STDP at Scale: CPU Bottlenecks and FPGA Offloading on a Zynq SoC

Maxime Carrière

Brain Language Laboratory, Freie Universität Berlin, Berlin, Germany

Spike-Timing Dependent Plasticity (STDP) is a central learning rule for Spiking Neural Networks (SNNs), but the precise conditions under which it becomes computationally expensive on conventional CPUs have not been systematically quantified, and lightweight FPGA implementations validated against reference simulators remain rare. This work makes two contributions.

CPU benchmark. We benchmark NEST 3.6 STDP synapses against static synapses across 1,950 parameter configurations (network size $N \in \{10\text{–}8{,}000\}$, connection probabilities $p \in \{0.05, 0.10, 0.20\}$, firing rates 10–40 Hz, 1–8 CPU cores). Overhead is negligible for small networks but grows to a $3.75\times$ slowdown at $N = 8{,}000$ ($p = 0.1$, Fig. 1A). Per-synapse cost exhibits a step increase above $N \approx 800$ neurons, consistent with the synapse table exceeding cache capacity and incurring main-memory traffic. Adding CPU cores reduces absolute runtime for both synapse types similarly, leaving the STDP overhead ratio unchanged across all core counts.

FPGA co-processor. A pipelined HLS kernel targeting the Xilinx Zynq XC7Z020 implements the online event-driven STDP rule using fixed-point arithmetic (Q1.15 weights, Q2.14 traces) and a 2,048-entry BRAM look-up table for exponential decay. The kernel achieves 500 ns per spike (50 cycles at 100 MHz), corresponding to 100 M syn/s for a 50×50 network, a $6\times$ reduction in compute latency relative to NEST (Fig. 1B), at 1.73 W board power. Hardware correctness is confirmed against a matched Python reference (MAE $= 1.91 \times 10^{-3}$, dominated by Q1.15 quantisation). The end-to-end system is currently bottlenecked by AXI-Lite host-transfer overhead (159 μs per spike); DMA batching is identified as the solution path.

Systems message. Both findings point to the same architectural conclusion: STDP performance at scale is governed by *where synaptic state lives and how often it must move*, not by multiply-accumulate cost. On the CPU, the bottleneck is cache pressure; on the current FPGA prototype, it is the host–accelerator interface. Future acceleration depends primarily on improving memory locality and data movement.

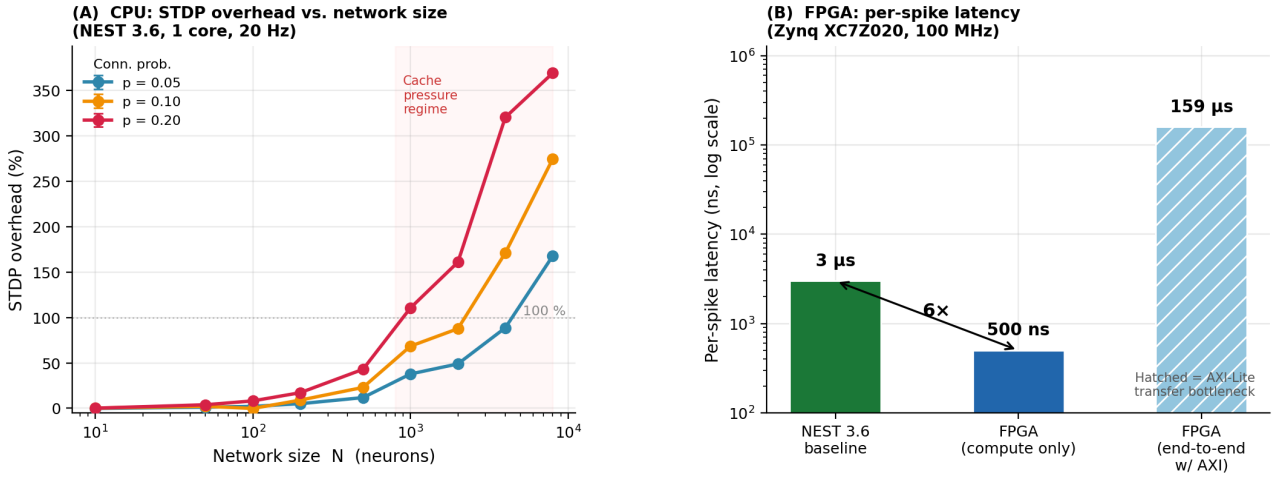

Figure 1: (A) STDP overhead in NEST 3.6 vs. network size; shaded region marks the cache-pressure regime above $N \approx 800$. (B) Per-spike latency on the Zynq XC7Z020: the FPGA compute kernel is $6\times$ faster than NEST; the hatched bar shows the current end-to-end cost dominated by AXI-Lite transfer overhead.