

Towards Asynchronously Triggered Spiking Neural Network on FPGA for Event-based Vision

Zhenyu Wu, Mo Song, Hayden Kwok-Hay So

Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong

{zhenyuwu, songmo, hso}@eee.hku.hk

Abstract—In this paper, the design and implementation of an asynchronously-triggered spiking neuron tailored for FPGA implementations is presented. The asynchronous triggering mechanism is designed to effectively preserve the spatial-temporal information captured by event-based vision sensors. Neurons of the target network model are implemented using the configurable logic blocks of an FPGA spatially and communicate through an asynchronous handshaking protocol. Without using a global clock to synchronize operations of the neurons, energy consumption and inference latency are minimized. The proposed design is composable and scalable for applications with different spiking neural network (SNN) configurations. Experimental results show that the proposed design is able to achieve 95.76% accuracy on the N-MNIST dataset while consuming less than 1 mW dynamic power on an Xilinx ZCU102 board.

Index Terms—Spiking neural network, asynchronous circuit, FPGA, dynamic vision sensor, neuromorphic computing

I. INTRODUCTION

Event-based vision sensors are bio-inspired devices that report light intensity changes local to each individual pixel as spiking events [1]. It is in contrast to conventional frame-based sensors that report the light intensity values across the entire sensing array as dense matrix frames at regular intervals. Consequently, the outputs of an event sensor are sparse and are produced asynchronously with spatio-temporal information of the lighting event encoded in the spikes.

To perform vision tasks with the sparse and asynchronous output from event sensors, previous works have predominantly relied on first aggregating the events into various kinds of 2D representations, which were subsequently processed using vision algorithms similar to those originally developed for frame-based input [2]–[5]. Although these pseudo-frame-based algorithms have demonstrated promising results, the high-resolution timing information of the spikes is often lost during the process. Moreover, the complexity of these conventional algorithms has made them less attractive for edge applications.

In this paper, an asynchronously triggered spiking neuron tailored for FPGA implementation is proposed. With that, spiking neural networks (SNNs) can be constructed spatially on FPGAs while the proposed neuron communicates by producing asynchronous spiking events. This asynchronous communication paradigm not only helps to preserve the high-resolution timing information of the spikes from the event

sensor, but it also allows energy-efficient and low latency inference to be performed on-chip. The end-to-end performance of SNN implemented on a Xilinx ZCU102 board was measured using multiple event vision datasets. The results show that the proposed SNN is able to operate with dynamic power well below 1 mW on-chip while resulting in up to 95.76 % accuracy on the N-MNIST dataset.

II. HARDWARE ARCHITECTURE

Figure 1 shows the schematic and the timing diagram of the proposed asynchronous neuron. In the proposed design, state transition is triggered by the rising or falling edge of input signals. The event trigger, denoted by a register symbol with 'T', is used to emit a short pulse at the falling edge of `event_in` if `is_event` is asserted. The delay units in the event trigger are implemented by cascading LUTs on FPGA.

A single-wire signal called `spike` is used to control the asynchronous flow of spikes across neurons. The rising edge of any of the `spike` signal from the n input neurons marks the beginning of the computation phase and `is_active` is raised to high. During this phase, the activation register will be set at the rising edge of each `spike`. Then the membrane potential of this synapse accumulates the weight of each of the connected inputs. At the same time, the membrane potential of the synapse decays exponentially over time according to a predefined rate. At the end of the computation phase, i.e. when the `is_active` returns to zero, the synapse fires if its membrane potential exceeds a threshold. The potential of the synapse is reset to zero after firing. At the circuit level, the firing of the event translates into asserting the `spike_out` signal in the neuron.

The inputs and output of the asynchronous neuron only carry spikes. Therefore, we can leverage the reconfigurability of FPGA to implement different network architectures by connecting the synapses together.

III. EXPERIMENT RESULT & EVALUATION

We have implemented various network architectures for different classification tasks on the Xilinx ZCU102 evaluation board. The network weights were trained using the STBP algorithm. We used three DVS datasets for evaluation: N-MNIST, Poker-DVS, and DVSGesture. During training, we only used events within the first 50,000 timestamps and used event frame transformation to aggregate events over 1,000 timestamps. When testing on hardware, we set the firing rate to

This work was supported in part by the Research Grants Council (RGC) of Hong Kong under the Research Impact Fund project R7003-21. This work was also supported by AI Chip Center for Emerging Smart Systems (ACCESS), sponsored by InnoHK funding, Hong Kong SAR.

