A 104.8TOPS/W One-Shot Time-Based Neuromorphic Chip Employing Dynamic Threshold Error Correction in 65nm

Luke R. Everson, Muqing Liu, Nakul Pande, and Chris H. Kim
Department of Electrical and Computer Engineering
University of Minnesota
Minneapolis, MN 55455 USA

***Abstract*- As neural networks continue to infiltrate diverse application domains, computing will begin to move out of the cloud and onto edge devices necessitating fast, reliable, and low power solutions. To meet these requirements, we propose a time-domain core using one-shot delay measurements and a lightweight post-processing technique, Dynamic Threshold Error Correction (DTEC). This design differs from traditional digital implementations in that it uses the delay accumulated through a simple inverter chain distributed through an SRAM array to intrinsically compute resource intensive multiply-accumulate (MAC) operations. Implemented in 65nmLP CMOS we achieve, to our knowledge, the lowest reported energy efficiency for a neuromorphic processor with 52.4TSOp/s/W (104.8TOp/S/W) at 0.7V with 3b resolution for an impressive 19.1fJ/MAC.**

I. INTRODUCTION

The ever-increasing demand for higher performance and energy efficiency in machine learning (ML) applications has driven an impressive range of ASICs [1-7] aimed at meeting the needs. Digital SoCs [3-5,7] have found success by restricting the weight resolution [8], changing memory access structures, and guarding operations when the input is zero. However, all require large registers to store intermediate results and complex multiplier blocks. An emerging trend [1,2] has been to employ time-domain circuits to implement dot-products; the main kernel for ML applications. Fig. 1 details how the dot-product is computed in the time-domain and in conventional digital implementations. In time-domain the delay is modulated by the application inputs and weights to generate proportional delays. These delays are accumulated and can be routed to a Time-to-Digital Converter (TDC) or counter to be processed for use in the deep learning application. Alternatively, the digital approach relies on many multiplier blocks and wide merging adders, typically in an array-like structure, to generate dot-products. The primary benefit of time-domain circuits is that the accumulate portion of the MAC is intrinsic to the architecture. Additionally, the processing unit can be realized as a collection of inverters making the area and active power consumption very low. Digital methods can leverage existing IP-blocks for multipliers and adders and do not require calibration, unlike the time-domain circuits. Additionally, digital circuits can handle higher bit operations more effectively due to the binary encoding. Previous time-domain neuromorphic chips have fundamental limitations. In [2], a digitally controlled oscillator was used to modulate the frequency, by switching capacitor loads representing the weights, while the number of cycles in a set sampling period was counted. While this closed loop structure has the benefit of canceling temporal noise, it must oscillate for many cycles to generate a result. Reference [1] is also a delay line based approach, but the outputs and weights are restricted to binary. More critically, their design has twice the area overhead do to the fact they utilize local reference delay lines instead of a global reference, and can limit the potential scalability of the architecture. In this work, we have addressed the shortcomings of previous designs by implementing Digitally controlled Delay Lines (DDLs) that are compared to

