

Configurable p-Neurons Using Modular p-Bits

Saleh Bunaiyan^{1,2†}, Mohammad Alsharif^{3,4†}, Abdelrahman S. Abdelrahman¹, Hesham ElSawy⁵,

Suraj S. Cheema⁶, Suhaib A. Fahmy⁴, Kerem Y. Camsari¹, and Feras Al-Dirini^{5,6*}

¹ECE, UCSB, Santa Barbara, CA, USA, ²EE, KFUPM, Dhahran, KSA, ³COE, KFUPM, Dhahran, KSA,

⁴CEMSE, KAUST, Thuwal, KSA, ⁵School of Computing, Queen’s University, Kingston, ON, Canada,

⁶Research Laboratory of Electronics, MIT, Cambridge, MA, USA,

[†]equally contributing authors, *email: aldirini@mit.edu

Abstract—Probabilistic bits (p-bits) have recently been employed in neural networks (NNs) as stochastic neurons with sigmoidal probabilistic activation functions. Nonetheless, there remain a wealth of other probabilistic activation functions that are yet to be explored. Here we re-engineer the p-bit by decoupling its stochastic signal path from its input data path, giving rise to a modular p-bit that enables the realization of probabilistic neurons (p-neurons) with a range of configurable probabilistic activation functions, including a probabilistic version of the widely used *Logistic Sigmoid*, *Tanh* and *Rectified Linear Unit (ReLU)* activation functions. We present spintronic (CMOS + sMTJ) designs that show wide and tunable probabilistic ranges of operation. Finally, we experimentally implement digital-CMOS versions on an FPGA, with stochastic unit sharing, and demonstrate an order of magnitude (10x) saving in required hardware resources compared to conventional digital p-bit implementations.

Index Terms—AI, decoupled, modular, MTJ, neural network, neuron, p-bit, p-computing, probabilistic, p-neuron, stochastic.

I. INTRODUCTION

With increasing demand for data-intensive computing, unconventional paradigms are emerging [1], including probabilistic computing, in which probabilistic-bits (p-bits) based on stochastic magnetic tunnel junctions (sMTJs) [2]–[9] are central building blocks. Recently, there has been increasing interest in the realization of neural networks (NNs) employing p-bits as stochastic neurons [10], [11], which have shown high energy-efficiency and low area requirements, compared to deterministic NNs [12]. The p-bit was first proposed as a modified version of magnetic random access memory (MRAM) technology [4], where the main difference is that the MTJ is made stochastic rather than being deterministic (Fig. 1 (a)).

In this work, we also adopt another inspiration from MRAM technology; the separation between – or the decoupling of – the write and read paths [13], [14]. We propose a novel decoupling approach in the p-bit between the stochastic path and the input path, such that the effect of each of the two paths on the p-bit response is independent of the other. Accordingly, each path can be engineered separately, resulting in a modular p-bit (Fig. 1 (b)) [15], [16]. We leverage this approach to customize the response of the p-bit, implementing probabilistic neurons (p-neurons) [17] with probabilistic versions of widely used activation functions in NNs; namely *Tanh*, *Logistic Sigmoid*, and *Rectified Linear Unit (ReLU)*. Transistor-level analog spintronic (CMOS + sMTJ) designs are shown in Fig. 2, while digital CMOS designs are shown in Fig. 5.

