Transporter: A 128×4 SPAD Imager with On-chip Encoder for Spiking Neural Network-based Processing

Yang Lin Claudio Bruschini Edoardo Charbon
Advanced Quantum Architecture Lab
École Polytechnique Fédérale de Lausanne
Neuchâtel, Switzerland
{yang.lin, claudio.bruschini, edoardo.charbon}@epfl.ch

***Abstract*—Single-photon avalanche diodes (SPADs) are widely used today in time-resolved imaging applications, however traditional architectures rely on time-to-digital converters (TDCs) and histogram-based processing, leading to significant data transfer and processing challenges. Previous work based on recurrent neural networks has realized histogram-free processing. To further address these limitations, we propose a novel paradigm that eliminates TDCs by integrating in-sensor spike encoders. This approach enables preprocessing of photon arrival events in the sensor while significantly compressing data, reducing complexity, and maintaining real-time edge processing capabilities. A dedicated spike encoder folds multiple laser repetition periods, transforming phase-based spike trains into density-based spike trains optimized for spiking neural network processing and training via backpropagation through time. As a proof of concept, we introduce *Transporter*, a 128×4 SPAD sensor with a per-pixel D flip-flop ring-based spike encoder, designed for intelligent active time-resolved imaging. This work demonstrates a path toward more efficient, neuromorphic SPAD imaging systems with reduced data overhead and enhanced real-time processing.**

I. INTRODUCTION

Single-photon avalanche diodes (SPADs) are highly sensitive photodetectors capable of detecting individual photons with precise timing resolution. They have become widely used in time-resolved imaging applications, such as fluorescence lifetime imaging microscopy (FLIM), light detection and ranging (LiDAR), and time-resolved Raman spectroscopy [1]. Traditional SPAD-based time-resolved imaging systems typically employ time-to-digital converters (TDCs) and histogram-building in the workflow [2], as shown in the top path of Fig. 1. However, the substantial volume of single-photon data generated by these systems presents challenges in data transfer, storage, and processing. To address this problem, artificial neural networks (ANNs) have been proposed as an alternative to histogram-building and/or data processing at the output of the TDC. Recurrent neural networks (RNNs) are deployed on the FPGA to process raw timestamps for LiDAR and FLIM [3], [4].

In this paper, we go one step further introducing spiking neural networks (SNNs) to eliminate TDCs and to simplify data and image processing. SNNs take spikes as input, making them well suited for processing pulses generated by SPADs without TDC processing, despite training challenges [5]. Unlike time-tagging methods, only single-bit data are transferred between the SPAD and the SNN, while the data stream can be further compressed using a dedicated spiking encoder. Herein, we propose a novel paradigm that integrates in- or near-sensor spike encoding with SPAD in time-resolved imaging, enabling end-to-end real-time processing at the edge [6]. The spike encoder folds multiple laser repetition periods temporally, converting phase-based spike trains into density-based spike trains, thus significantly compressing the data while retaining the spiking nature essential for SNN processing. The compressed spike trains also facilitate the training based on backpropagation through time (BPTT). Based on the new paradigm, we introduce *Transporter*, an imager inspired by an episode of *Star Trek: The Next Generation* (S6E4, *Relics*). *Transporter* is a proof-of-concept 128×4 SPAD sensor with per-pixel D flip-flop (DFF) ring-based spike encoder, specifically designed for intelligent active time-resolved imaging.

II. TRANSPORTER

Transporter is composed of a clock module, 128 rows of 1×4 SPAD-Counter-Ring, and a readout module, as shown in Fig 2. The clock module, located on the top, provides and distributes the clock to all the rings for operation and readout. For each row of 1×4 SPAD-Counter-Ring, the 1×4 SPADs are located in the middle, while two passive-quenching and passive-recharging (PQPR) circuits and rings are placed on either side. The rings can be organized into four columns, where a tri-state buffer bus goes through each of them and the data are read out by the readout module at the bottom.