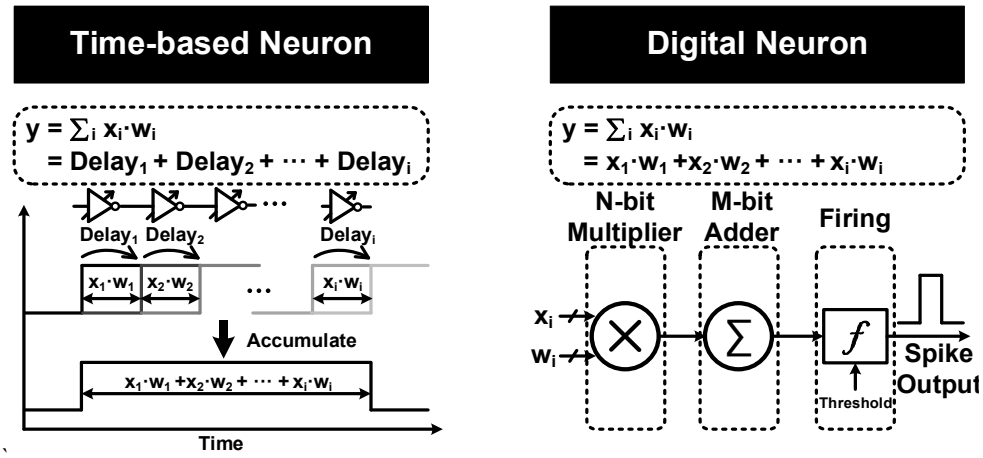


Fig. 1. Time-based neurons utilize the delay through basic circuit elements such as inverters to implement the dot-product. Digital neurons use conventional Boolean logic for arithmetic operations. Both architectures can be mapped to deep learning applications.

This research was supported in part by the National Science Foundation under award number CCF-1763761 and IGERT grant DGE-1069104.

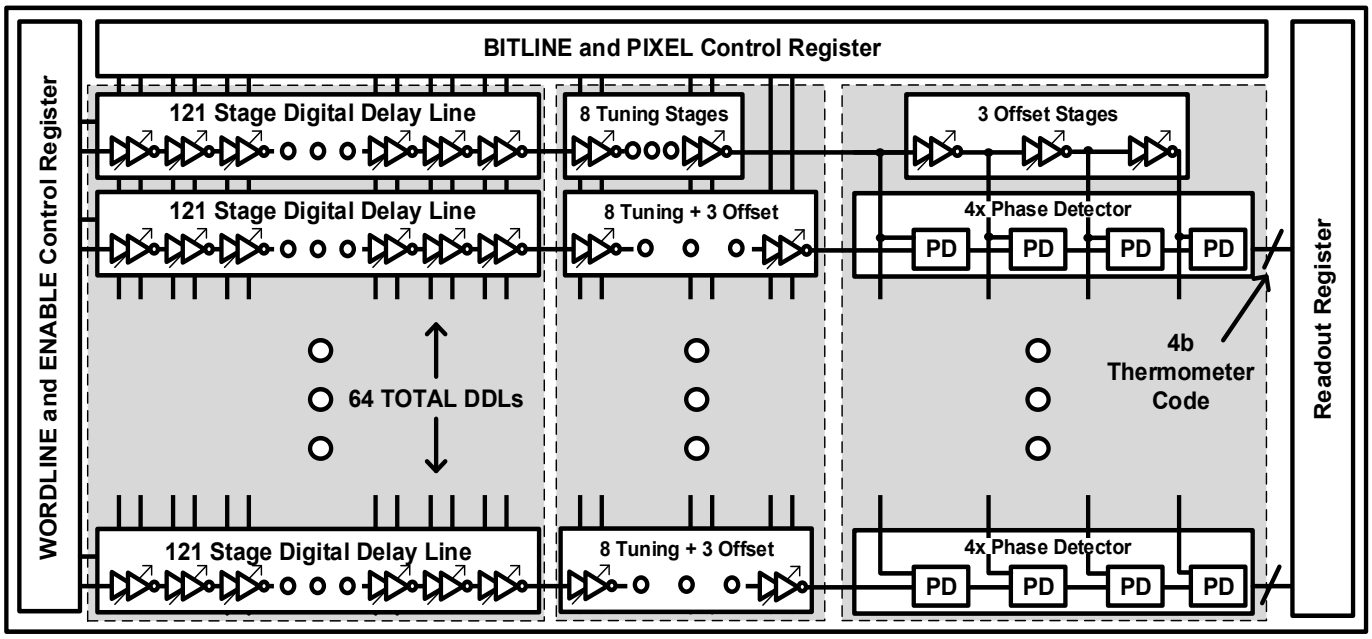


Fig. 2. Top level schematic of the time-based neuromorphic core. Layout is based on SRAM array. The core contains 64 DDLs each with 129 stages and each has a 4b phase detector that is compared to the reference DDL.

a shared reference delay to compute multi-bit MACs. We will explain the operating concept, novel accuracy boosting technique DTEC, and describe the chip measurement results in the following sections.

