A Scalable Time-based Integrate-and-Fire Neuromorphic Core with Brain-Inspired Leak and Local Lateral Inhibition Capabilities

Muqing Liu, Luke R. Everson, and Chris H. Kim
University of Minnesota, Minneapolis, MN 55455, USA

***Abstract* — A fully scalable light-weight integrate-and-fire neuromorphic core with brain-inspired leak and local lateral inhibition features is implemented in 65nm. The core computes the neural net algorithm entirely in the time domain using standard digital circuits. A parallel two-layer architecture realized using the proposed core achieves a 91% handwritten digit recognition accuracy. The 0.24mm^2 neuromorphic core including 64 digitally controlled oscillator (DCO) circuits consumes 320.4μW per DCO at a maximum throughput of 746M pixels/s.**

***Index Terms* — Time-based circuits, integrate-and-fire, neuromorphic, leak, local lateral inhibition.**

I. Introduction

Neuromorphic brain-inspired computing using customized hardware has offered an attractive solution to overcome the von Neumann bottleneck. Many approaches have been proposed to model the neural processing elements in hardware, aiming at higher power and area efficiency. Early approaches relied on analog circuits to mimic synapse and neuron functions [1]. The main drawback of using analog circuits to implemental brain-inspired computing models is that they are sensitive to noise and process variation, so homogeneity and precision cannot be guaranteed for large scale designs. Scaling of CMOS technology also poses a challenge for analog designs. Digital implementation of neural nets has been a more popular approach lately. Compared to analog neural nets, they are less vulnerable to noise and process variation, and can benefit from technology scaling, enabling massively parallel neuromorphic ASIC systems such as IBM's Truenorth [2]. At the same time, device researchers have been exploring emerging materials and technologies to emulate the synapse function. For example, memristors and RRAM devices have been reported to be well suited for implementing synapses with improved circuit density and reduced power consumption. Despite the potential for entirely new neuromorphic implementations, their deployment has been limited due to fabrication constraints. This paper presents the idea of implementing neuron functions entirely in the time domain using standard digital circuits. Time-based circuits have several known advantages over voltage-mode or current-mode circuits such as high precision, compact area, low power consumption, excellent compatibility with advanced CMOS technologies, and low operating voltages.

II. Time-based Integrate-and-Fire DCO Neuromorphic

Fig. 1 shows the proposed time-based integrate-and-fire DCO neuromorphic core. The main innovation is that it

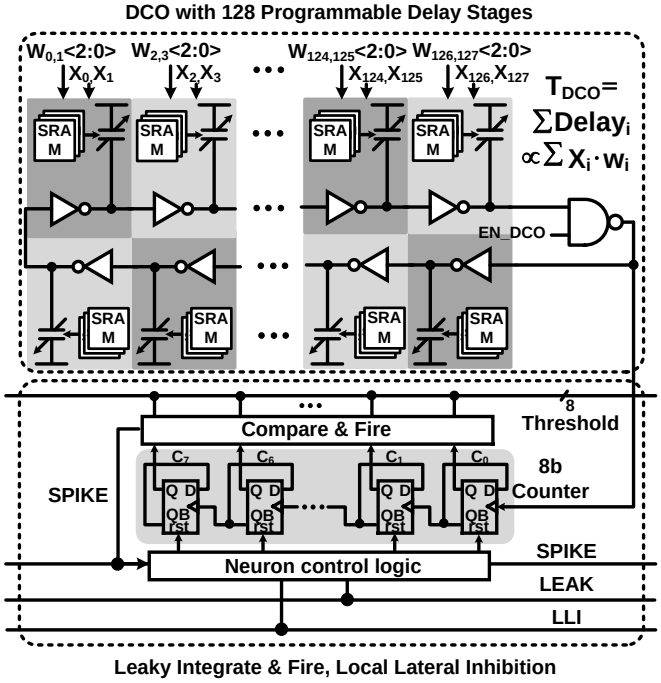


Fig. 1. Conceptual circuit diagram of the proposed time-based integrate & fire (I&F) DCO neuromorphic core.

