University of BRISTOL

Building Brains with ARM Processors and FPGAs

By Felipe Galindo Sanchez

Supervisor: Dr. J L Nunez-Yanez

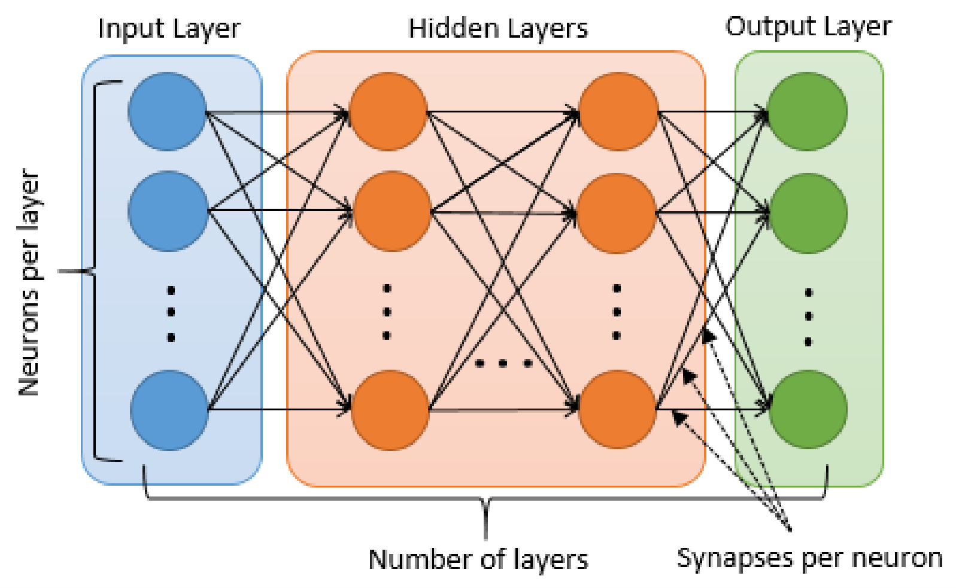
Introduction

- ✓ **High-performance** simulations are required for understanding **biological systems** with neural networks of billions of neurons such as the **human cerebral cortex**.
- ✓ **FPGA** and **GPU**-based implementations of **spiking neural networks (SNN)** have become attractive [1] for addressing **high-performance** and **low-power requirements.**
- ✓ **Deep learning** techniques [2] have become appealing for applications with large amounts of data processing such as **speech recognition**, **computer vision**, **drug discovery**, among others.

Objectives

- ✓ Research **neuron models** and **network topologies:** computational complexity, power consumption and biological plausibility.
- ✓ Apply **high-level synthesis (HLS)** optimizations to the proposed **spiking neural network (SNN)** using Vivado tools.
- ✓ Analyze the overall **performance** and **power consumption** of the solutions after implementing the different simulation **architectures**.
- ✓ Evaluate **deep learning algorithms** for the resolution of specific tasks.

