
- Significantly more parallelizable compute on die
- Progressive refinement of compute accuracy

Unique Combination of Spintronics and Stochastic Computing On-Die

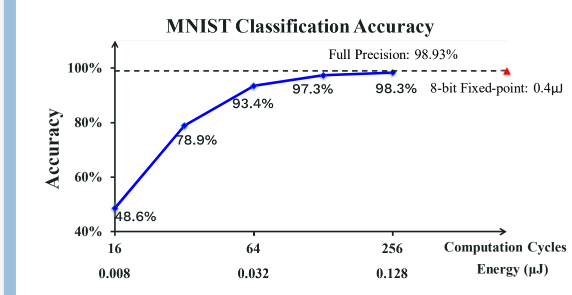
- Magnetic stochastics solves bit-stream generation problem of stochastic computing
- Logic-process compatible spintronics back-end allows fine-grain and intimate memory-compute interface





Application and Impact

- Tightly coupled computing, storage, and stochastics
- Very low area and power
- Runtime energy-latency-accuracy tradeoff on same hardware
- Uniquely beneficial to a large class of applications that have to process streaming data and would very much prefer quick results



- Significant improvement over state-of-the-art on energy and latency

Metrics	State-Of-The-Art ASIC	Anticipated results
Embedded Memory (fJ/b,ns)	90, 3	1, 1 (90x, 3x)
256DFT Energy, Latency (nJ, us)	6.2, 102	0.4, 20 (16x, 5x)
VGGNet Energy, Latency (mJ, ms)	108, 460	1.7, 2 (64x, 230x)

References

[1] Grezes, C., et.al. "Ultra-low switching energy and scaling in electric-field-controlled nanoscale magnetic tunnel junctions with high resistance-area product." Applied Physics Letters 108, no. 1 (2016): 012403.

[2] Wang, S., et.al. "Hybrid VC-MTJ/CMOS non-volatile stochastic logic for efficient computing." In Proceedings of the Conference on Design, Automation & Test in Europe, pp. 1442-1447. European Design and Automation Association, 2017.