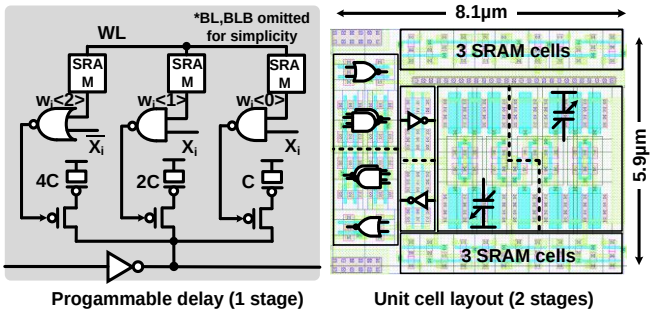


Fig. 2. Detailed implementation of programmable delay stage and unit cell layout.

computes $y = \sum_i x_i \cdot w_i$ purely in the time domain using programmable delay stages, where x_i is the input data and w_i is the synaptic weight. The detailed implementation of programmable delay stages and the unit cell layout of 2 delay stages are shown in Fig. 2. Each input stage of the DCO is

composed of an inverter and binary-weighted MOSFET capacitors controlled by an input pixel and a 3-bit weight, which are stored in SRAM cells. Input pixels determine whether a stage is activated or not, and weights determine how many capacitors are turned on as load in that stage. Weight 100_2 is defined as weight zero. If weights are less than 100_2 (i.e. 001~011_2), fewer load capacitors are turned on, reducing the delay of that stage. These weights represent excitatory synapses. To the contrary, weights greater than 100_2 (i.e. 101~111_2) represent inhibitory synapses. Delay of all stages accumulates naturally in the DCO loop and is converted to an oscillation frequency, which is fed to an 8-bit counter. The counter increments every DCO cycle, and when the counter value reaches a target count, a spike is generated and the counter self-reset. The spike count will be recorded and used as the output of each DCO unit. The counting and thresholding blocks can implement the integrate and fire using simple hardware. The measurement precision of the time based DCO circuit can be easily programmed by changing the spiking threshold. With a higher spiking threshold for instance, a smaller frequency difference can be detected at the cost of higher energy dissipation. The DCO circuit is also very robust against jitter, since the jitter will be averaged out over many DCO cycles.

Fig. 3 shows the overall architecture of the time-based neuromorphic core with 64 parallel DCO circuits. The array is divided into 8 groups, each consisting of 8 DCO circuits, to realize the local lateral inhibition feature (described in next section). Each DCO can be enabled or disabled independently allowing us to activate any number of DCOs simultaneously. The proposed neuromorphic core compares the raw spike count of each DCO to determine which neuron output is dominant. Due to process variation however, different DCOs have slightly different oscillation frequencies for identical inputs. So it is critical that the DCO frequencies are uniform to start with. Unlike process variation, voltage and temperature variation affect all DCOs in the same way, so although the absolute count may vary, the dominant neuron will stay the same under V and T variation. In our design, 7 of the 128 DCO stages are reserved for frequency trimming while the remaining 121 stages are used for the normal neural net function. For frequency calibration, all DCOs are configured using the same pixel inputs and weights, and then the frequency counts are measured for a fixed time period. By tuning the weights of the 7 frequency trimming stages, we can ensure that the baseline DCO frequencies are the same. Measurement results in Fig. 4 (left) show that after calibration, the frequency variation of 10 DCOs reduces from 1.17% to 0.10%. After a one-time calibration, the frequency variation remains low as voltage and temperature shifts affect all DCO transistors in the same way. Fig. 4 (right) shows the mean and 3σ error bars of the frequency count when different number of stages are activated. Measured results confirm good linearity.

