

Feasibility Study for a Dense Analog Asynchronous Spiking Neural Network in a 22 nm Process with On-chip Synaptic Time Dependent Plasticity Unsupervised Learning

E. Fagnière •

An analog integrated implementation of a spiking neural network (SNN) in a 22 nm process is proposed and designed at transistor level. Simulation of a behavioral asynchronous VHDL model shows potentially promising dynamics while first layout attempts on successfully simulated transistor-level schematics suggests that dense SNN on a single chip (~10k synapses/mm²) is feasible and could lead to adaptive neural filtering at sensor level.

The human brain consumes an average of around 20 Watts, whereas today's AI system consumes tens of thousands of Watts to perform complex, but highly specific tasks. The *neuromorphic* approach, that takes more precise inspiration from the way the brain works, has occupied researchers for several decades, but interest has waned over the last 20 years. The main reason being the lightning progress made by AI thanks to the exponential growth in the performance of available hardware and the explosion in digitized data collected by companies exploiting it.

To revisit such concepts, an analog integration of a spiking neural network (SNN) in an advanced 22 nm process is investigated. It could ideally complement a companion bio-inspired low-power integrated analog cochlear filter [1] to implement speech recognition or be used for any other time signal preprocessing.

Besides its feasibility, the aim of the project is to explore the applicability of such an SNN for unsupervised learning. A recurrent architecture suited for time series such as speech signal is targeted. The main advantage of inter-neuron communication via asynchronous short logical pulses, the *spikes*, lies in its much lower power consumption than the one of the classical Artificial Neural Networks (ANN). Storing the synaptic weights of the neurons as analog charges on capacitors is the most challenging aspect of the approach, due to their unavoidable leakage. To alleviate it, each synaptic weight is continuously adapted by a local mechanism called *Synaptic Time Dependent Plasticity (STDP)*, also found in biological neurons. It is expected to perform an unsupervised learning avoiding the need of the biologically unplausible and very power-hungry back-propagation (BP) algorithm fueling supervised learning in today's ANN.

Figure 1 shows the architecture of an L-layer SNN, each k of them made of N_k fully recurrent neurons. Each neuron j has $N_k + N_{k-1}$ synapses and STDP blocks. A synapse is made of a single MOS transistor working in the weak to moderate inversion regime as a current source, whose value I_{ij} , corresponding to the synaptic weight w_{ij} , is controlled exponentially by the charge stored on its gate capacitor C_w . Each time t_{sp} a spike of duration T_{sp} occurs at input x_i of neuron j, it connects the current source I_{ij} to a common node, the *dendritic tree*, on which weighted spikes $q_{ij} = I_{ij}T_{sp}$ sum spatially. Against a controlled *Leakage* I_{Lj} , they also temporally accumulate (*Integrate*) on the neuron's capacitor C_j until its potential p_j reaches a given threshold θ_j at which it resets while the neuron *Fires* a spike on its output y_j (*LIF* neuron model). To each synapse is attached a STDP block, which increases or decreases the charge on synapse w_{ij} 's C_w by sourcing or sinking

a current pulse whose value decreases exponentially with the time interval between spikes occurring on x_i (*pre-synaptic*) and on y_j (*post-synaptic*) whether they are causal or anti-causal.

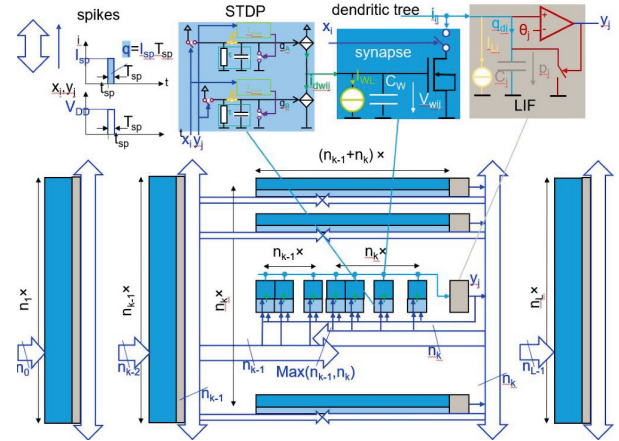


Figure 1: Proposed architecture and blocks of multi-layer recurrent SNN.

Before designing these cells at transistor level, their behavioral and asynchronous VHDL model was developed with which an arbitrary sized SNN can be built and simulated. Figure 2 illustrates the results for a 5-input, 4-layers (5-10-20-4) SNN.

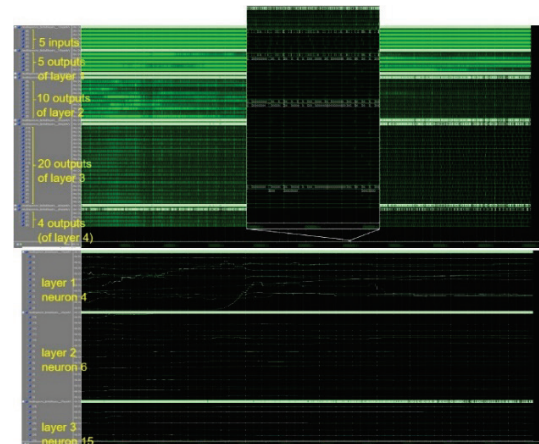


Figure 2: Modelsim simulation result of behavioral VHDL model. Top-plots are spike raster, bottom plots show synaptic weight time evolution.

Transistor-level design in GF22FDX was validated by simulations and a first layout attempt indicates a synapse-STDP block size of about 100 μm^2 , consuming a mere 100 nW of power.

This study was carried out by the author as visiting professor from the University of Applied Science of Fribourg.

• University of Applied Science of Western Switzerland (HES-SO), School of Engineering and Architecture of Fribourg (HEIA-FR)

[1] E. Fagnière, A 100 channel analog CMOS auditory filter bank for speech recognition, ISSCC 2005