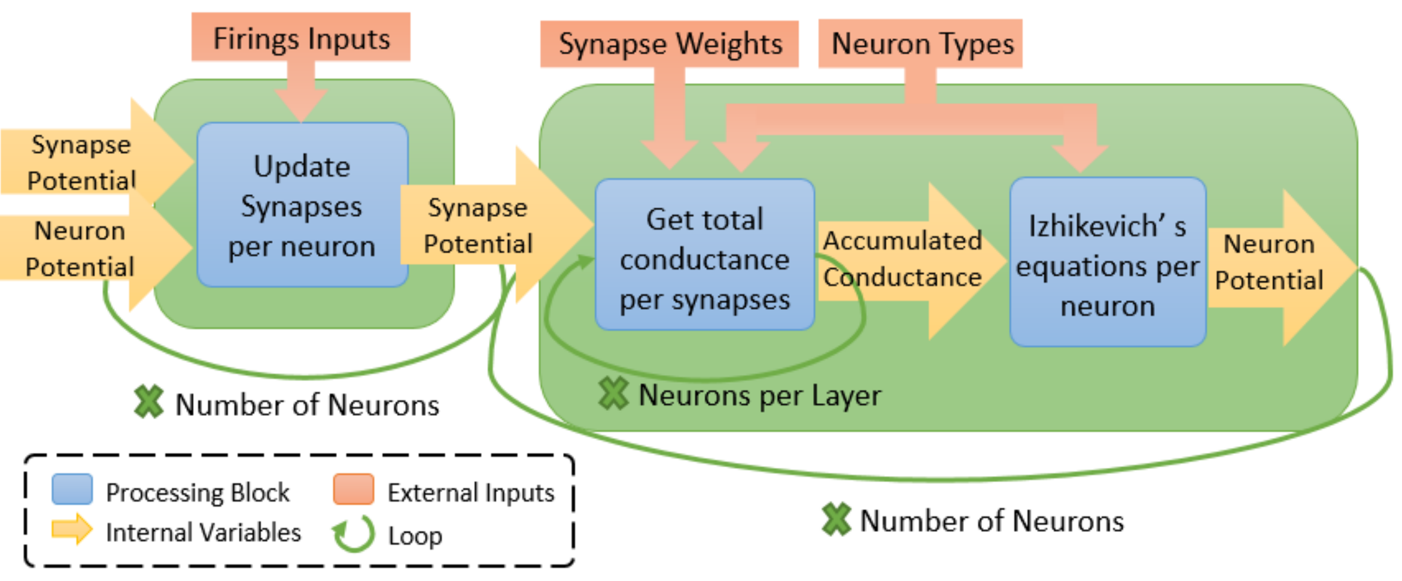
Implementation

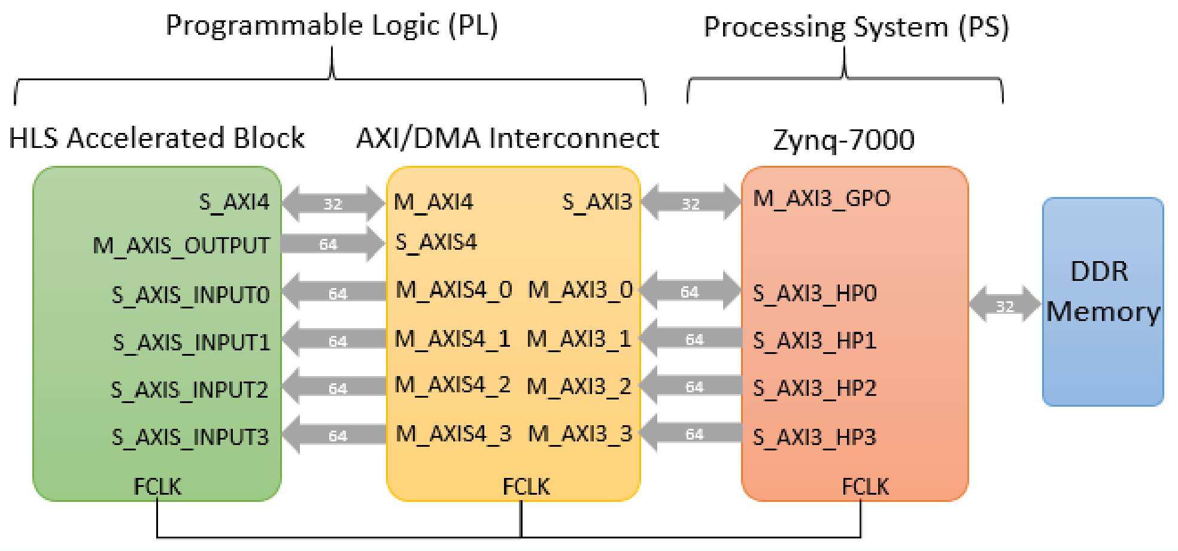


- ✓ **Feed-forward** network based on neuron model described by **Izhikevich** [3].
- ✓ For a N x N network, $\boldsymbol{N^2}$ **neurons** and $\boldsymbol{N^3}$ **synapses** are being simulated.
- ✓ **Models** are simulated with a step time of $\mathbf{0.5}\,\boldsymbol{ms}$**,** while the **overall block is 1** $\boldsymbol{ms}$**.**