III. LEAK AND LOCAL LATERAL INHIBITION

Brain-inspired leak and local lateral inhibition (LLI) features are also implemented in our design, and they can enhance the contrast between neuron outputs. Fig. 5 illustrates the working principle of leak and LLI. When the leak feature is enabled, the LSB of a counter is not only reset by the spikes, but is also reset periodically by a low-frequency LEAK signal. This has the effect of gradually decreasing the stored count value, mimicking

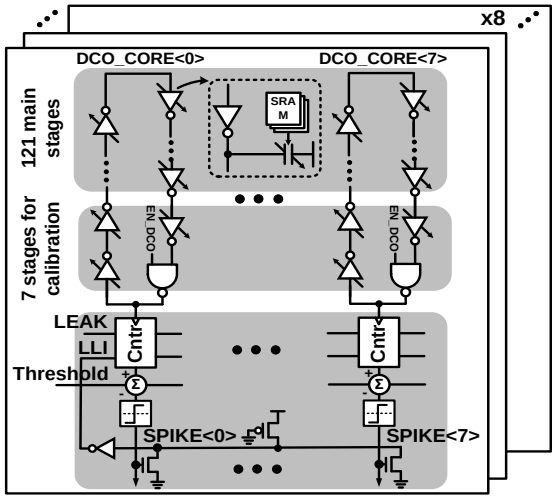


Fig. 3. Time-based DCO neuromorphic architecture.

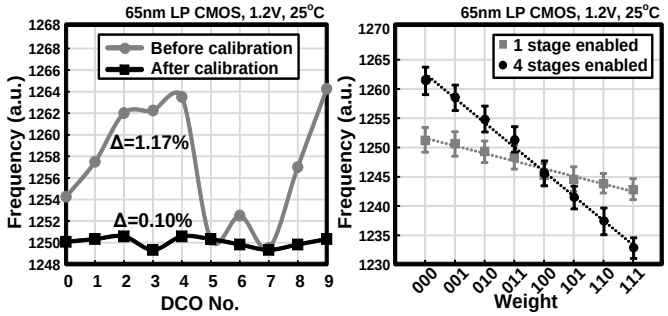


Fig. 4. Measurement results of DCO frequency compensation and linearity test.

a leaky neuron. Note that the period of the LEAK signal need not be very accurate, but must be several times longer than that of the DCO. The main benefit of the leak operation is that it can increase the relative difference between DCO counts as shown in Fig. 6 (left). The frequency difference between the fastest

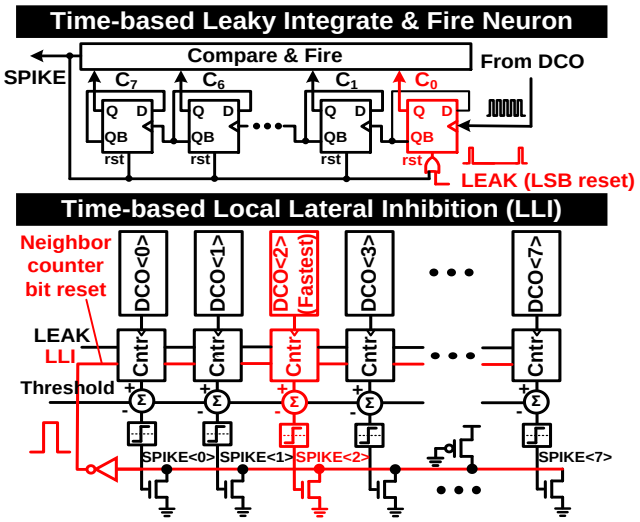


Fig. 5. Illustration of time-based leaky neuron and local lateral inhibition (LLI) operation. The LSB is periodically reset using a low-frequency clock to realize a leaky neuron, while bits in neighboring counters are reset by each DCO to mimic the local lateral inhibition behavior.

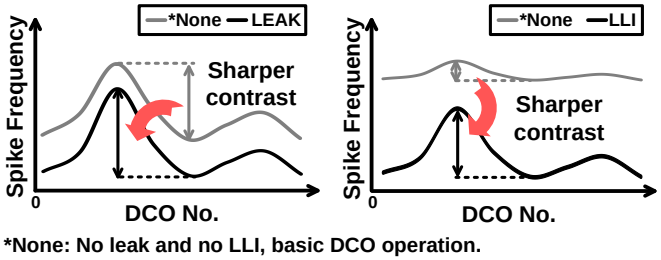


Fig. 6. Effects of leak and local lateral inhibition (LLI) features.

