Ultra-Low Power Probabilistic IMT Neurons for Stochastic Sampling Machines

M. Jerry[1], A. Parihar[2], B. Grisafe[1], A. Raychowdhury[2], and S. Datta[1]
[1]University of Notre Dame, Notre Dame, IN; [2]Georgia Institute of Technology, Atlanta, GA; Email: mjerry@nd.edu

Abstract: Stochastic sampling machines (SSM) utilize neural sampling from probabilistic spiking neurons to escape local minima and prevent overfitting of training datasets [1]. This enables improved error rates compared to deterministic implementations, and, in turn, enables lower bit precision, decreased chip area, and reduced energy consumption. In this work, we experimentally demonstrate: (i) Insulator-to-Metal Phase Transition (IMT) neurons with record low peak operating power of 11.9μW at V_{DD}=0.7V; (ii) the IMT in vanadium dioxide (VO_2) provides a natural probabilistic hardware substrate for realizing a compact stochastic IMT neuron for SSMs; (iii) implementation of SSM for pattern recognition on MNIST database [2] using experimentally calibrated device modeling. These results are compared to a 22nm CMOS ASIC which shows stochastic IMT neuron based SSMs result in a 4.5x reduction in system power consumption.

Introduction: Neural networks are primarily implemented on high power clusters or GPUs. However, the ubiquitous use of neural networks in data processing for character recognition, speech-to-text translation, and classification motivates the development of energy-efficient hardware tailored to their algorithmic requirements. Advances in stochastic algorithms show the energy-performance benefit of probabilistic network elements (which act to regularize the network and propel the system out of local minima (Fig. 1) [1]) in SSMs. Implementing such networks with CMOS require dedicated hardware for random number generation (RNG) and numerous multiply-accumulate (MAC) functions (Fig. 2). This in turn limits the energy and area efficiency of a traditional CMOS based SSM. In this work, we experimentally demonstrate a probabilistic hardware kernel for implementing SSMs based on stochastic IMT neurons. We harness the fundamental threshold switching variations of VO_2 to demonstrate the properties of IMT neurons map directly to the algorithmic requirements of SSMs (Fig. 2), sigmoidal spiking probability and firing rates.

Low Power IMT Neuron: Fig. 3 shows the IMT neuron structure where VO_2 is serially connected to the drain of a MOSFET in a 1T1R structure [3]. Fig. 4 shows the trends of IMT neuron peak input power and average switching voltage (V_{IMT}) with the device size. Record low peak power (11.9μW) and V_{DD} (0.7V) are achieved at L_{VO2}=100nm for an IMT neuron. Fig. 3 benchmarks this work against other published results [4]–[6] hightlighting the reduced power, operating voltage, and first demonstration of a truly stochastic neuron.

Stochastic IMT Neuron: VO_2 devices exhibit time-variant cycle-to-cycle fluctuations in the thresholding switching voltage (V_{IMT}) (Fig. 6). We verify the mechanism behind stochastic switching in VO_2 using an experimentally calibrated 2D-heterogenous network (Fig. 5). The VO_2 device is simulated as a rectangular grid of domains (45*x*84), where domains are independently capable of undergoing an IMT or MIT based on both the local potential (electrical) and temperature (thermal) [7]. Simulation results in Fig. 5(b) show the variations in V_{IMT} result from spatial and potential variations in the nucleation point of the metallic filament. V_{IMT} as a function of cycle number is shown in Fig. 6 emphasizing that the variations are not a result of V_{IMT} drift. Fig. 7 shows the model accurately captures the experimentally measured DC characteristics and V_{IMT} distribution. The IMT neuron operating principal is shown in Fig. 8(a), where the state of the IMT neuron is determined by the electrical load line. When the transistor load line crosses the stable low resistance state (solid line) the IMT neuron remains in the resting state. However, as V_{GS} increases the transistor load-line periodically (due to cycle-to-cycle V_{IMT} variations) crosses both unstable arms (dashed line) of the VO_2 characteristics which results in probabilistic spiking due to occasional oscillations in the VO_2 conductance. The DC load line analysis is confirmed by time domain measurements in Fig. 8(b) where the neuron output is measured over a time envelope for a constant V_{GS}. From this the required neuron response for SSMs is extracted in Fig. 9(a-c), where the IMT neuron exhibits the required sigmoidal instantaneous spike probability and firing rate as a function of V_{GS} and an exponential firing rate when normalizing for the refractory period. An experimentally calibrated noise model (Fig. 9(d)) reproduces the measured results, accounting for, V_{IMT} fluctuations (dominates), thermal, flicker, and shot noise.