- ✓ **Critical latency** is defined by the input throughput of the $\boldsymbol{N^3}$ **synapses**.
- ✓ **256 synapse-bits** are **processed parallel** per cycle.
- ✓ The **hardware block latency** is defined as:

$$n_{total}^2 \cdot \left(1 + roundup\left(\frac{n_{layer} \cdot w_{bits}}{256}\right)\right)$$

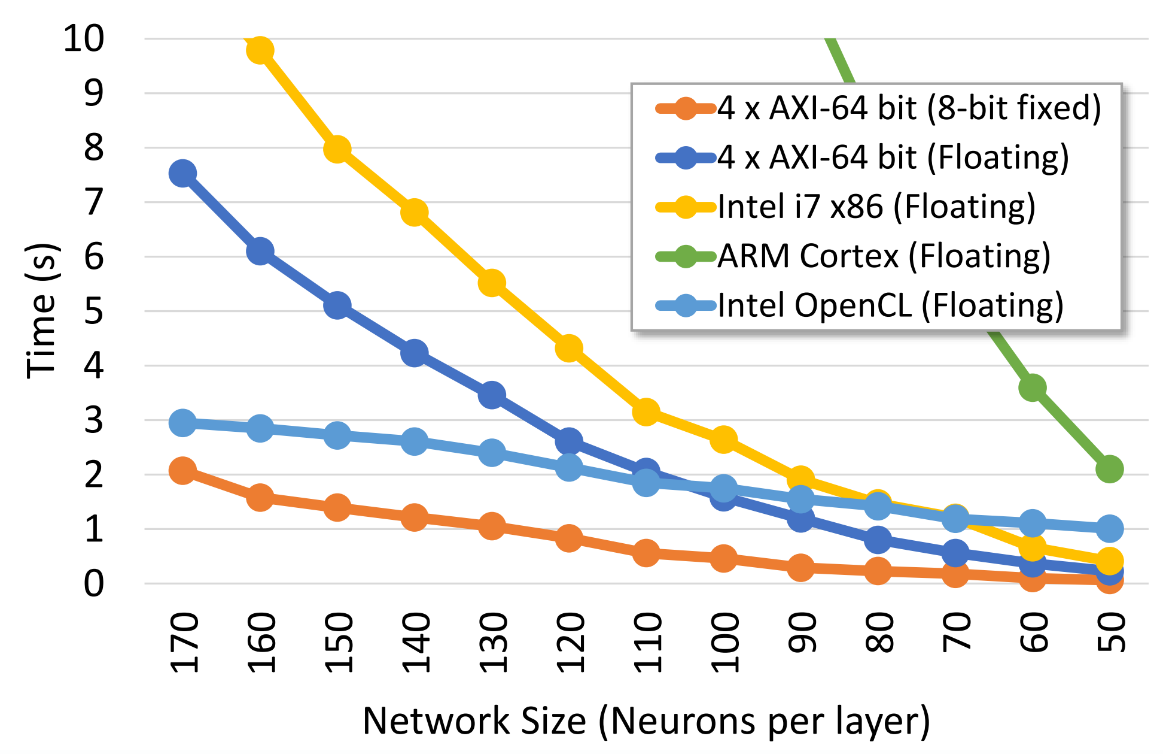




- ✓ **Data** flows though **4 x AXI ports** of **64-bits** and **1 x AXI lite** of **32-bit** for **control.**
- ✓ Overall **system** is evaluated in a **Zynq-7000 SoC** [4] containing a dual core **ARM Cortex-A9** and an **Artix-7 FPGA.**