DCO and the other DCOs becomes larger with the leak operation, which enhances the contrast between counts. Lateral inhibition is a phenomenon in which the active neuron strives to suppress the activities of its neighbors. In this design, every 8 DCO cores are grouped together to realize LLI. The inhibition amount (count decrease) is determined by which bits of the neighbor counter are reset. Once a DCO in the group spikes, there is a pulse generated as LLI signal, which resets specific bits in the neighboring counters. The fastest DCO in the group resets the other DCOs more often than it is reset by the other DCOs, enhancing the contrast between different DCO outputs, which is illustrated in Fig. 6 (right).

IV. Digit Recognition Application and Measurement Results

A test chip was fabricated in a 1.2V, 65nm LP CMOS process to demonstrate the time-based neuromorphic core in real hardware. Due to chip size limitations, we opted for a single core implementation. However, a multi-core architecture can be realized to handle deep neural network algorithms by simply tiling several DCO cores and operating them in parallel. Fig. 7 shows a 2-layer test architecture for handwritten digit recognition used to showcase the versatility of the proposed core. Handwritten text images were obtained from the MNIST database [3]. The original image size is 28x28 pixels. We first shrink the image to 22x22 pixels by removing 3 pixels on each side as they contain little information. Each image is then divided into 4 patches so that they can be assigned to different cores for increased throughput and accuracy. The first layer of the neural net can extract 60 distinct features from each patch. The outputs of the 4 patches from the first layer are summed, encoded, and used as the inputs for the second classifier layer. Weights of both layers are trained off-chip using supervised learning and downloaded to the chip. Fig. 8 compares the

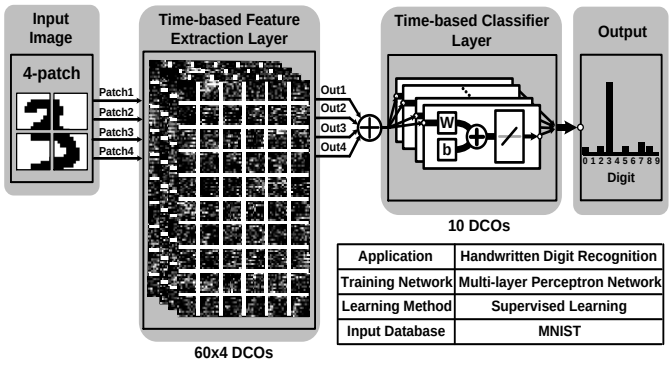


Fig. 7. Multi-layer digit recognition test architecture and summary for time-based neuromorphic chip demonstration.

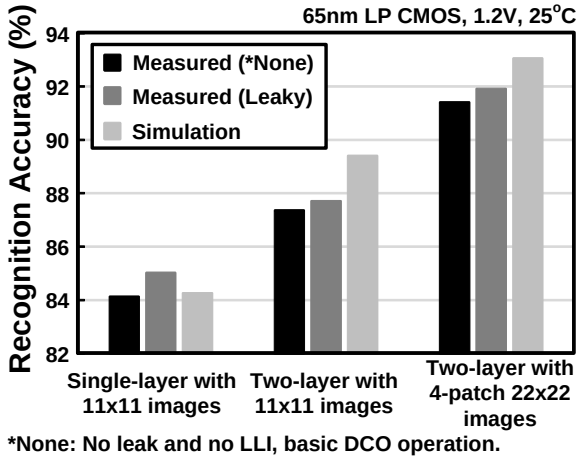


Fig. 8. Measured accuracy results of handwritten digit recognition application.