A. Clock Module

The clock module is designed to supply either a high-speed clock to the ring, delivering exactly 128 pulses per laser period, or a low-speed clock for readout. An internal 3-stage NAND-based ring oscillator generates the high-speed clock for ring operation. A separate VDD is supplied to improve the stability of the ring oscillator. Alternatively, an external clock can be selected. The clock is distributed

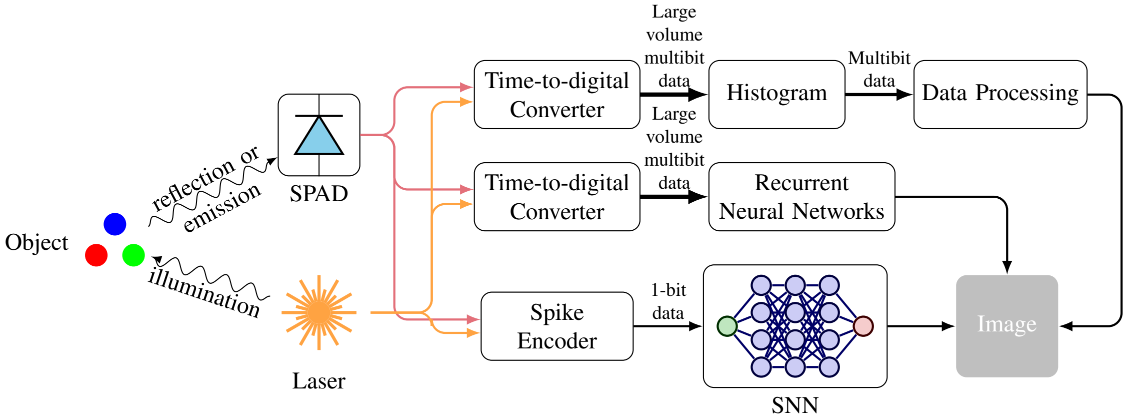
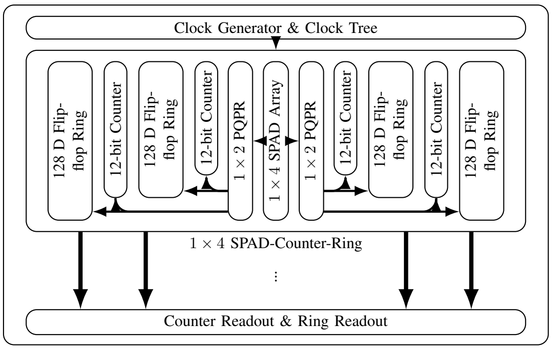


Fig. 1: Workflow of the traditional SPAD TCSPC system, RNN-based system, and the proposed *Transporter* system. In a traditional SPAD TCSPC system, the laser illuminates the object and sends a reference signal to the TDCs. The SPAD detects the reflected or emitted light from the object. The reference and detection signals from the SPAD are directed to TDCs for time-tagging. The timestamps are often histogrammed and transmitted to a PC/Mac for data processing, where the image is reconstructed. For the RNN-based system, the timestamps are directly processed within or near the sensor, then the images are sent to PC/Mac. In the proposed *Transporter* imager, the on-chip spike encoder synchronizes with the laser, and receives pulses upon photon arrival. The encoder keeps collecting photons until the readout is launched, then the pulses are sent to the SNN, which is deployed either on the chip or the FPGA, where the images are reconstructed.



(a) High-level schematic of the *Transporter* sensor

Fig. 2: *Transporter* is composed of a clock module, 128 rows of 1×4 SPAD-Counter-Ring, and a readout module.

to the four columns of rings with a clock buffer H-tree, and further goes to the 128 rings.

Ensuring exactly 128 pulses per laser period is crucial for maintaining data integrity within the ring. Any deviation, whether an excess or shortage of pulses, would result in the loss of relative timing information between spikes. To address this, a gating mechanism is implemented to enforce the 128-pulse constraint. This mechanism utilizes a 7-bit counter, composed of a 2-bit synchronous counter and a 5-bit asynchronous counter, designed to operate at gigahertz frequencies. Once the counter reaches 128, the clock is disabled and remains so until the next laser synchronization signal arrives, at which point it is re-enabled.