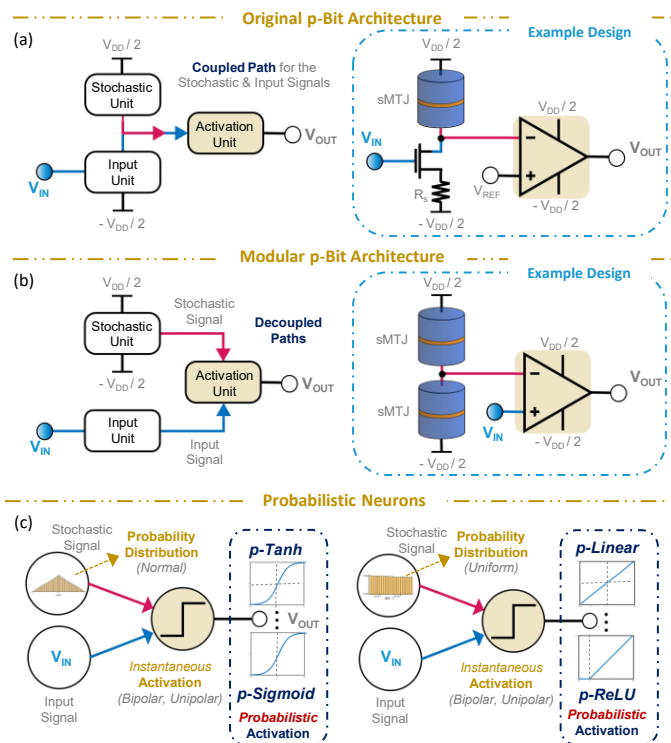


Fig. 1. P-neurons via decoupled modular p-bits. (a) Original p-bit architecture where the stochastic and input paths are coupled. The original design is shown, where the two paths are coupled at the drain node. (b) Proposed architecture that decouples the stochastic path from the input path, decoupling the design of these two paths. A design example based on a dual sMTJ voltage divider cell is shown. (c) P-neurons with different probabilistic (time-average) activation functions by modular engineering of the stochastic and the activation units.

II. DECOUPLED ARCHITECTURE FOR MODULARITY

In the original “*Coupled Architecture*” of the p-bit changing the input directly alters the stochastic signal generated by the stochastic unit. A common example is the original p-bit design [4], shown in Fig. 1 (a), where the input voltage V_{IN} directly affects the stochastic response at the drain node [4], [7]. On the other hand, in our “*Decoupled Architecture*”, the effect of the stochastic unit response is decoupled from the input signal (Fig. 1 (b)). A dual sMTJ voltage divider (2M cell) is shown in Fig. 1 (b) as an example design of the stochastic unit. Decoupled designs can be utilized to generalize the p-neuron

