# Hybrid digital-neuromorphic architecture integration for low-power applications

Michelangelo Barocci
*Politecnico di Torino*
Turin, Italy
michelangelo.barocci@polito.it

***Abstract*—Innovative computing paradigms based on machine learning are revolutionizing the world of IoT, where devices that were once seen as end-nodes connected to the internet to acquire and transmit data now are required to also perform computation. This poses new challenges for developers, since edge devices need to satisfy strict requirements in terms of energy efficiency and limited computational resources. Spiking Neural Networks (SNN), thanks to their capability of emulating the data processing of biological neurons, are promising candidates for low-power edge applications. In this work-in-progress paper we would like to explore the co-existence of a traditional processor with a SNN accelerator through the SPI protocol used for configuring the SNN. Respectively PULPissimo, a RISC-V-based microcontroller, and ReckOn, an open source neuromorphic processor. Later on, we present our next steps towards integrating ReckOn as a co-processor inside PULPissimo.**

***Index Terms*—Neuromorphic computing, Neuromorphic hardware, IoT, RISC-V**

## I. Introduction

Neuromorphic computing was initially developed as a way to accurately simulate the human brain in order to accelerate neuroscientific research, but the way neuromorphic architectures model neurons and synaptic interconnections to elaborate sparse data and make accurate predictions moved the focus towards creating machine learning models that are able to perform tasks such as pattern recognition, keyword spotting and constraint satisfaction problems. The capability of neuromorphic systems to implement Spiking Neural Networks (SNN) paved the way to the development of event-driven sensors, that are capable of transducing input data into spiking signals that could be received by neuromorphic architectures. The direct integration of event-driven sensors and architectures makes them the perfect candidates for on-the-edge applications, where power budgets are strongly limited.
In order to take full advantage of the neuromorphic processors' potential, an external processing unit may still be needed, to either save the network configurations, control the operations or manage the data transfers. Ideally such computing units are accessible via the cloud, but not always an internet connection is available, or the continuous data exchange may create a bottleneck, thus limiting the neuromorphic hardware real-time operations. A solution can be identified in the integration of traditional processors with neuromorphic co-processors, taking advantage of high speed memory units for quickly accessing the network's weights, or to store inference data without the need to access an internet connection. In this work in progress, we propose a preliminary interconnection between two open source architectures: a RISC-V single core processor, PULPissimo and a neuromorphic processor, ReckOn.
The two open source models are written in `Verilog` and `SystemVerilog`, HDL languages mainly aimed at describing hardware for FPGA synthesis and programming.

PULPissimo is a low-power, single-core digital microcontroller based on the RISC-V ISA, developed by ETH of Zurich and University of Bologna [1], released in 2018. It has a energy efficiency of $433\,\mathrm{MOPS/mW}$ and is capable of implementing a 32 bit core with a 4-stage or 2-stage pipeline.
The I/O communications are handled by a $\mu$DMA that ensures the correct data exchange between the tightly coupled L2 embedded memory and the peripherals (I2C, I2S, SPI, UART and others) in an autonomous way [2], thanks to the dedicated portion of memory in the L2 bank. The presence of the SPI module makes PULPissimo a great candidate for interfacing with ReckOn.
In addition, PULPissimo is provided with a Hardware Processing engines (HWPE) interface. HWPEs are architectures aimed at accelerating specific tasks by providing dedicated hardware. The advantages of using HWPEs lie in the tight coupling of the engines with the memory and the specific protocols that can be used for data transfer to and from the memory, for controlling the engine, or to move data efficiently inside the datapath. In this specific field of application, making use of the HWPE support in PULPissimo could allow the neuromorphic co-processor to access the data autonomously by avoiding long programming times set by the SPI.

ReckOn is a spiking recurrent neural network simulator capable of online learning in seconds-long timescales and low power consumption, released by Frenkel et al. [3] in 2022. The network can be implemented with up to 256 input/recurrent LIF neurons and 16 output Leaky-and-Integrate neurons. This allows ReckOn to be used for both classification and regression tasks.
What makes ReckOn optimal for embedded applications is: 1) The optimized weights access that takes advantage of the sparse nature of spiking data, thus minimizing the power consumption (less than $200\,\mu\mathrm{W}$ for learning and inference) and the number of SRAM data readings. 2) The external SPI interface that is used to configure the spiking RNN parameters and to read/write from the SRAM blocks, which makes it compatible with most digital architectures. The input spiking

data can be sent to ReckOn with a 4-phase handshake Address Event Representation (AER) 8 bit bus.