B. SPAD and Pixel Circuits

The SPAD features $\phi 10$ μm round active area. A PQPR circuit, composed of two NMOS transistors, is employed for each SPAD. The photon detection probability and the dead time can be tuned by adjusting the gate voltage of these transistors. Two inverters are cascaded at the output to digitize the pulse. The pixel pitch is restricted by the height of the *Transporter* ring, which is 25 μm in our implementation. The horizontal pitch is set equal to the vertical pitch to maintain isotropic pixel spacing, though it can be reduced to 15 μm. As a result, the fill factor is 12.6%.

C. Transporter Ring

The *Transporter* ring functions as the spike encoder as shown in Fig. 1. In this work, we adopt DFFs as the delay element described in [6]. A simplified schematic is illustrated in Fig. 4. 128 DFFs are cascaded like a circular shift register, except that an OR gate is inserted between the two ends, with its second input connected to the SPAD output for spike injection. The DFFs are arranged into two rows, with clock buffer H-tree placed on the top and bottom. As mentioned earlier, the accurate clock timing for the DFFs is crucial to the integrity of the spikes. Considering the setup and hold time at four corners, the maximum operation frequency is 1GHz.

In this proof-of-concept *Transporter*, only 128 DFFs are used. As these DFFs store single-bit data, the ring could be saturated in case of high photon flux. Therefore, a ring stopper, based on the ripple counter and the clock gate, is designed to prevent the ring from saturation. With the stopper, the injection of the spike into the ring will be stopped when the photon count rises above 128 or 256.

III. Simulation Results

The chip will still be in the manufacturing process when this paper is published and will be tested in the near future. In the following subsections, the post-layout simulation results of the chip are shown, as well as the simulation of the SNN on the PC.

A. Clock Module

In the post-layout simulation, the ring oscillator generates a clock at 1.068 GHz (ss: 0.77 GHz, ff: 1.346 GHz). The simulation result of the gated clock is shown in Fig. 5a. One can observe the clock is disabled when the pulse count reaches 128 in every laser period (100 ns).

B. SPAD and Pixel Circuits

Given the operation voltage $V_{OP} = 25\text{V}$, breakdown voltage $V_B = 20\text{V}$, cascode voltage $V_{cascode} = 3.3\text{V}$, the quenching voltage V_q is swept from 100 mV to 1 V with 10 equal steps. The result is shown in Fig. 5b. By adjusting V_q, the dead time can be tuned from a few nanoseconds to hundreds of nanoseconds.

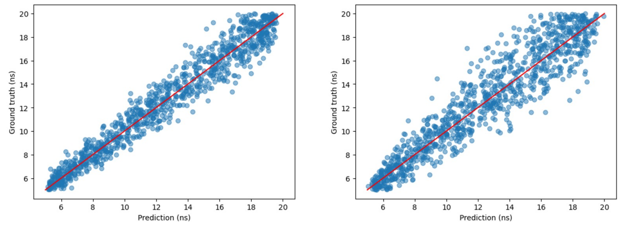
C. Transporter Ring

The simulation result of the test structure for *Transporter* ring is shown here. The collection of photons takes 2 μs, given a ring clock of 1 GHz. The readout of the spikes takes 1 μs, given a readout clock of 200 MHz. The laser period is 128 ns. With a simulated photon arrival with a period of 130 ns, the ring is supposed to receive the photons every two stages. The result is shown in Fig.5c. During the accumulation period, 13 spikes are inserted into the ring, which are read out through the tri-state buffer bus. One can observe that the relative timing information is well preserved within the ring.

D. Spiking Neural Networks for FLIM

Based on the specifications of *Transporter*, a spiking neural network (SNN) is designed, trained, and evaluated for fluorescence lifetime estimation. The SNN consists of a single-node input layer, a 512-neuron hidden layer, and a single-node output layer. Training is performed using backpropagation through time (BPTT) with surrogate gradients.

For the training and evaluation dataset, it is assumed that lifetime ranges from 5 to 20 ns. For each data entry, 256 timestamps are histogrammed and binarized as input for the SNN. Scenarios with no background noise and 10% noise are considered. The result is shown in Fig. 3. The mean average percentage errors are 4.82% and 5.29%, respectively.