Results & Analysis

Performance Latency



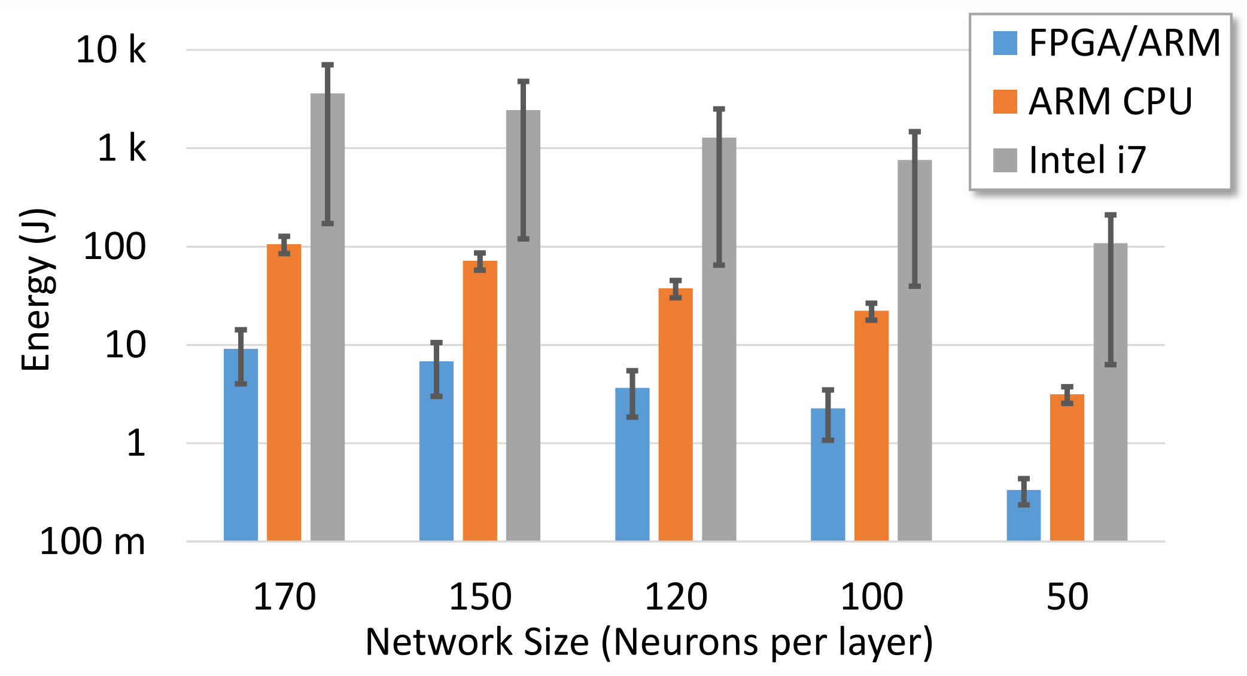
FPGA accelerated version using **4 x AXI** ports of 64-bit is faster than a traditional solution (e.g. Intel i77 x86):

- ✓ **6 times faster** using a **8-bit fixed** point precision.
- ✓ $\mathbf{1.7}$ **times faster** using a **floating** point precision.