II. One-Shot and Time-Based Neuromorphic Concept

Conventionally, Boolean computations are used to realize arithmetic operations in hardware. However, time domain circuits can also be used at an advantage of lower area and power per processing unit, and reduced design complexity. The kernel of all ML algorithms can be distilled into a dot product; $y = \sum xw + b$, where x is an input vector, w a weight matrix, and b is a bias, or offset vector. Our high level architecture is shown in Fig. 2. An input pulse is presented on the left side of the core and the delay of each stage is modulated based on the application inputs. Each stage has 8 delay units (DU) with output taps which the pulse travels through as seen in Fig. 3. The number of DU enabled depends on the weight, stored locally in SRAM cells, and the input pixel, which is applied across the array on the bitlines. Each DU has two inverters to retain consistent polarity between stages. This is critical in the event that the rising and falling propagation delays are not matched, as well as ensuring correct polarity at the TDC. The output tap is realized as a complex tristate gate and the functionality is described in Fig. 4. The first column shows the circuit schematic and corresponding connections between the different DUs. The right four columns show the activated paths, shown with black lines, depending on the values of the input and weights. DU_5 is the nominal stage delay, and is activated through the right branch of the circuit when the pixel is not present, representing "zero delay." The right table shows the mapping between the algorithm weights and the delays realized in the chip. When the input is present the

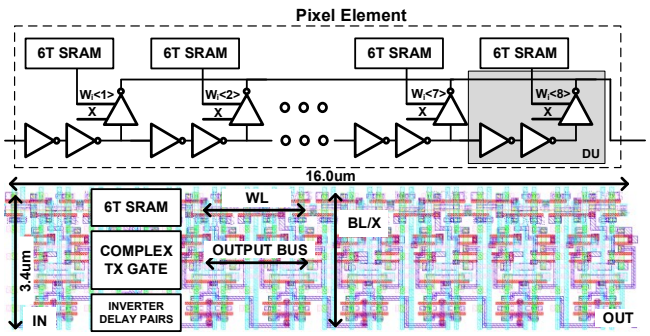


Fig. 3. (Top) Schematic of pixel stage. Complex tristates (Fig. 4) drive output bus. (Bottom) Layout of pixel stage.

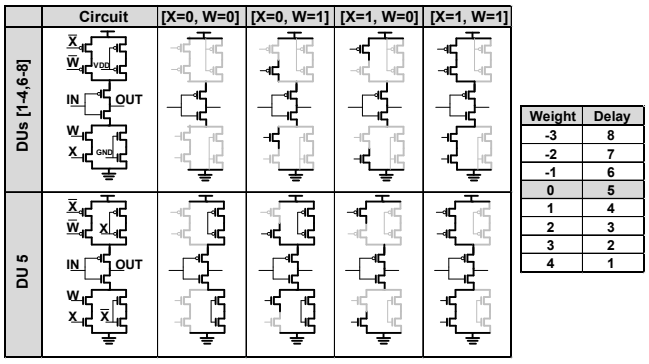


Fig. 4. Left table shows the complex tristate connections used to implement dot product based on input and weights. Right table shows 3b weight-delay mapping.

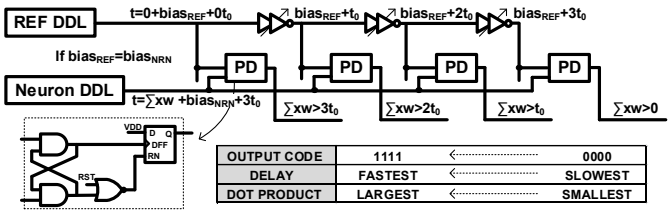


Fig. 5. Details of delay to dot product relationship.

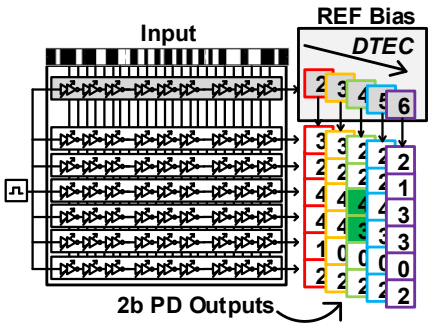


Fig. 6. DTEC concept. Reference DDL bias is increased to elucidate the strongest DDL.