SSM Neural Network Model: Using the noise model developed in Fig. 9 we exploit the IMT neuron level stochasticity to enable probabilistic firing of neurons in a 784×500×10 network and map unsupervised learning and inference from the MNIST handwriting dataset as in [1] (Fig. 11). IMT neurons reduce the error rate by 7.5% for 100k training sets. For large data-sets (>200K) stochasticity prevents over-fitting and improves classification accuracy by 4-5% even when the baseline accuracy is close to 90% (Fig. 12).

Benchmarking with CMOS: We perform a quantitative analysis of the power dissipated in SSM implementations using stochastic IMT neurons and 22nm CMOS ASIC with 16-bit data paths (Fig. 13). For inference tasks at matched network accuracy and memory (SRAM) power consumption (72mW) the 22nm CMOS ASIC requires 376mW while the stochastic IMT neuron accelerator sees a 4.5x reduction in operating power requiring only 82mW (Fig. 14). When excluding memory, stochastic IMT neurons reduce power dissipation by 30x (304mW to 10mW) over the 22nm CMOS ASIC.

Conclusion: Stochastic IMT neurons are demonstrated for the first time and directly mapped to the agorithmic requirements of stochastic sampling machines. The stochastic IMT neuron displays record low power and operating votlage. Using an experimentally calibrated circuit model we implement an IMT neuron based SSM, which results in a reduction of 7.5% in the error rate for unsupervised learning on the MNIST handwriting database. Further, the IMT neuron based SSM results in a 4.5x power reduction compared to a 22nm CMOS ASIC.

References: [1] S. Sheik, *ISCAS*, 2016, pp. 2090–2093. [2] Y. LeCun, *IEEE Sig. Proc. Mag.*, 1998. [3] M. Jerry, *DRC*, 2016, pp. 1–2. [4] J. Lin, *IEDM*, 2016, pp. 2–5. [5] A. A. Sharma, *VLSI*, 2016, pp. 2–3. [6] K. Moon, *IEDM*, 2015, p. 17.6.1-17.6.4. [7] H. Madan, *ACS Nano*, vol. 9, no. 2, pp. 2009–17, 2015 [8] J. Frougier, *VLSI*, 2016, pp. 1-2.

Acknowledgment: This project was supported by the National Science Foundation under grant 1640081, and the Nanoelectronics Research Corporation (NERC), a wholly-owned subsidiary of the Semiconductor Research Corporation (SRC), through Extremely Energy Efficient Collective Electronics (EXCEL), an SRC-NRI Nanoelectronics Research Initiative under Research Task IDs 2698.001 and 2698.002.

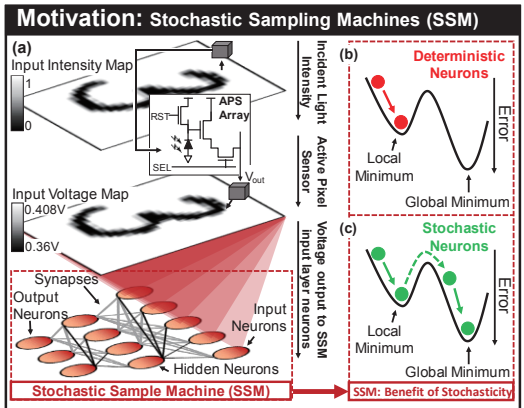


Fig. 1: (a) Image sensor input to stochastic sampling machine (SSM) network. For deterministic neurons (b) the system remains in a local minimum while stochastic neurons (c) enable the network to escape local minima and reach the global minimum of the error surface.