(a) No background noise (b) 10% background noise

Fig. 3: Simulation results of spiking neural networks for fluorescence lifetime estimation. The parameters correspond to the ones in the *Transporter*, i.e., timestep of 128 and clock period of 1 ns. The lifetime ranges from 5 to 20 ns. 256 photons are collected for each inference.

IV. Conclusion and Outlook

In this work, we demonstrate the proof-of-concept *Transporter* sensor that integrates a spike encoder for active time-resolved SPAD imaging, which explores a potential approach for realizing neuromorphic SPAD imaging systems. The suitability of *Transporter* can be further enhanced by increasing the clock frequency, extending the encoder ring with more units, and utilizing simpler devices such as dynamic random access memory (DRAM). Our vision is to develop an extended encoder ring operating at a higher clock frequency, enabling real-time, TDC-free, and energy-efficient processing for advanced applications.

Acknowledgment

The authors acknowledge the financial support provided by US Department of Energy.

References

[1] Claudio Bruschini, Harald Homulle, Ivan Michel Antolovic, Samuel Burri, and Edoardo Charbon. Single-photon avalanche diode imagers in biophotonics: review and outlook. *Light: Science & Applications*, 8(1):87, 2019.

[2] Peter Kapusta, Michael Wahl, and Rainer Erdmann. Advanced photon counting. *Springer Series on Fluorescence*, 15, 2015.

[3] Tommaso Milanese, Jiuxuan Zhao, Brent Hearn, and Edoardo Charbon. Histogram-less direct time-of-flight imaging based on a machine learning processor on FPGA. In *Proceedings of the International Image Sensor Workshop (IISW)*, 2023.

[4] Yang Lin, Paul Mos, Andrei Ardelean, Claudio Bruschini, and Edoardo Charbon. Coupling a recurrent neural network to SPAD TCSPC systems for real-time fluorescence lifetime imaging. *Scientific Reports*, 14(1):3286, 2024.

[5] Jason K Eshraghian, Max Ward, Emre O Neftci, Xinxin Wang, Gregor Lenz, Girish Dwivedi, Mohammed Bennamoun, Doo Seok Jeong, and Wei D Lu. Training spiking neural networks using lessons from deep learning. *Proceedings of the IEEE*, 111(9):1016–1054, 2023.

[6] Yang Lin and Edoardo Charbon. Spiking neural networks for active time-resolved SPAD imaging. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8147–8156, 2024.