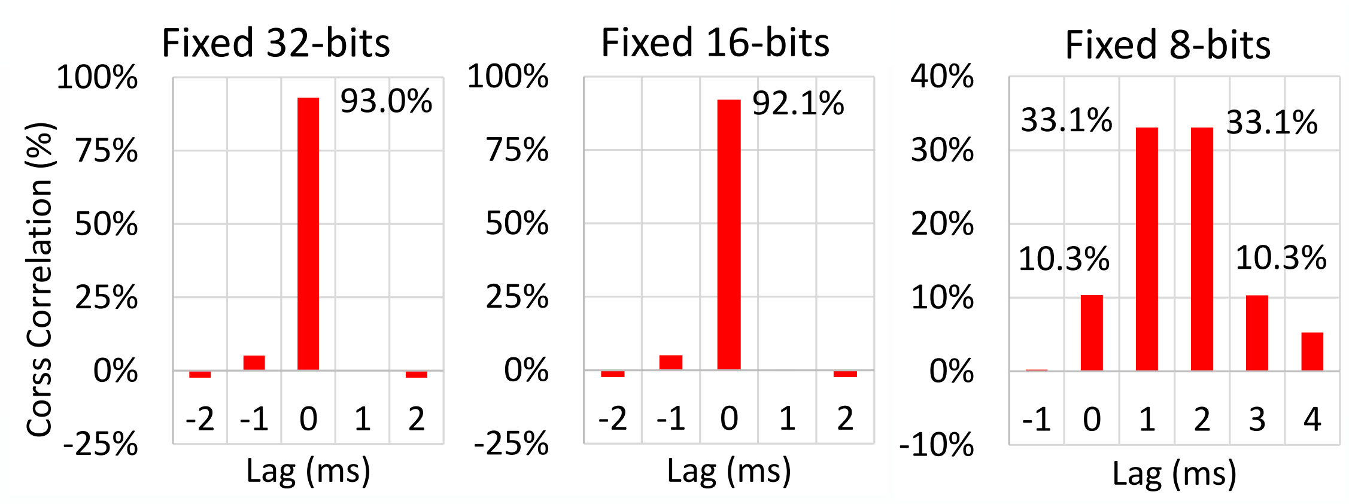
Energy Consumption

Based on the energy required by a traditional solution (e.g. Intel i7 mobile processor), the FPGA/ARM accelerated implementation requires only:

- ✓ From $\mathbf{0.1}\%$ **to** $\mathbf{2.3}\%$ using a **8-bit fixed** point precision.
- ✓ From $\mathbf{0.2}\%$ **to** $\mathbf{8.4}\%$ using a **floating** point precision.



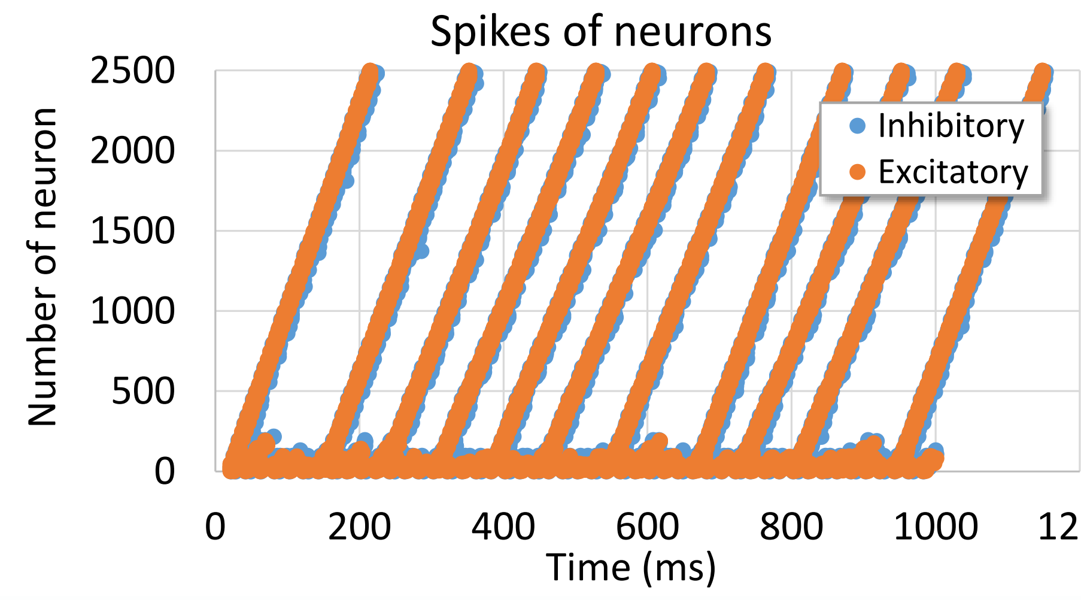
Spike timing cross-correlation of floating and fixed precision