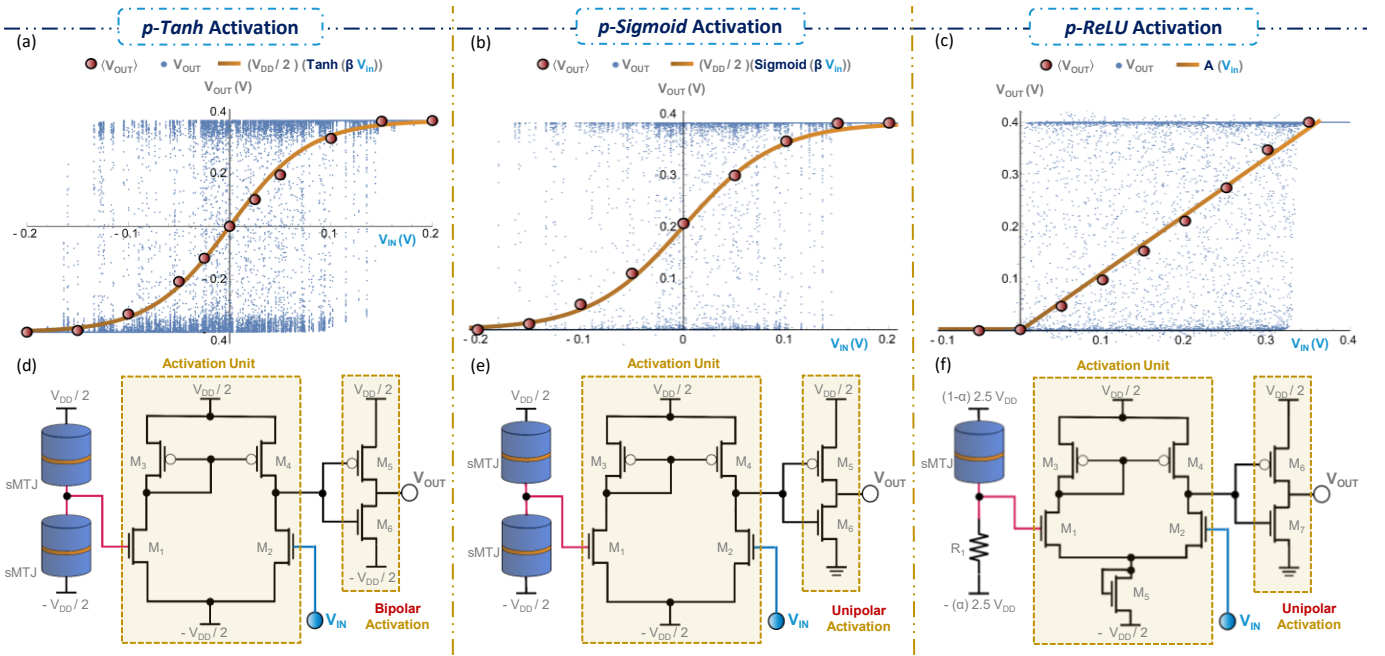


Fig. 2. Probabilistic Neurons: spintronic (CMOS + sMTJ) design examples. (a)–(c) response of p-neurons (blue dots: instantaneous response, large orange circles: time-averaged response) with p -Tanh, p -Sigmoid, and p -RELU activation functions, respectively. Non-bipolar data points are due to the limited slew-rate of the amplifier [18]–[21]. (d) and (e) transistor-level circuit designs for implementing p-neurons with p -Tanh and p -Sigmoid activation functions, respectively. The stochastic unit is a dual sMTJ voltage divider (2M cell). For p -Sigmoid, the activation unit was re-engineered to obtain and optimize unipolar activation of the p-neuron through the reduction of the W/L of M_6 by a factor of three. (f) transistor-level circuit design for implementing a p-neuron with a p -RELU activation function, using a single sMTJ + single resistor (1M1R) cell as the stochastic unit, with $R_1 = 0.35/G_0$ and $\alpha = 0.155$. For all designs $V_{DD} = 0.8$.

response to any probabilistic response with proper engineering of the stochastic unit and the instantaneous activation unit, as shown in Fig. 3 (a)–(b) and Fig. 2. In this work, we exploit the modularity of the decoupled architecture, showing that not only can we implement spintronic-based modular p-neuron designs, but digital CMOS versions also, which we experimentally realize using a Field-Programmable Gate Array (FPGA), with customizable and tunable probabilistic activation functions, as well as shared stochastic units.

III. SPINTRONIC DESIGNS (CMOS + STOCHASTIC MTJs)

The first decoupled design example employs a stochastic unit of two identical sMTJs connected in series in a 2M cell (Fig. 3 (a)), which we simulate in HSPICE using the stochastic Landau-Lifshitz-Gilbert (sLLG) equation [22]. We consider an sMTJ with perpendicular magnetic anisotropy (PMA) and no effective field ($\Delta_B \approx \vec{H}_{eff} \approx 0$). The magnetization \hat{m} of such an sMTJ was theoretically proven to have uniform distribution over all directions [23]. Hence, the sMTJ conductance $G(t)$ will also be uniform. We assume a free layer diameter of 22 nm, a polarization of 0.7, and a low G_0 to minimize spin-transfer torque (STT); preserving the uniform randomness of $G(t)$, while keeping other simulation parameters as in previous work [4]. Then we couple this stochastic unit to a differential amplifier followed by pull up (PU) and pull down (PD) networks, constructed using an inverter (Fig. 2 (d)). All CMOS transistors are simulated using the HP 14-nm FinFET predictive technology model (PTM) [24]. The time-

averaged response of this design implements a probabilistic Tanh (p -Tanh) activation function as shown in Fig. 2 (a), where the supply voltage swing is bipolar. Note that this realization eliminates the requirement for matching the sMTJ with the transistor. By fixing the stochastic branch and engineering the

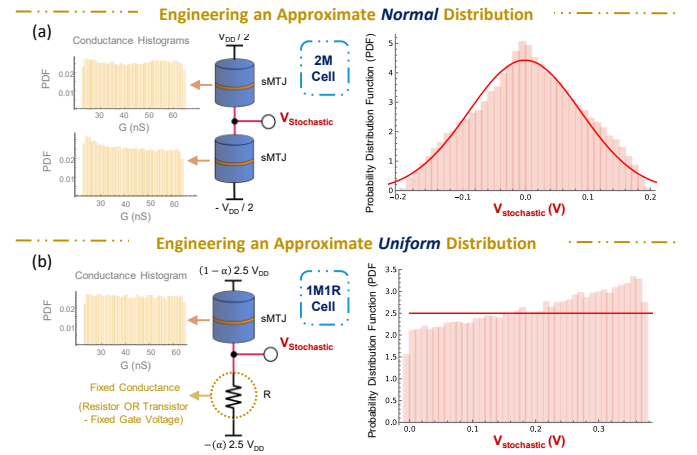


Fig. 3. Engineering the stochastic unit: (a) A dual sMTJ stochastic unit - 2M cell (left), consisting of two sMTJs in series (each with a uniform conductance distribution). The probability distribution of $V_{Stochastic}$ of the 2M cell that is approximately normal (right). (b) A stochastic unit that consists of one sMTJ and one fixed resistor connected in series - 1M1R cell (left), with $R_1 = 0.35/G_0$ and $\alpha = 0.155$. The probability distribution of $V_{Stochastic}$ of the 1M1R cell that is approximately uniform (right).