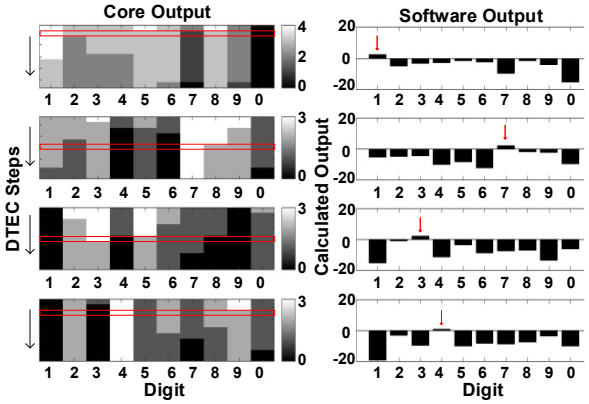


Fig. 7. Colormaps (left) show outputs from core after DTEC. Each row corresponds to results from successive values of the bias. The red rectangle highlights where DTEC has identified the dominant output. Bar plots (right) show expected results from software.

left branch is enabled in the DU corresponding the weight of the stage. The accumulation in the MAC is achieved naturally as the pulse passes sequentially through the DDL, stage by stage. The layout of each DU in the stage is pitch matched to a 6T SRAM so the layout is regular, compact, and scalable. The bias vector is applied in the same way for the last eight units. Additionally, it can be used to tune process variation, so that during evaluation those pixels are always activated. Fig. 5 shows the relationship between the time domain computation in the chip and the expected arithmetic output. The phase detector output maps roughly to the Relu transfer function. When the reference pulse beats the neuron rising edge all four thermometer bits are zero, regardless of the magnitude. The transfer function between the four bits is linear and then clips, or saturates, once the neuron pulse is faster than all the offsets.

III. Dynamic Threshold Error Correction (DTEC)

In this design we opt for a 2 bit TDC due to the optimal tradeoff between small area and low power, and strong architecture performance. In networks with "winner-take-all" topologies, such as the last stage of classification networks, ambiguous predictions can occur. Unclear outputs in this work can stem from limited resolution between phase detector trip-points or activity outside of the range of the phase detector. To mitigate this issue, we propose a Dynamic Threshold Error Correction (DTEC) technique which increases the effectiveness of the 2bit TDC. As shown in Fig. 6, when two or more DDLs have the same output, DTEC works by increasing the threshold bias delay which moves the trip point of the phase detectors.

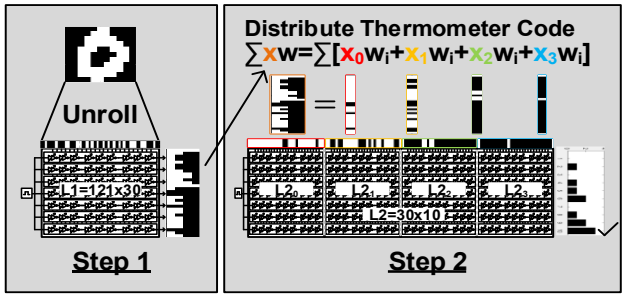


Fig. 8. Dataflow for multilayer time-based deep neural network demonstrated in this work.

DTEC is dynamic due to the fact that the bias sweep would be terminated after the third evaluation, when the dominant DDL was identified from the phase detectors. Additionally, DTEC can be stopped after a fixed number of steps if no dominant DDL emerges to conserve power. In Fig. 7 the top row of colormaps shows ambiguous predictions from the core, while successive rows show the output as DTEC is applied. Red rectangles highlight where DTEC has successfully identified the target. Analysis of results from the 3b single layer application (section IV) show that by applying just two DTEC steps 81.64% of the correctible errors are recovered. This comes at a cost of just 0.41 additional evaluations per image. After the one-shot evaluation 73% of all images have a dominant output. The remaining 2,668 images begin DTEC and after the first step 46% are resolved and 37% after the second step leaving less than 1,000 images ambiguous. Thus, 4,108 DTEC evaluations improves the total accuracy from 69.16% to 82.14%. If three DTEC steps are applied 88.8% of errors can be recovered at an overhead of 0.51, demonstrating the dynamic scalability of the technique. DTEC is an economical and scalable approach to significantly improve application performance.