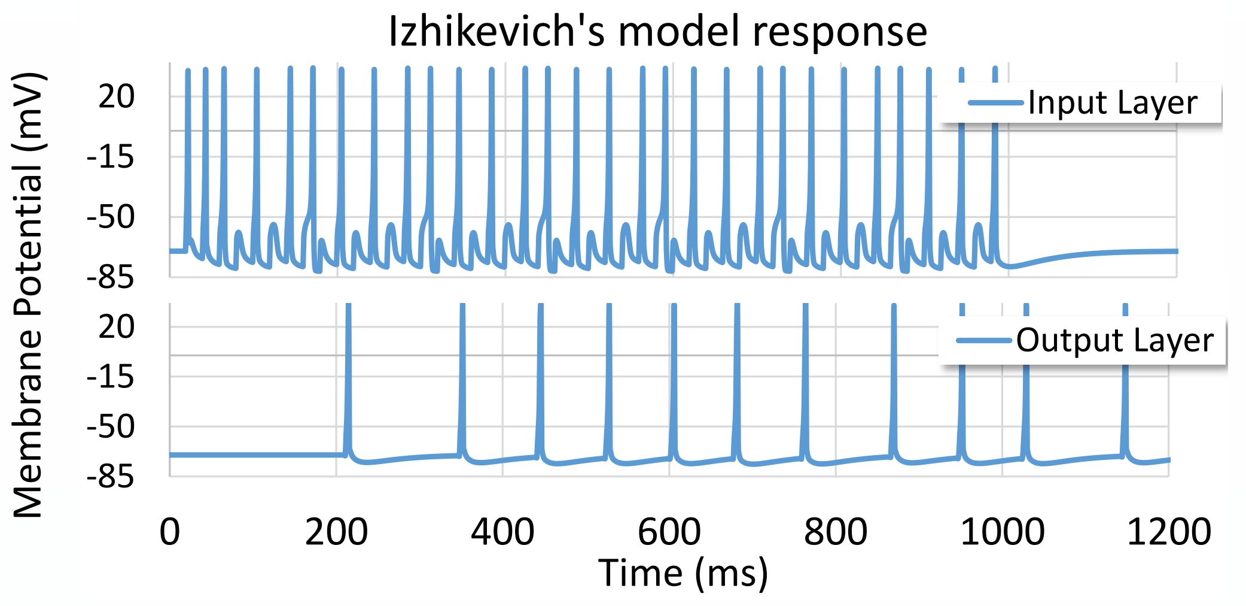


- ✓ **90% of spikes** are correlated precision with a **zero delay** against a **32** and **16-bits fixed** implementation.
- ✓ **87% of spikes** are distributed with a **delay of** $\mathbf{1.5} \pm \mathbf{2}\ \boldsymbol{ms}$ for a **8-bit fixed** precision.
- ✓ Autocorrelation **error** of firing rates in the output later of a 30x30 network is less than $\mathbf{0.0437}\%$

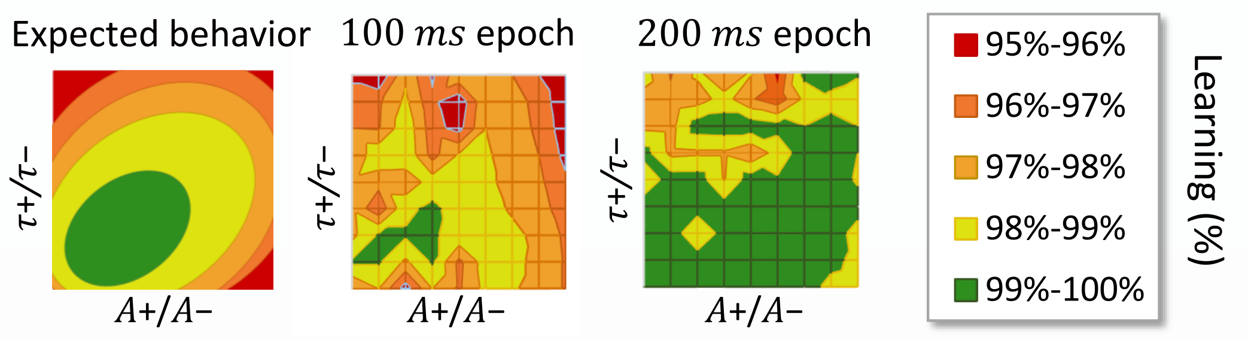
Simulations

A $\mathbf{1000}\ \boldsymbol{ms}$ simulation of a **50 x 50** network **fully connected** with **random** stimuli, weights and a distribution of $\mathbf{90\%/10\%}$ **excitatory** and **inhibitory** neurons is represented with the spikes of the $\mathbf{2,500}$ **neurons** and $\mathbf{125,000}$ **synapses** simulated, along with the **neuron model response** of neurons in the **input** and **output** layer.