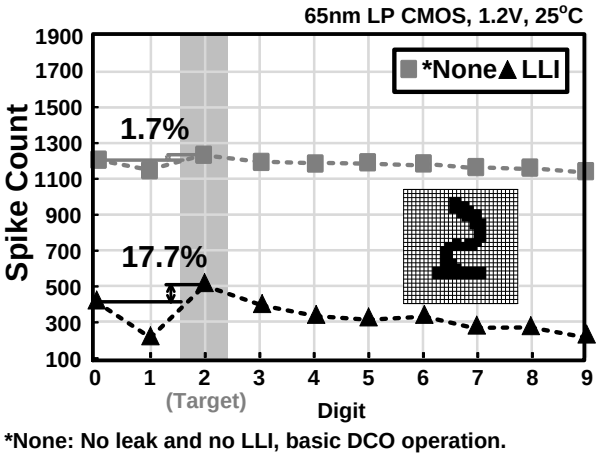


Fig. 9. Measured results of digit recognition application with local lateral inhibition (LLI) feature enabled.

accuracy between different configurations. The 2-layer architecture with 4-patch inputs (=22x22 pixels) achieves a recognition accuracy of 91.4%. With the leak feature enabled, the accuracy increases modestly to 91.9%. The measured accuracy is comparable to software simulation results. As seen from the measurement results in Fig. 8, the recognition accuracy of a single-layer architecture increases from 84.1% to 85.0% after enabling the leak feature, while the accuracy doesn't improve as much in the two-layer architecture. This is because in the two-layer architecture, we have more weights at our disposal to improve the contrast between the neuron outputs. This makes the leak feature less effective. Fig. 9 shows the output spike count with and without the LLI feature for an image containing handwritten digit "2". The spike count difference between digit "2" and runner-up digit "0" is improved from 1.7% to 17.7% using LLI.

Table I shows the performance comparison with recent neuromorphic chip designs [4-8]. It is worth noting that an apples-to-apples comparison between our time-based scheme and traditional ASIC chips can be tricky. Here, we chose to present metrics (e.g. spikes/s/DCO) specific and relevant to our design. The proposed DCO array can generate 3.09×10^{11} / 16 =

TABLE I. PERFORMANCE COMPARISON

	This work	ISSCC'16 [4]	VLSI'16 [5]	ISSCC'15 [6]	ISSCC'14 [7]	CICC'11 [8]
Application	Hand writing recognition	Object detection + intention prediction	Object recognition	Big data analysis	Pattern recognition	Hand writing recognition
Function	Multi-layer perceptron network	Deep neural network	Deep neural network	Deep neural network	Unsupervised online clustering	Restricted Boltzmann Machine
Circuit Type	Time-based	Analog + Digital	Digital	Digital	Analog + Floating gate	Digital
Technology	65nm	65nm	40nm	65nm	0.13µm	45nm
Area	0.24mm^2 (64 DCOs)	16.0mm^2	1.4mm^2	10.0mm^2	0.36mm^2	4.2mm^2
Voltage	1.2V	1.2V	0.9V	1.2V	3V	0.85V
Frequency	99MHz (nominal DCO freq.)	250MHz	240MHz	200MHz	8.3kHz	-
Power	320.4 µW/DCO	330mW	140.9mW	185mW	11.4µW	45pJ/spike
Performance	99M ÷ N spikes/s/DCO (*N=spiking threshold)	502.0GOPS	898.2GOPS	411.3GOPS	0.012GOPS	-
Power Efficiency	309G ÷ N spikes/s/W (*N=spiking threshold)	862GOPS/W	6.37TOPS/W	1.93TOPS/W	1.04TOPS/W	-

