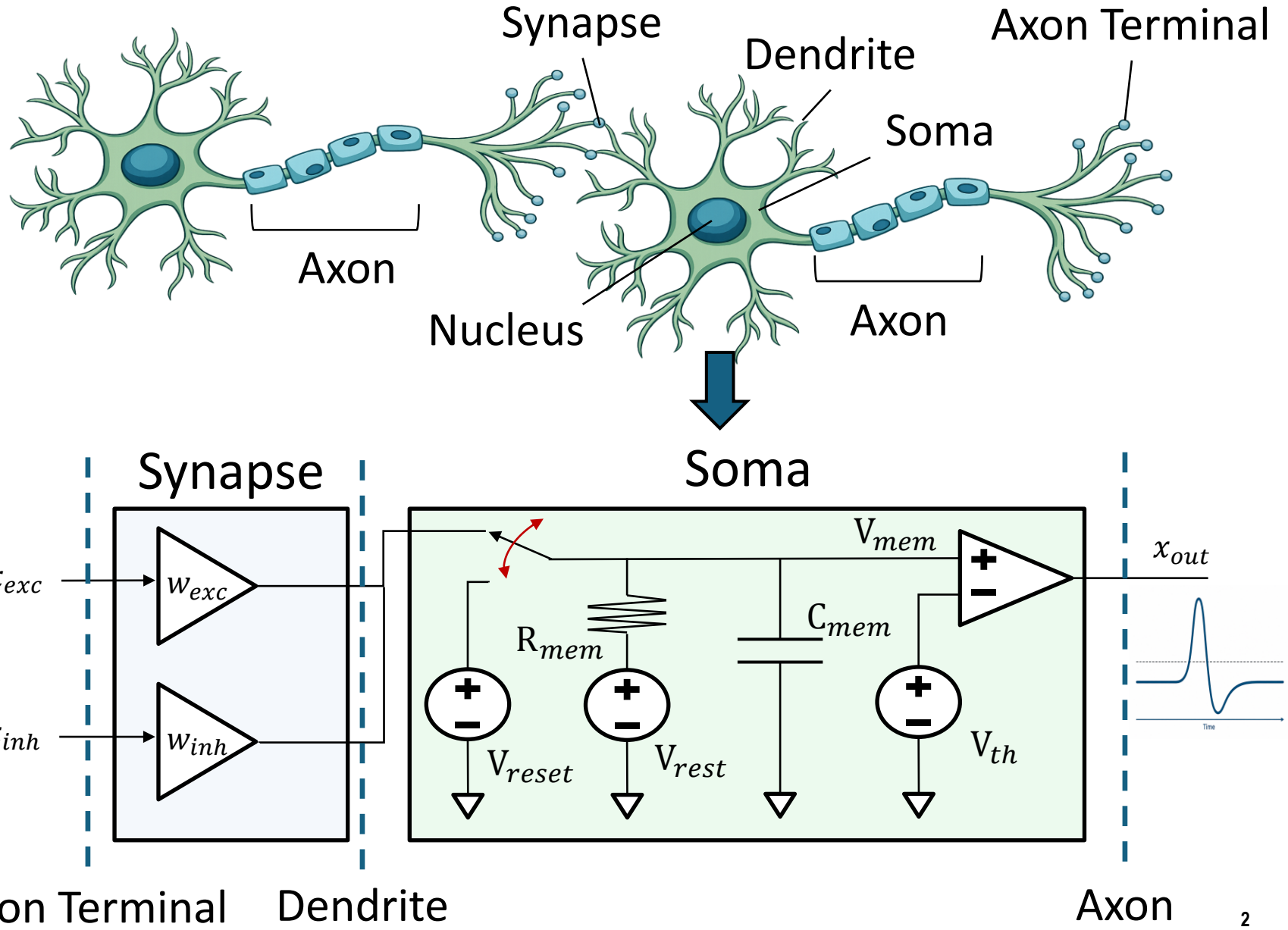
Energy efficient FPGA Spiking Neural Networks

Dimitrios Sekertzis

Electrical and Computer Engineering
Democritus University of Thrace, Xanthi, Greece

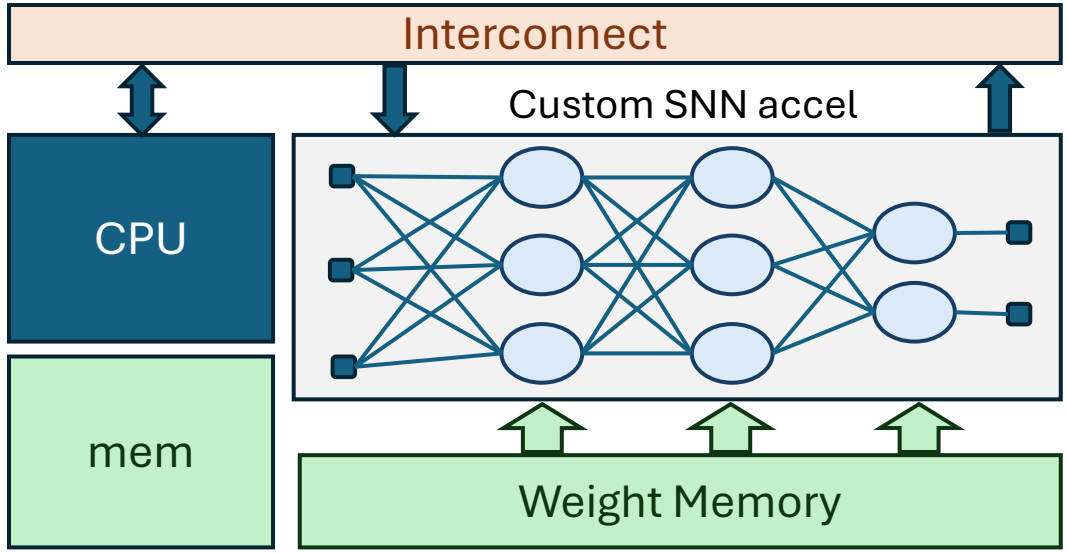
Spiking Neural Networks (SNNs)