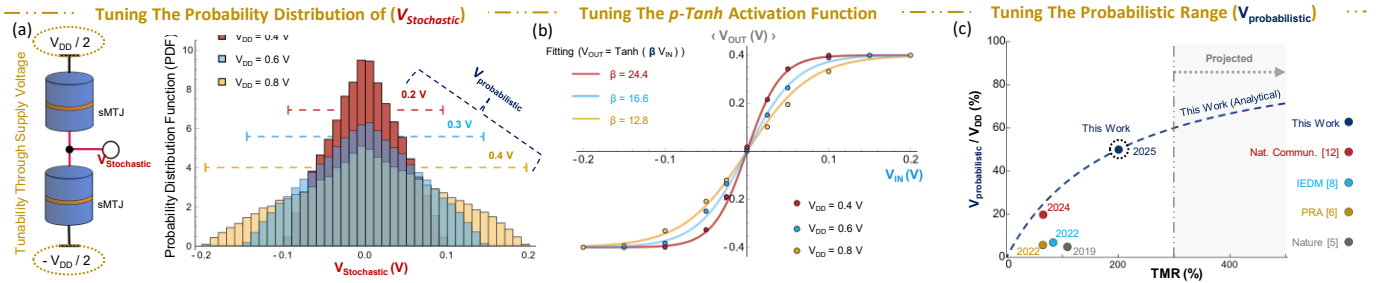


Fig. 4. Tunability of the p-neuron’s probabilistic activation function. (a) For a stochastic unit of two sMTJs in series (2M cell), the characteristics of the $V_{Stochastic}$ probability distribution (mean and variance) can be tuned by V_{DD} . (b) Tuning the p-neuron’s probabilistic range of the p -Tanh activation function. The p -Tanh tunability is controlled by a scaling factor β . (c) Analytical limit that describes the tunability of the probabilistic range ($V_{probabilistic}$) as a function of the sMTJ tunneling magnetoresistance (TMR). The other data points refer to other experimental realizations of p-bits.

activation unit, we can adjust the activation function to be a probabilistic sigmoid (p -Sigmoid), as in Fig. 2 (b) and (e).

Another design example is implemented by replacing one of the sMTJs with a fixed resistance in the stochastic unit, resulting in a 1 sMTJ + 1 resistor (1M1R cell) stochastic unit. In this example a probabilistic ReLU (p -ReLU) activation function is realized (see Fig. 2 (f)), which is achieved by adjusting the voltage $V_{Stochastic}$ between R_1 and the sMTJ to have a uniform distribution (Fig. 3 (b)). The uniform random conductance of the sMTJ does not result in an ideal uniform distribution at $V_{Stochastic}$ in the 1M1R cell, however, it can be adjusted to become near-uniform by choosing a suitable value for R_1 and extending the supply voltage of the stochastic unit as shown in Fig. 3 (b). Other approaches and technologies used in encryption [25]–[30] and sensing [31]–[34], can also provide such a distribution. The time-averaged response of the p -ReLU activation function is shown in Fig. 2 (c).

IV. PROBABILISTIC RANGE TUNABILITY

Using the supply voltage V_{DD} across the two sMTJs, we can shrink or stretch the range of input voltages to which the p-neuron has a stochastic response ($V_{Stochastic}$), as shown in Fig. 4 (a). We define this range of input voltages here as the probabilistic range of the p-neuron ($V_{probabilistic}$). Hence, using V_{DD} , we can tune this probabilistic range of operation of the p-neuron, as shown in Fig. 4 (b). The value of $V_{probabilistic}$ relative to the overall supply voltage, obtained using the 2M cell stochastic unit, can be described as a function of the sMTJ’s tunneling magnetoresistance (TMR):

$$V_{probabilistic}/V_{DD} = TMR/(2 + TMR) \quad (1)$$

where $TMR = (R_{AP} - R_P)/R_P$, R_{AP} and R_P are antiparallel and parallel resistances of the sMTJ, respectively. This analytical limit of $V_{probabilistic}/V_{DD}$ is plotted in Fig. 4 (c), showing that, at high TMR (around 300 %), a probabilistic range of around 60 % of V_{DD} can be attainable. Although higher values of TMR were reported for stable MTJs [35], [36], sMTJs are still at early stages and their limits are yet to be discovered. Nonetheless, the plot shows an enhancement in $V_{probabilistic}$ for the modular p-neuron, across a range of TMR values, compared to previous designs reported in the literature.