IV. Test Chip Measurements and Application Details

We evaluate the core on the MNIST benchmark [9]. Fig. 9 shows the comparison of classification accuracy on an 11x11 image for single and two layer networks between expected simulated software results, one-shot evaluation, and DTEC. To reduce the 28x28 grayscale images to 11x11 binary images, 3 pixels are sliced from all four sides of the image. Then, a fixed resizing command is applied, and finally the pixels are binary thresholded. Fig. 8 shows how the core can be used in a multi-layer deep neural net application. Each bit of the thermometer

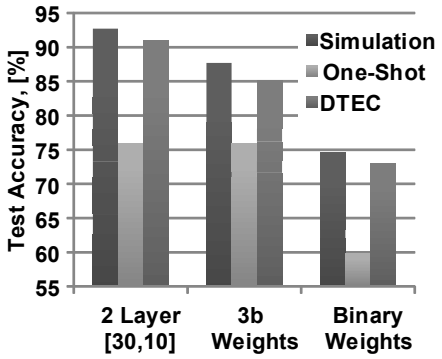


Fig. 9. Results on 11x11 MNIST for 3b two layer, 3b and binary

TABLE I. COMPARISON TABLE

	This Work		A-SSCC'16 [1]	CICC'17 [2]	ISSCC'17 [3]	ISSCC'17 [4]	ISSCC'16 [5]	ISSCC'16[6]	Science'14[7]
Chip Architecture	Time-Based		Time-Based	Time-Based	Digital	Digital	Digital	Sw. Cap	Digital
Algorithm Target	FCDNN & CNN		FCDNN & CNN	FCDNN & CNN	FCDNN & CNN	FCDNN & FFT	CNN	CNN & SGD	FCDNN & CNN
Technology [nm]	65		65	65	28 FDSOI	40	65	40	28
Chip Area [mm^2]	0.644		3.61	0.24	1.87	7.1	12.25	0.012	430
Precision* [b]	[B,T,2,3]		B	3	[4-16]	[6-32]	16	3	[B,T]
On-Chip SRAM [kB]	8.06		20	3	144	270	181.5	[-]	256MB
VDD [V]	1.2 (Nom.)	0.7 (E_{MAx})	1	1.2	0.6	0.65	0.82	1	0.85
Frequency [MHz]	1700	285	23041	792	200	19.3	250	1000	0.001
Energy Efficiency** [TSop/s/W]	36.2	52.4	48.2	2.47	5.0	0.19	.18	3.86	0.04
Hardware Efficiency [GE/PE][1]	38.4		76.5	33.2	7456	18269	50637	288	6.5

*B=Binary, T=Ternary **Synaptic Op=MAC

code is expanded as the input in the next layer. The input is divided into four segments, and the weight matrix is copied four times ($L2_0$-$L2_3$), which gives each bit equal weighting. In the example shown in Fig. 8, 30 neurons in layer 1 yield a 120 bit input to layer 2. By applying DTEC, the ambiguous results are almost completely recovered and the slight loss in accuracy is due to output differences smaller than a single tuning bit. Fig. 10 shows the tradeoff between power consumption and nominal stage delay for various supply voltages. Power is kept exceptionally low because rarely are more than two stages switching at a time in a DDL. A wide operating voltage range is enabled, due to the all-digital time-based design choices. If the design incorporated pipelining, it could achieve even greater throughput. That is, multiple pulses could be pushed into the DDL and the input could shift as well. This is ideally suited for Convolutional Nets where a weight filter slides across an image. In this case, the image could slide across the weights while input pulses are applied to the DDL. Die photo and design specs are highlighted in Fig. 11. Table I shows strong performance compared with state of the art. All comparisons

