Kairos: A Novel Neuromorphic Architecture for Efficient Simulation of SNNs with Long-range Synaptic Delays

Amirreza Movahedin[(1,2)], Lennart P. L. Landsmeer[(2,1)], Said Hamdioui[(2)], and Christos Strydis[(1,2)]

[(1)] NeuroComputing Lab, Department of Neuroscience, Erasmus Medical Center

[(2)] Department of Quantum and Computer Engineering, Delft University of Technology

Recent advances in training Spiking Neural Networks (SNNs) have promoted variable-length synaptic delays to a first-class network parameter, improving temporal expressiveness and performance on temporally rich tasks. Yet, architectural support for long-range delays remains highly inefficient. Existing neuromorphic architectures generally follow one of two design approaches, dictated by whether spike delays are applied before or after the synapses. *Presynaptic* designs store delayed spikes compactly but repeatedly perform costly source-to-destination, synaptic resolutions across timesteps. *Postsynaptic* designs resolve spike destinations immediately but rely on delay buffers whose memory cost and underutilization grow with maximum supported delay. In effect, both design approaches struggle to scale efficiently to larger, trainable synaptic delays.

We present *Kairos*, a new design approach that rethinks delay handling at the architectural level by combining compressed presynaptic storage for long-term spikes (PreSpike memory) with a bounded postsynaptic short-term delivery window (WeightSum memory). This combination allows for a novel spike-delivery method where source-to-destination resolution is performed periodically. At the microarchitecture level, Kairos further improves performance by handling spike deliveries in parallel (across neurons and delay ranges), with parallelism being maximized based on the range of encountered delays per executed workload, in addition to aggressive eviction of delivered spikes.

To evaluate our proposed neuromorphic architecture, we build a configurable cycle-level, architectural simulator that models Kairos and other, competitive state-of-the-art architectures for fair energy, latency, and area comparisons and evaluate all of them across diverse SNN benchmarks. Against Intel Loihi 2 (postsynaptic design), Kairos supports at least 5× larger delays at the same memory and energy budgets and exhibits a 4.6× higher energy-latency product. Overall, Kairos achieves on average 78× and 2.5× improvements in the energy-latency product over presynaptic and postsynaptic designs, respectively, while generally demonstrating sublinear scaling with an increasing supported delay range.

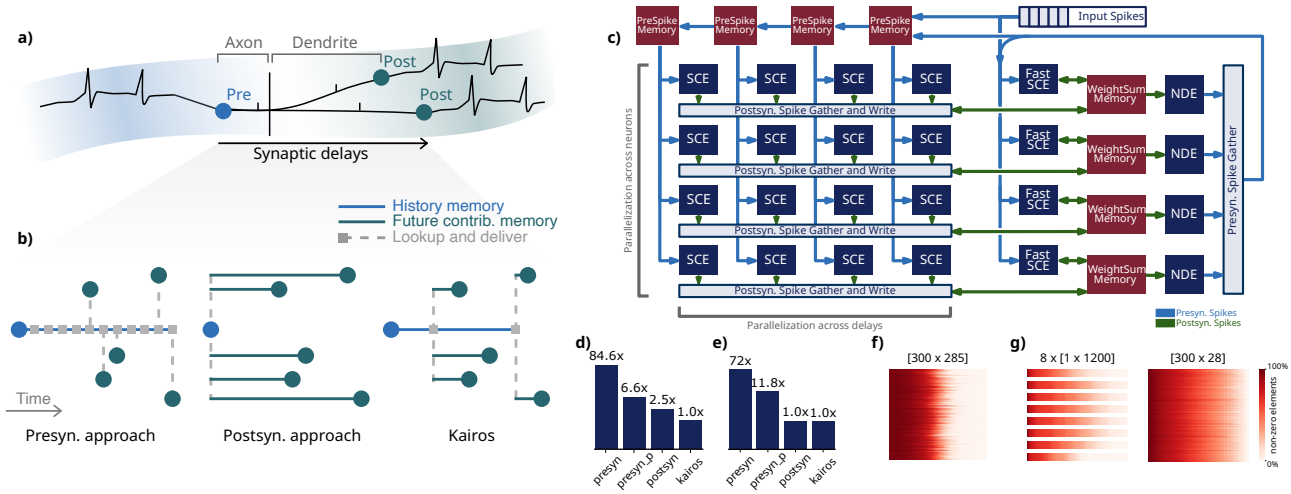

Figure 1: a) Synaptic delays in SNNs, b) Current and our proposed methods for handling delays, c) Kairos architecture, d) Average energy consumption, e) Average latency, f) Postsynaptic memory utilization, g) Kairos utilization of memories