V. DIGITAL CMOS FPGA IMPLEMENTATIONS

To demonstrate the generality of our approach, we reproduce the probabilistic activation functions obtained earlier using digital CMOS implementations on an FPGA, employing linear-feedback shift registers (LFSRs) for the stochastic units. Since an LFSR naturally generates a uniform (pseudo) random variable, we can approximate the response of the p -Tanh and p -Sigmoid activation functions by using a stochastic unit that employs two 32-bit LFSRs and sums their output using a 32-bit adder. This addition of two uniform distributions constructs an Irwin–Hall distribution that approximates the Gaussian/normal distribution needed for these probabilistic activation functions. The generated stochastic sum (usually referred to as the “random number” in digital systems) is compared to the digital input string (I_{IN}), represented by a 32-bit fixed point unsigned binary representation, using a 32-bit digital comparator. On the other hand, for the p -ReLU and probabilistic linear (p -Linear) p-neurons, only a single LFSR is required to construct the stochastic unit, as it generates the needed random number with a uniform distribution. In both cases, 2’s complement binary representation is used.

The digital designs of each of these p-neurons are shown in Fig. 5 (e)–(h), building upon previous work [37]. However, all of these modular p-neuron designs, contrary to earlier designs [37], do not require a lookup table (LUT) to achieve the required activation function; instead, the input is directly compared to the output of the stochastic unit. It is the stochastic signal’s probability distribution that controls and customizes the activation function, minimizing FPGA hardware resource requirements (Fig. 5(n)). For the p -ReLU p-neuron, a rectification unit is needed within the activation unit, as shown in Fig. 5(g), which is a multiplexer that is enabled by the most significant bit of the input I_{IN} (Fig. 5(g)).

VI. STOCHASTIC UNIT SHARING

The experimental results shown in Fig. 5 (a)–(d) not only reproduce the desired activation functions, but also demonstrate how – enabled by their modular design – multiple p-neurons can share the same stochastic unit, yet still generate different activation functions (Fig. 5). This is achieved by independently engineering each p-neuron’s instantaneous activation unit, as