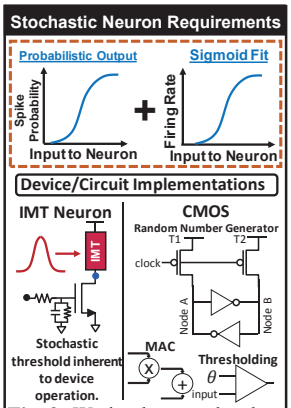
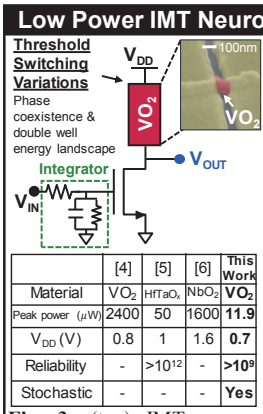


Fig. 2: We implement ultra-low power stochastic IMT neurons capable of implementing SSMs. The results are benchmarked against a 22nm CMOS ASIC.



	[4]	[5]	[6]	This Work
Material	VO_2	$HfTaO_x$	NbO_2	VO_2
Peak power (μW)	2400	50	1600	**11.9**
V_{DD} (V)	0.8	1	1.6	**0.7**
Reliability	-	>10^{12}	-	**>10^9**
Stochastic	-	-	-	**Yes**

Fig. 3: (top) IMT neuron structure with SEM inset. (bottom) Benchmarking IMT neurons demonstrated in this work and published results.

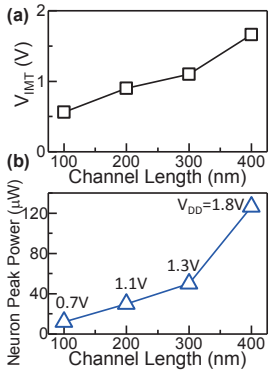


Fig. 4: (a) IMT trigger voltage (V_{IMT}) and (b) IMT neuron power scale with VO_2 channel length (L_{VO2}), allowing record low power.

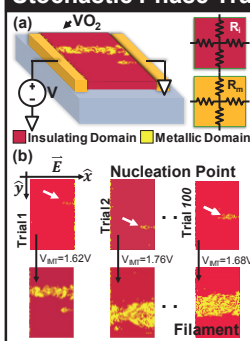


Fig. 5: (a) 2D heterogeneous resistive network model reveals (b) cycle-to-cycle variations in V_{IMT} occur from variations in the nucleation phase transition.

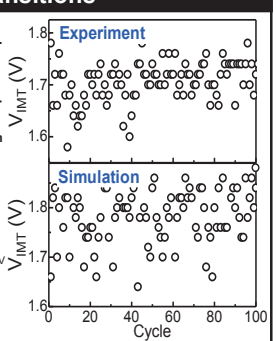


Fig. 6: Cycling shows variations are not a result of V_{IMT} drift. Supported by >10^9 endurance [8].

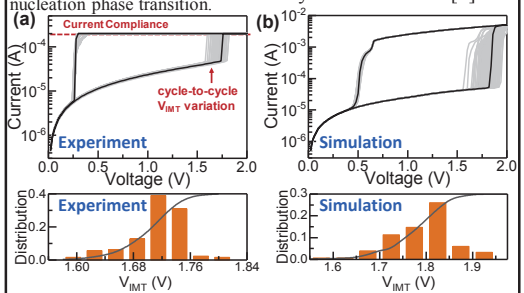


Fig. 7: (a) Experimental variations in the VO_2 DC characteristics and their probability distributions accurately reproduced by the network model in (b).

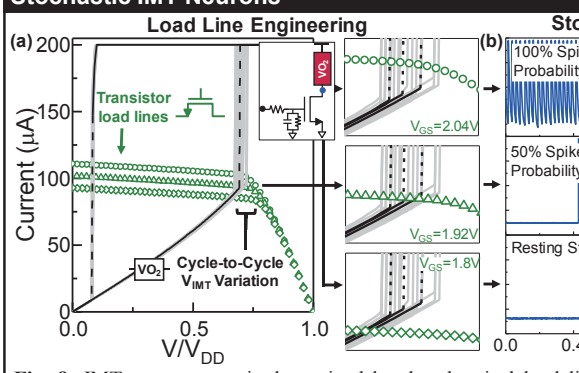


Fig. 8: IMT neuron state is determined by the electrical load line of the transistor. As V_{GS} increases such that the transistor load-line periodically crosses both unstable arms (dashed line) of the VO_2 characteristics, probabilistic spiking occurs due to occasional oscillations in the VO_2.