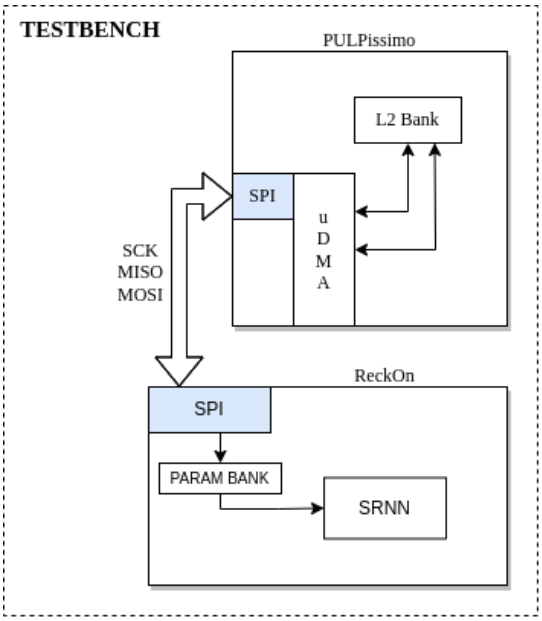


Fig. 1. Schematic representation of the PULPissimo - ReckOn integration: the microcontroller directly accesses the parameter bank to program and configure the spiking RNN inside ReckOn.

## II. Current Status

As shown in Fig. 1, the integration of ReckOn with the PULPissimo processor is made through the SPI port, which is controlled by the autonomous $\mu$DMA subsystem. The SPI transactions for communicating with ReckOn follow a 32 bit command - 32 bit data scheme, where the command contains the R/W request, the target (a specific SRAM, or the parameter bank) and the address (a specific neuron, weight or parameter). The 32 bit data packets to be transferred are fetched sequentially through the $\mu$DMA TX channel directly from the L2 memory and sent to ReckOn to write on the parameters bank registers. PULPissimo acts as SPI master and drives the MOSI and SCK pins, while ReckOn, the slave, drives the MISO channel when requested (although this is not the case).

The programming of PULPissimo can be made by writing the correct $\mu$DMA registers: each transaction on the two main channels, TX and RX, should be configured by providing the L2 address of the data buffer, its length in bytes and some specific commands such as the number of bits to be transferred or the enable signal. These informations must be written on specific registers, respectively `SPIM_TX_SADDR`, `SPIM_TX_SIZE` and `SPIM_TX_CFG` for the TX channel, and similarly to the RX one. The configuration of the CMD channel is also required on the specific registers to provide the $\mu$DMA commands to start the transactions.

All the TX, RX, CMD registers are accessible through the provided libraries `udma.h` and `udma_spim.h`, which include the necessary functions to perform the desired transfers' configurations.

Once the C code has been compiled, the verification of the correct configuration of ReckOn can be made through a QuestaSim hardware simulation.

## III. Thesis Completion

The integration between traditional computing elements and neuromorphic co-processors looks promising. Further steps will involve running tests with the structure proposed in this work and the exploration of more advanced and performant ways to integrate the two architectures.

After the efficacy of the SPI is proven, a more efficient solution will be implemented, where the neuromorphic co-processor will be tightly coupled with the L2 memory inside PULPissimo, exploiting its HWPE interface capabilities. This solution will avoid the need to program the spiking RNN of ReckOn before each use, and will allow the processor to command it directly. In order to make this possible it will be necessary to operate directly on the source code of ReckOn, making it compatible with the HWPE template and the various interface protocols, like in Figure 2.

Other developments will include the integration of other different open source neuromorphic architectures with PULPissimo, in order to benchmark different solutions and individuate the most appropriate fields of applicability.

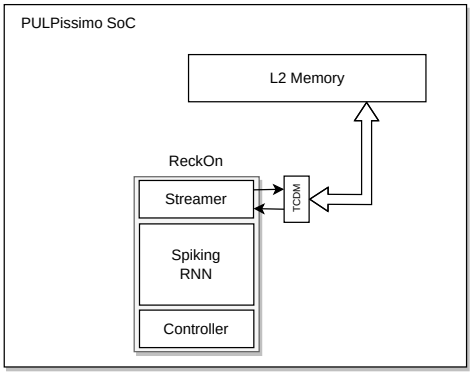


Fig. 2. A possible way in which a neuromorphic co-processor like ReckOn could be integrated inside PULPissimo through the HWPE interface, where the data and spiking RNN parameters are fetched through the Tightly-Coupled Data Memory (TCDM) interconnect.

## References


[1] Pasquale Davide Schiavone et al. "Quentin: an Ultra-Low-Power PULPissimo SoC in 22nm FDX". In: *2018 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*. 2018, pp. 1–3. DOI: 10.1109/S3S.2018.8640145.

[2] Antonio Pullini et al. "$\mu$DMA: An autonomous I/O subsystem for IoT end-nodes". In: *2017 27th International Symposium on Power and Timing Modeling, Optimization and Simulation (PATMOS)*. 2017, pp. 1–8. DOI: 10.1109/PATMOS.2017.8106971.

[3] Charlotte Frenkel and Giacomo Indiveri. "ReckOn: A 28nm Sub-mm2 Task-Agnostic Spiking Recurrent Neural Network Processor Enabling On-Chip Learning over Second-Long Timescales". In: *2022 IEEE International Solid- State Circuits Conference (ISSCC)*. Vol. 65. 2022, pp. 1–3. DOI: 10.1109/ISSCC42614.2022.9731734.