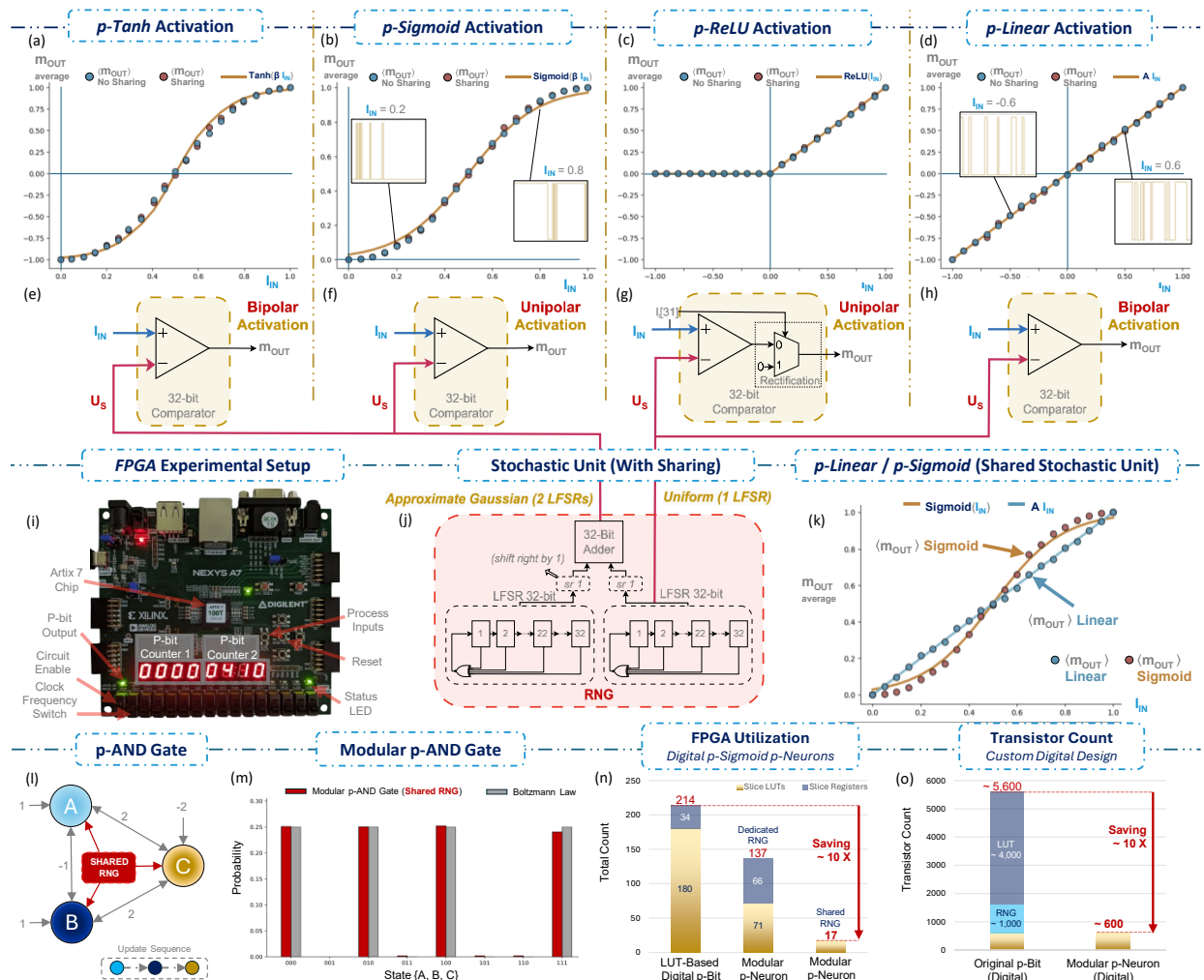


Fig. 5. Digital p-neurons and stochastic unit sharing experiments. (a)–(d) Time-averaged response of digital p-neurons with (a) p -Tanh (b) p -Sigmoid (c) p -ReLU and (d) p -Linear activation functions, implemented on an FPGA. (e)–(h) The p-neuron digital designs employing 32-bit comparators, which compare the input string I_{IN} with the random number generated by the stochastic unit. (i) Experimental setup for the FPGA implementation. (j) Shared stochastic unit. (k) The result of an FPGA experiment where one p -Sigmoid p-neuron and one p -Linear p-neuron receive their random numbers, with different probability distributions, from a single shared stochastic unit. (l) A modular probabilistic AND (p-AND) gate with 3 p-neurons that share a single RNG as their stochastic unit. The interconnections between the p-neurons are shown with weights on bidirectional arrows and biases on unidirectional arrows. (m) Histogram showing the probability distribution of visiting the correct p-AND gate states, and the expected Boltzmann distribution. (n) Savings in FPGA hardware resource utilization. (o) Comparison of the estimated transistor count in custom digital designs of modular p -Sigmoid p-neurons with original FPGA p-bits [12].

shown in Fig. 5 (e)–(h). This sharing of LFSRs further minimizes FPGA hardware resource requirements, as shown in Fig. 5 (n), reaching beyond an order of magnitude compared to conventional FPGA p-bits. Moreover, in custom p-neuron digital designs, leveraging stochastic unit sharing and the fact that no LUT is needed leads to each additional digital p-neuron requiring almost one order of magnitude less transistors, when compared to conventional LUT-based digital p-bit implementations [12], as shown in Fig. 5(o). Stochastic unit sharing can also be implemented between multiple p-neurons with different classes of activation functions, where each neuron requires a different stochastic unit as in Fig. 5(k). Power savings are also expected; but require future analysis.

An example 3 p-neuron circuit is shown in Fig. 5(l), implementing a probabilistic AND (p-AND) gate that can be operated in both forward and reverse modes. All p-neurons have

p -Sigmoid activation functions and are configured in a fully connected Boltzmann machine configuration. All p-neurons share the same stochastic unit. The probability distribution of visiting states for the 3 p-neuron network, shown in Fig. 5(m), confirms the network’s functionality as a p-AND gate, and matches the expected Boltzmann distribution.

VII. CONCLUSION

We presented spintronic (CMOS + sMTJ) and digital p-neuron designs based on modular p-bits, exhibiting a range of configurable probabilistic activation functions that are highly tunable, and a modular ability of stochastic unit sharing. The digital p-neuron designs do not require LUTs, providing further savings in hardware resources. Our work paves the way toward accessible and scalable hardware implementations of Probabilistic Neural Networks using modular p-neurons.