*N=16 in our measurements

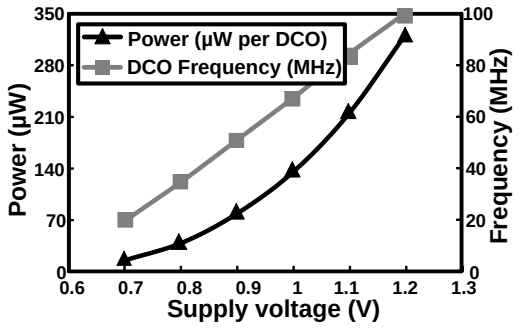
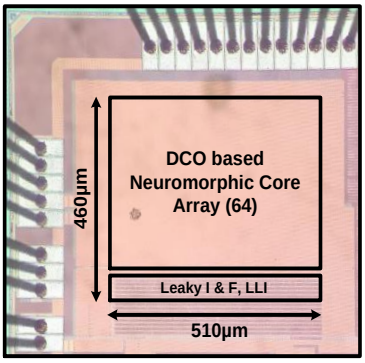


Fig. 10. Measured power consumption and DCO frequency of the test chip.



Technology	65nm LP CMOS
Core Size	510µm x 460µm = 0.24mm^2
VDD range	0.7V~1.2V
# of Neurons	64 (single core)
# of Synapses	8192 (8K)
Throughput	746Mpixels/s ÷ N (1.2V) 148Mpixels/s ÷ N (0.7V) (*N=spiking threshold)
Power	320.4µW (per DCO, 1.2V) 17.5µW (per DCO, 0.7V)

*N = 16 in our measurements

Fig. 11. Test chip micrograph and performance summary.

$1.93x10^{10}$ spikes per second per watt, for a spiking threshold value of 16. Fig. 10 shows the measured power consumption and DCO frequency under different supply voltages. The test chip has a wide operating voltage range of 1.2V to 0.7V. The DCO circuit oscillates at 99MHz consuming 320.4µW under a nominal 1.2V supply. At 0.7V, the DCO oscillates at 20MHz consuming 17.5µW. Fig. 11 shows the chip micrograph and performance summary.

V. CONCLUSION

In this paper, we present the idea of implementing neuromorphic function purely in time domain with programmable delay stages. Brain-inspired leak and local lateral inhibition (LLI) is also presented. The time-based neuromorphic core is tested with digit recognition application and achieves a 91.4% recognition accuracy. The energy-efficiency and versatility of the presented time-based DCO neuromorphic core makes it a promising building block for future large scale deep neural network applications.

ACKNOWLEDGMENT

This research was supported in part by NSF IGERT grant DGE-1069104.

REFERENCES

[1] C. Mead, "Neuromorphic Electronic Systems," *Proc. IEEE*, vol. 78, no. 10, pp. 16229-1636, Oct. 1990.

[2] P. Merolla, *et al*., "A Million Spiking-Neuron Integrated Circuit with a Scalable Communication Network and Interface," *Science*, vol. 345, pp. 668-673, Aug. 2014.

[3] LeCun, et al., "The MNIST Database of Handwritten Digits."(1998).

[4] K. J. Lee, *et al*., "A 502GOPS and 0.984mW Dual-Mode ADAS SoC with RNN-FIS Engine for Intention Prediction in Automotive Black-Box System," *IEEE Int. Solid-State Circuits Conf.* (*ISSCC)*, 2016.

[5] P. Knag, *et al*., "A 1.40mm2 141mW 898GOPS Sparse Neuromorphic Processor in 40nm CMOS:, *VLSI Circuits, 2016 IEEE Symposium on*, pp. 180-181, Jun. 2016.

[6] S. Park, et al., "A 1.93TOPS/W Scalable Deep Learning/Inference Processor with Tetra-Parallel MIMD Architecture for Big-Data Applications," *IEEE Int. Solid-State Circuits Conf.* (*ISSCC)*, 2015.

[7] J. Lu, *et al.*, "A 1TOPS/W Analog Deep Machine-Learning Engine with Floating-Gate Storage in 0.13um CMOS," *IEEE Int. Solid-State Circuits Conf.* (*ISSCC), 20*14.

[8] P. Merolla, et al., " A Digital Neuromorphic Core Using Embedded Crossbar Memory with 45pJ per Spike in 45nm," *IEEE Int. Custom Integrated Circuits Conference (CICC)*, 2011.