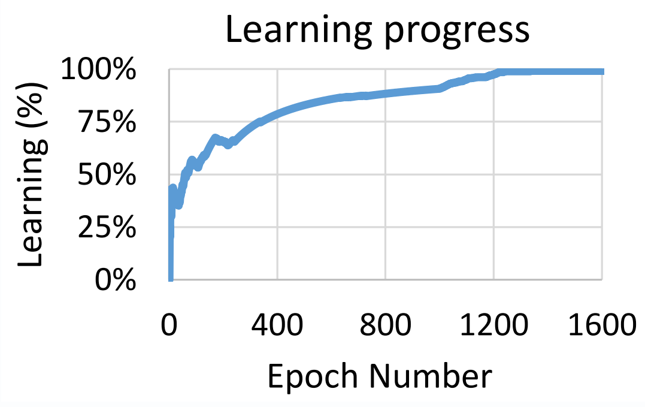
Spike Timing Dependent Plasticity (STDP) Learning

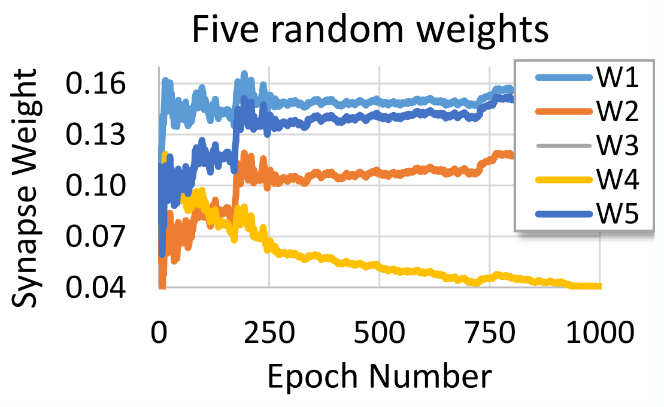


- ✓ The **faster learning** depends on the STDP learning amplitude $\boldsymbol{A_+/A_-}$ and time constants $\boldsymbol{\tau_+/\tau_-}$.
- ✓ The **longer** the **epoch time** is; the **faster** the **learning** will be.

XOR Benchmark

- ✓ **3-6-1** topology with a **first-to-spike** encoding learns a **XOR** with a **98%** accuracy in around **1,200 epochs**.
- ✓ Changes in weight **decrease** through the epochs.





Conclusions

- ✓ The **FPGA-based** neuron network achieves **higher performance** with **lower power consumption** than implementations using **high-end Intel processors.**
- ✓ **High-level synthesis (HLS)** significantly **increases productivity** for any design complexity compared with using native hardware description language (e.g. VHDL or Verilog)
- ✓ **Numerical precision** can increase the **performance** and **reduce energy** without a significant impact on the desired results.
- ✓ A **Hebbian-based learning algorithm** has been applied to the network implemented in order to perform **specific tasks.**

References

[1]. S. Areibi G. Lacey and G.W. Taylor, "Deep Learning on FPGAs: Past, Present, and Future," CoRR, abs/1602.04283, 2016

[2]. LeCun, Y., Bengio, Y., & Hinton, G. (2015, May). Deep learning [Review]. Nature, 521, 436-444.

[3]. E. Izhikevich, "Simple model of spiking neurons," IEEE Trans. Neural Netw. IEEE Transactions on Neural Networks, vol. 14, no. 6, pp. 1569–1572, 2003.

[4]. "Zynq-7000 All Programmable SoC Overview (DS190)," Xilinx, November-2015. [Online]. Available: http://www.xilinx.com/support/documentation/data_sheets/ds190-Zynq-7000-Overview.pdf. [Accessed: 2016].