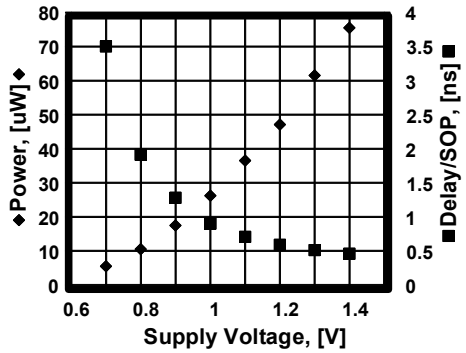
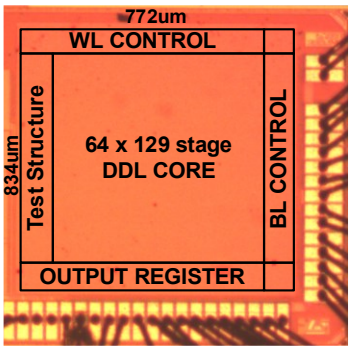


Fig. 10. Power consumption of a DDL and delay/stage versus VDD.



Process	65nm LP CMOS
Core Area	0.644mm^2
VDD	0.7-1.4V
# of Neurons	64
# of Synapses	8256 (8K)
Throughput	1.7Gpixels/s/DDL
Power	47.1uW/DLL
Energy/SOP	27.6fJ

Fig. 11. Die photo and summary with reported metrics at 1.2V.

are made at the highest reported energy efficiency operating point. We report the best energy efficiency and a very competitive gate equivalent count for each processing unit at half the size of [1]. Our chip is scalable in voltage, weight resolution, and is versatile in that it is able to tackle fully connected deep networks as well as convolutional nets.

V. CONCLUSION

We described a time-based neuromorphic core based on one-shot DDLs in 65nm LP CMOS and proposed an error recovery technique, DTEC. It uses inverter delays to compute the dot product kernel, making it ideally suited for ML applications. The proposed core is validated on the MNIST dataset and achieves near simulated prediction accuracy on single and multi-layer networks after applying our error correction technique, DTEC. Maximum energy efficiency of 54.2TSOPs/s/W with 3b resolution at 0.7V makes the proposed architecture attractive for edge devices.

REFERENCES

[1] D. Miyashita, S. Kousai, T. Suzuki and J. Deguchi, "Time-domain neural network: A 48.5 TSOp/s/W neuromorphic chip optimized for deep learning and CMOS technology," *2016 IEEE Asian Solid-State Circuits Conference (A-SSCC)*, Toyama, 2016, pp. 25-28.

[2] M. Liu, L. R. Everson and C. H. Kim, "A scalable time-based integrate-and-fire neuromorphic core with brain-inspired leak and local lateral inhibition capabilities," *2017 IEEE Custom Integrated Circuits Conference (CICC)*, Austin, TX, 2017, pp. 1-4.

[3] B. Moons, R. Uytterhoeven, W. Dehaene and M. Verhelst, "14.5 Envision: A 0.26-to-10TOPS/W subword-parallel dynamic-voltage-accuracy-frequency-scalable Convolutional Neural Network processor in 28nm FDSOI," *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, 2017, pp. 246-247.

[4] S. Bang *et al.*, "14.7 A 288μW programmable deep-learning processor with 270KB on-chip weight storage using non-uniform memory hierarchy for mobile intelligence," *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, 2017, pp. 250-251.

[5] Y. H. Chen, T. Krishna, J. Emer and V. Sze, "14.5 Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, 2016, pp. 262-263.

[6] E. H. Lee and S. S. Wong, "24.2 A 2.5GHz 7.7TOPS/W switched-capacitor matrix multiplier with co-designed local memory in 40nm," *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, 2016, pp. 418-419.

[7] P.A. Merolla *et al.*, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668-673, Aug. 2014.

[8] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized Neural Networks," 2016 NIPS, Barcelona, Spain, 2016, pp. 4107-4115, Dec. 2016.

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE*, 86(11):2278-2324, Nov. 1998.