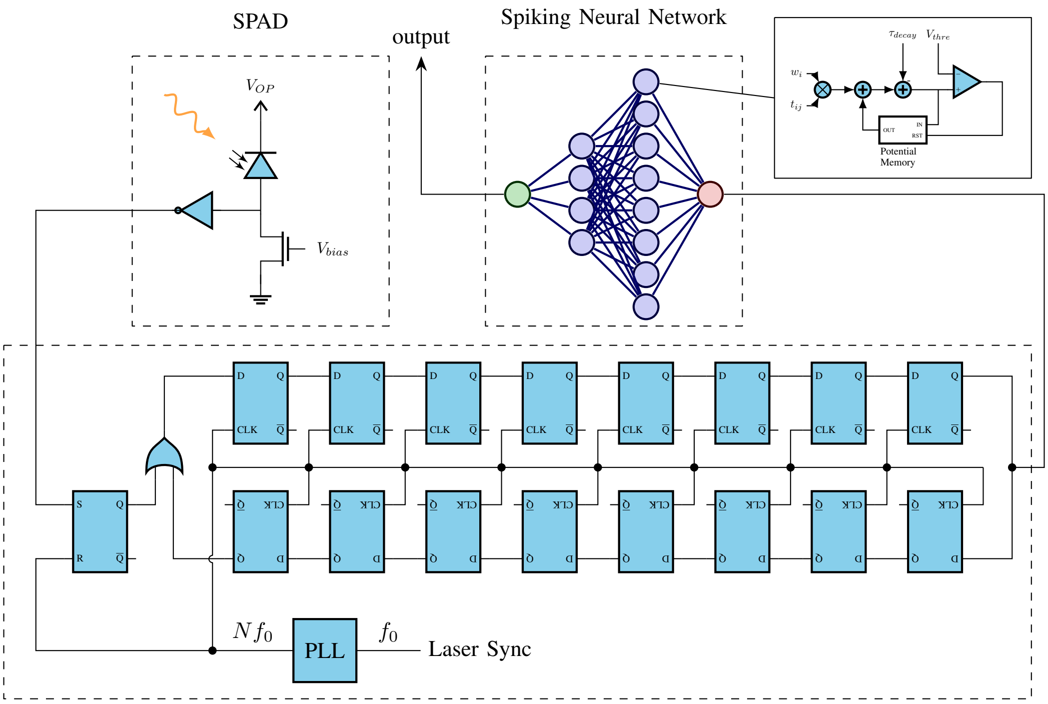
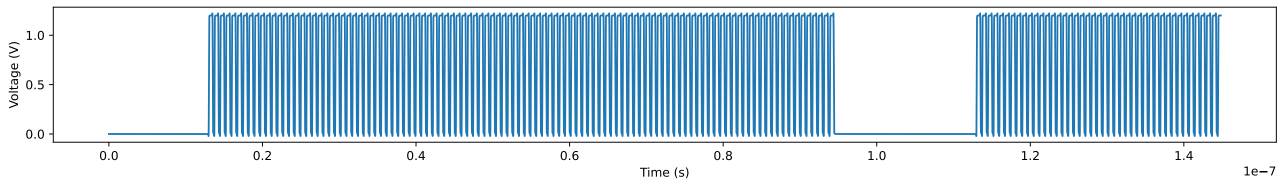
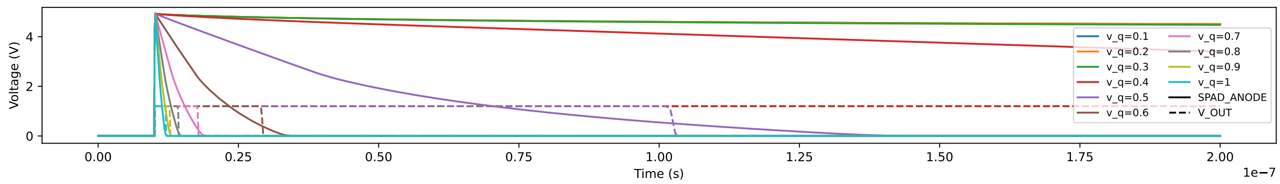


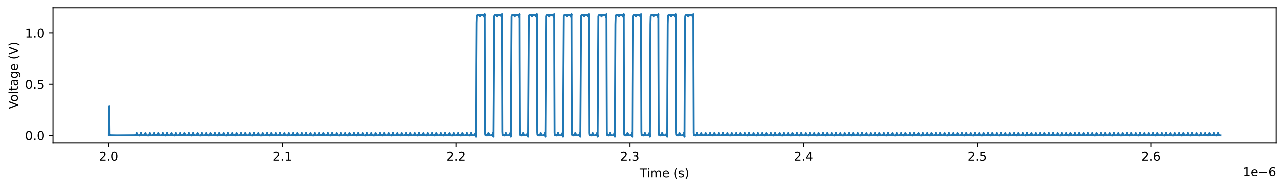
Fig. 4: The proposed DFF ring-based *Transporter* system. A SPAD with PQPR is adopted for photon detection. The pulse is injected into the ring, composed of N DFFs (only 16 DFFs are plotted for simplicity), with an OR gate and keeps circulating. Ideally, the synchronization signal from the laser is divided by N by a phase-locked loop (PLL) and drives all the DFFs. After the exposure, the clock is replaced by a low-speed readout clock and the pulses are read from the DFFs in serial.



(a) The output of the clock module with the internal ring oscillator and gating.



(b) The anode and output of a SPAD with different V_q. The solid lines represent the voltage at the anode of the SPAD, and the dash lines represent the voltage digitized by inverters. Different colors represent different V_q.



(c) Readout of the spikes in the ring. The readout clock is 200MHz, so it takes 6.4 μs to read out 128 DFFs. There are 13 spikes injected into the ring every two delay elements.

Fig. 5: Post-layout simulation results