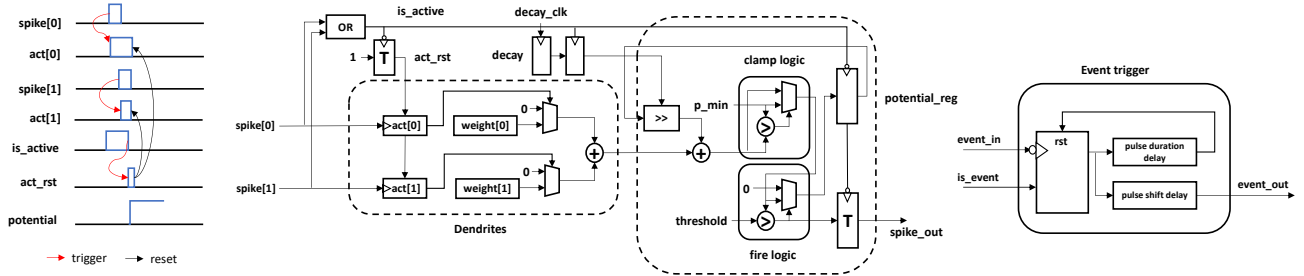


Fig. 1. The circuit and timing diagram of a 2-input asynchronous neuron

TABLE I
COMPARISON WITH PREVIOUS WORK ON FPGA

		Minitaur[6]	Skydiver[7]	[8]	[9]	This work
Platform		Spartan-6 LX150	XC7Z045	XC7VX690T	XC7Z045	ZCU102
Type of tasks		frame-based	frame-based	frame-based	both	event-based
Accuracy		92% @ MNIST	98.5% @ MNIST	98% @ MNIST	82.32% @ N-MNIST	95.76% @ N-MNIST
Frequency		75 MHz	200 MHz	-	250 MHz	-
Resource	LUT	-	45,986	-	10,052	39,709~165,071
	FF	-	20,544	-	8,505	6,499~33,076
	BRAM	200	262	-	131	0
	DSP	64	0	-	32	0
Power	Static	1.1 W	-	0.36 W	-	0.33~0.37 W
	Dynamic	0.4 W	-	0.34 W	-	≤ 1 mW
	Total	1.5 W	0.96 W	0.70 W	0.94 W	0.33~0.37 W

1 MHz, which is the same as the one of the real DVS camera for better emulation. Therefore, the classification latency is 50 ms (calculated as 50,000 / 1 MHz). For power measurement, we used the PL-side INA226 power monitor on the ZCU102 board.

We compared our work with several existing works on FPGA, and the comparison results are shown in Table I. Since we have implemented different network architectures, there are some variations in resource and power consumption. The LUT and FF consumption increases linearly with the growth of the weight precision and number of the input synapses as shown in Figure 2. Due to the asynchronous method we adopted, the circuit activities are reduced to the minimum. Results show our design has good flexibility for different tasks with minimal power consumption.

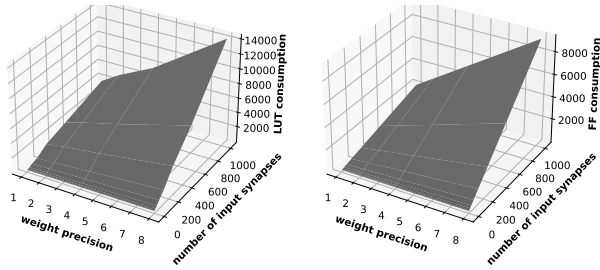


Fig. 2. LUT and FF consumption of a single neuron vs. weight precision and number of the input synapses

REFERENCES

- [1] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Tabá, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, p. 154–180, jan 2022.
- [2] R. Tapiador-Morales, J.-M. Maro, A. Jimenez-Fernandez, G. Jimenez-Moreno, R. Benosman, and A. Linares-Barranco, "Event-based gesture recognition through a hierarchy of time-surfaces for FPGA," *Sensors*, vol. 20, no. 12, p. 3404, 2020.
- [3] A. Sironi, M. Brambilla, N. Bourdis, X. Lagorce, and R. Benosman, "HATS: Histograms of averaged time surfaces for robust event-based object classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [4] Y. Gao, S. Wang, and H. K.-H. So, "REMOT: A hardware-software architecture for attention-guided multi-object tracking with dynamic vision sensors on FPGAs," in *Proceedings of the 2022 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, ser. FPGA '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 158–168.
- [5] —, "A reconfigurable architecture for real-time event-based multi-object tracking," *ACM Trans. Reconfigurable Technol. Syst.*, apr 2023.
- [6] D. Neil and S.-C. Liu, "Minitaur, an event-driven fpga-based spiking network accelerator," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 22, no. 12, pp. 2621–2628, 2014.
- [7] H. Wang, Z. He, T. Wang, J. He, X. Zhou, Y. Wang, L. Liu, N. Wu, M. Tian, and C. Shi, "Triplebrain: A compact neuromorphic hardware core with fast on-chip self-organizing and reinforcement spike-timing dependent plasticity," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 16, no. 4, pp. 636–650, 2022.
- [8] J. Zhang, H. Wu, J. Wei, S. Wei, and H. Chen, "An asynchronous reconfigurable snn accelerator with event-driven time step update," in *2019 IEEE Asian Solid-State Circuits Conference (A-SSCC)*, 2019, pp. 213–216.
- [9] H. Wang, Z. He, T. Wang, J. He, X. Zhou, Y. Wang, L. Liu, N. Wu, M. Tian, and C. Shi, "Triplebrain: A compact neuromorphic hardware core with fast on-chip self-organizing and reinforcement spike-timing dependent plasticity," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 16, no. 4, pp. 636–650, 2022.