Fully asynchronous neural network for multi-core neuromorphic processor

Laurent Chen[(1)], Bernhard Vogginger[(2)] and Guangzhi Tang[(1)]

[(1)] Maastricht University, [(2)] TU Dresden

Neuromorphic processors achieve energy efficiency through asynchronous, event-driven computation. Each core processes arriving spikes without waiting for a global clock. Yet deployed neuromorphic networks remain synchronous: matrix multiplications force each layer to complete before the next begins, imposing time-step barriers (SpiNNaker2) or global synchronization checkpoints (Loihi) that prevent effective parallelism across neuromorphic cores.

We introduce an asynchronous event-driven (AED) deep learning framework, built on JAX and MPI, that eliminates matrix multiplication while retaining gradient-based backpropagation. Each layer runs as an independent process that wakes only when events arrive, integrates events into neural states, and immediately emits its output events; no global barrier is needed, and multiple layers compute concurrently. Events carry continuous values rather than binary spikes, enabling standard gradient-based training within a fully event-driven simulation. A suite of sparsity mechanisms achieves greater than 99% event sparsity for asynchronous integration while preserving accuracy.

Evaluated across MLP, CNN, and gated RNN architectures, the framework matches synchronous baselines on six benchmarks: NMNIST (97.48%), SMNIST (93%), SHD (67.80%), DVS Gesture (78.99%), N-CARS (83.79%), and a primate reaching intracortical neural decoding task (R^2 0.53). Deploying fully asynchronous inference on SpiNNaker2 produces outputs that match simulation exactly, confirming that trained networks run directly on neuromorphic silicon without adaptation.

These results establish AED as a principled route to scalable neuromorphic inference. Because every operation is event-driven and barrier-free by construction, multi-core processors can exploit their full parallelism rather than emulating a synchronous pipeline. Future work targets hardware-in-the-loop training, where the neuromorphic substrate participates directly in gradient computation.

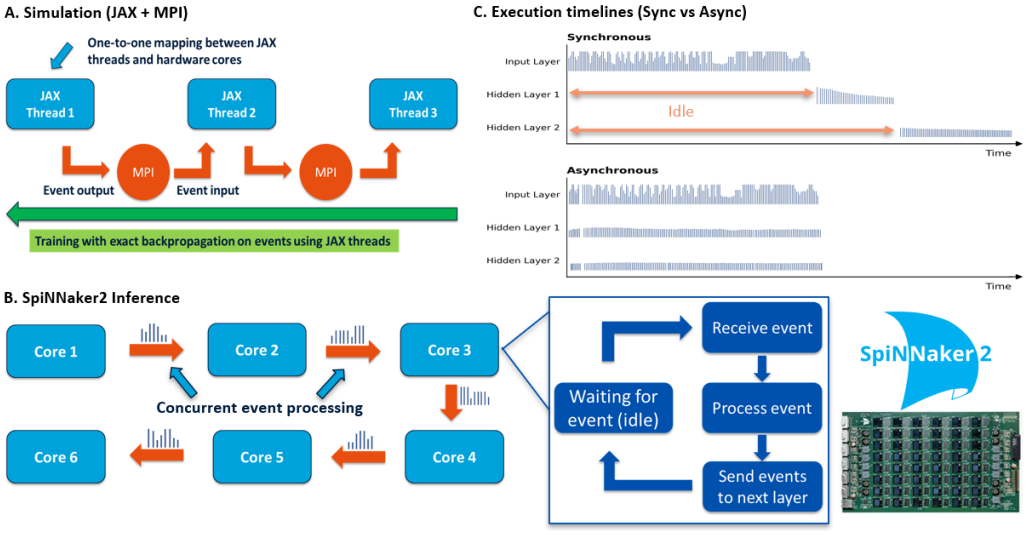

Figure 1: **Illustration of the AED framework.** (A) In simulation, each JAX thread maps one-to-one to a hardware core and communicates via MPI using sparse event messages, while remaining fully compatible with exact backpropagation. (B) During inference on SpiNNaker2, each core runs as an independent event-driven process, idle until an event arrives, then immediately processing and forwarding it to the next layer without any global synchronization. (C) Example execution timelines show how removing global synchronization eliminates layer-wise idle time which enables concurrent computation across all layers and reduces end-to-end latency.