Non-Gradient Neural Dynamics for Real-Time Edge Intelligence: A 53 μs Inference Engine Without Surrogate Gradients

Jianwei Lou[1]

[1]RailMind Systems, Neuss, Germany. j.lou@railmind.eu

Motivation

Edge-deployable predictive maintenance (PdM) demands real-time inference ($<$1 ms), online adaptation without retraining, and instantaneous task switching—requirements poorly served by both conventional ML pipelines (offline training, GPU dependence) and current spiking neural network (SNN) approaches that rely on surrogate gradients [6] or backpropagation through time for training. We present a non-gradient neural dynamics engine that meets all three requirements on commercial off-the-shelf (COTS) hardware costing under $60, achieving 53 μs end-to-end latency—within the operating envelope of dedicated neuromorphic ASICs such as Loihi 2 ($\sim$10 μs)—while providing auditable, cross-industry benchmarks.

System Overview

The engine implements a population of N computational units with D-dimensional state vectors updated exclusively via local, bio-inspired plasticity rules—**no surrogate gradients**, no backpropagation, no global loss function. Population dynamics are governed by: (i) *local Hebbian-type plasticity* for unsupervised content learning (cf. [7] for related sequence learning in SNNs), (ii) *discrete population-level gating* for context-selective activation, and (iii) *energy-regulated neural turnover* for structural adaptation under finite resource budgets. The frozen model occupies 50 KB; the full pipeline requires 112 KB and no GPU.

Results

We evaluate across three independent industrial domains using standard benchmark datasets (Table 1), with *identical* model weights and *zero retraining* between domains.

Table 1. Cross-industry edge benchmarks (identical 50 KB model, zero retraining).

Domain / Dataset	AUC	Latency	Power	Task Switch	Platform
Bearing PdM (CWRU)	0.998	53 μs	—	0 ms	x86 sim
Aerospace (ESA-ADB)	0.911	53 μs	—	0 ms	x86 sim
Turbofan (NASA C-MAPSS)	0.958	53 μs	—	0 ms	x86 sim
Bearing PdM (CWRU)	0.996	103 μs	2.73 W	0 ms	RPi5 ($60)

Key findings: (1) AUC $\geq$ 0.91 across all three domains *without retraining*—genuine zero-shot transfer. (2) Latency of 53 μs (x86) / 103 μs (RPi5 at 2.73 W) places the engine within 10$\times$ of Loihi 2 ASIC targets [2], on $60 hardware. (3) Task switching is instantaneous (0 ms) via discrete gating.

Neuromorphic Relevance

All primitives—local Hebbian plasticity, discrete gating, energy-regulated turnover—map onto SNN substrates without surrogate gradients [6] or BPTT, with structural parallels to Hebbian sequence learning [7] and cortical column gating. The 50 KB model fits in on-chip SRAM, making it a candidate for memristive crossbar arrays [3]; the 53 μs software latency provides a quantitative target for hardware acceleration.

References

1. Schuman C.D. et al., *Nat. Comput. Sci.*, 2, 10–19, 2022.
2. Davies M. et al., *Proc. IEEE*, 109(5), 911–934, 2021.
3. Indiveri G. et al., *Proc. IEEE*, 99(12), 2021–2039, 2011.
4. Lansner A. et al., *Biol. Cybern.*, 101, 227–240, 2009.
5. Smith W.A. et al., *Mech. Syst. Signal Process.*, 35, 188–203, 2013.
6. Neftci E.O. et al., *IEEE Signal Process. Mag.*, 36(6), 51–63, 2019.
7. Bouhadjar Y. et al., *PLoS Comput. Biol.*, 18(6), e1010233, 2022.