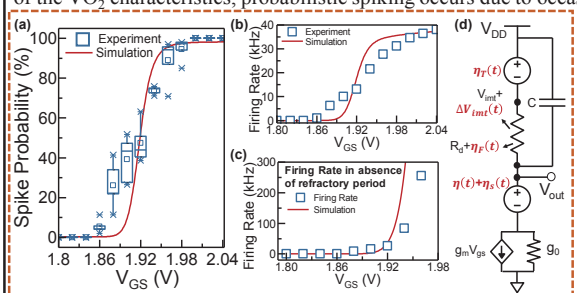


Fig. 9: Extracted spike probability and firing rates (from Fig. 8) display the required (a-b) sigmoidal and (c) exponential response for neural sampling in SSMs. (d) Comprehensive noise model fits the measured data.

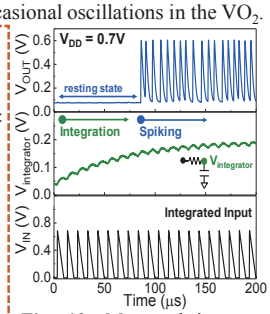


Fig. 10: Measured integrate and fire (I&F) response of IMT neuron at V_{DD}=0.7V.

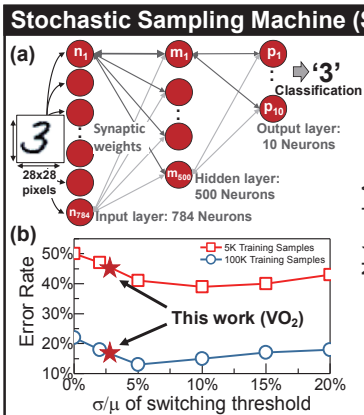


Fig. 11: (a) SSM network implemented using IMT neuron model from Fig. (9) on MNIST. (b) IMT neurons reduce the error rate by 7.5% for 100k training sets.

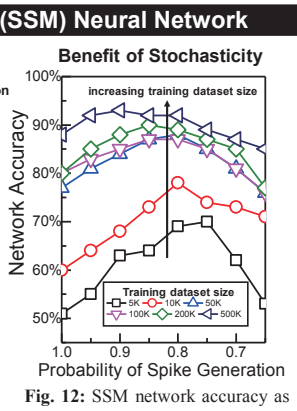


Fig. 12: SSM network accuracy as a function of neuron stochasticity (p=1=deterministic) show that stochastic neurons can increase the network accuracy by up 25% (for small training datasets - 5K).

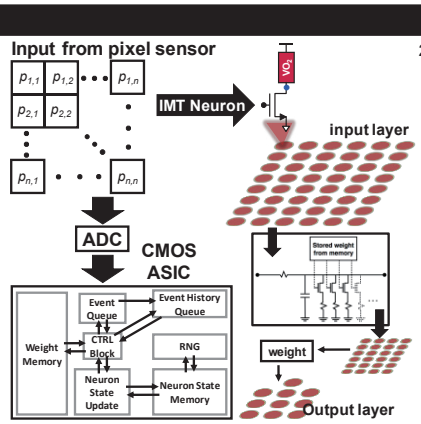


Fig. 13: Schematics of SSM based processing scheme for 22nm CMOS ASIC and IMT neuron. MNIST handwritten digit database used as input (28x28 pixels).

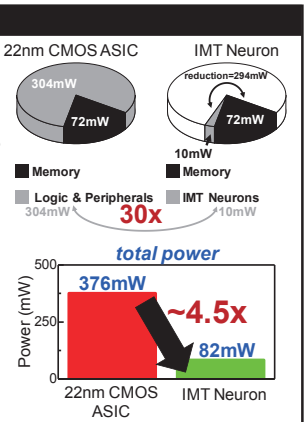


Fig. 14: IMT neurons result in 4.5x reduction in power over 22nm CMOS ASIC with matched memory power and network accuracy.