- **Brain inspired machine learning**
  - Models the behavior of biological neurons
  - Information is transmitted through discrete spikes
- **Leaky Integrate and Fire (LIF) model**
  - **Leakage** causes the potential to decay over time
  - Membrane potential **integrates** incoming spikes
  - A **spike** is generated when the threshold is reached
  - Membrane potential is **reset** after firing

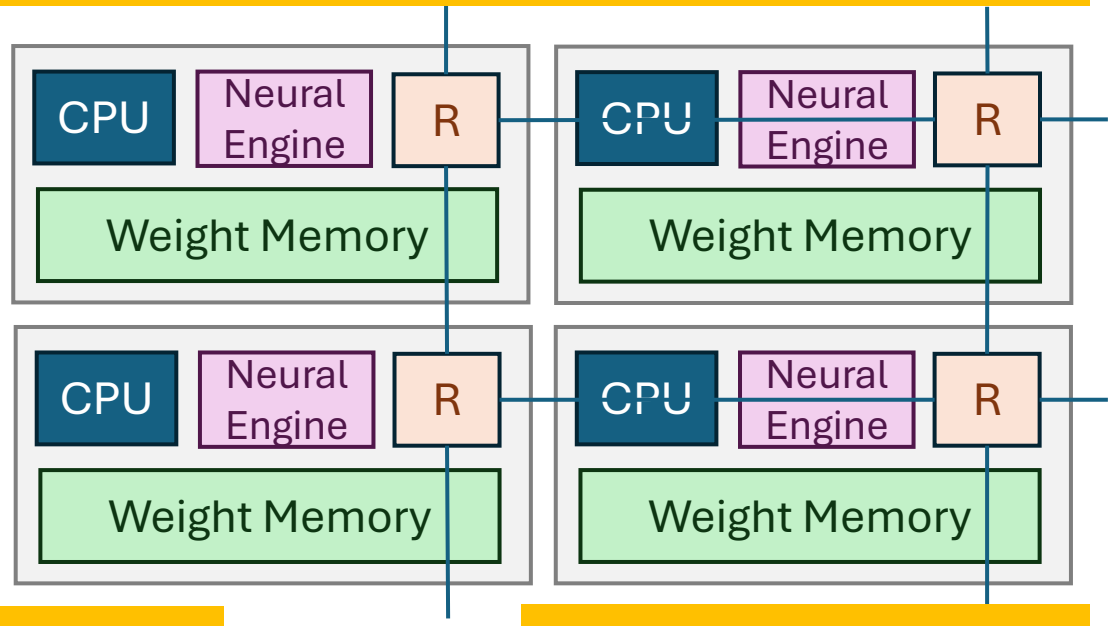


SNNs in hardware

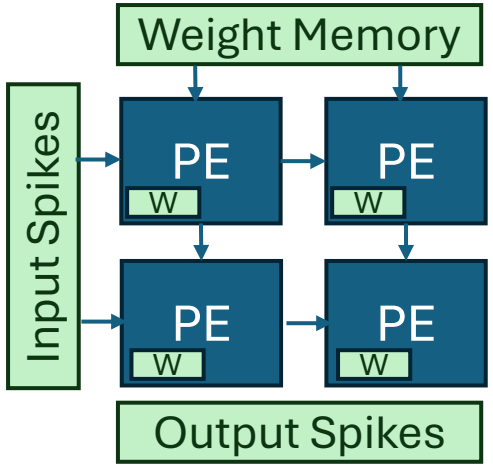
CPU orchestrator with custom SNN accel



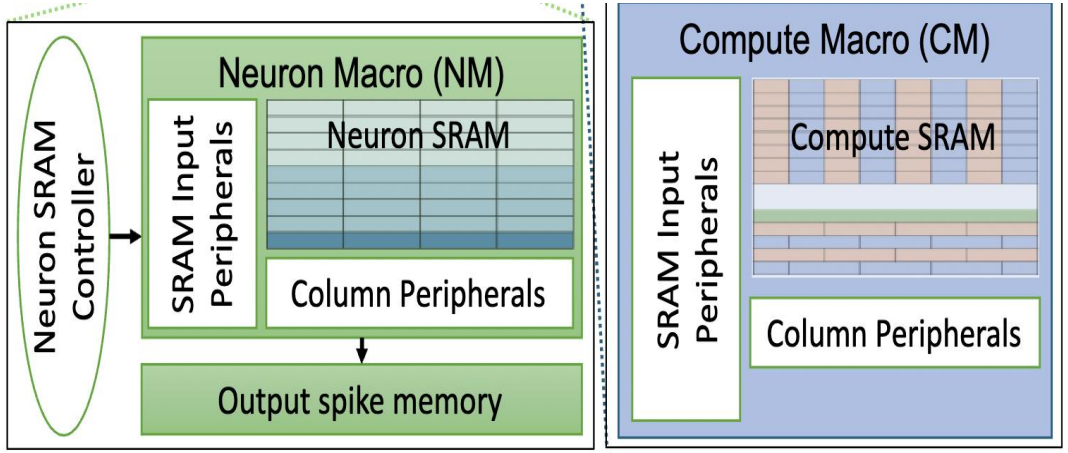
Small CPUs – Neural engines – Regural NoC



Systolic SNN accel



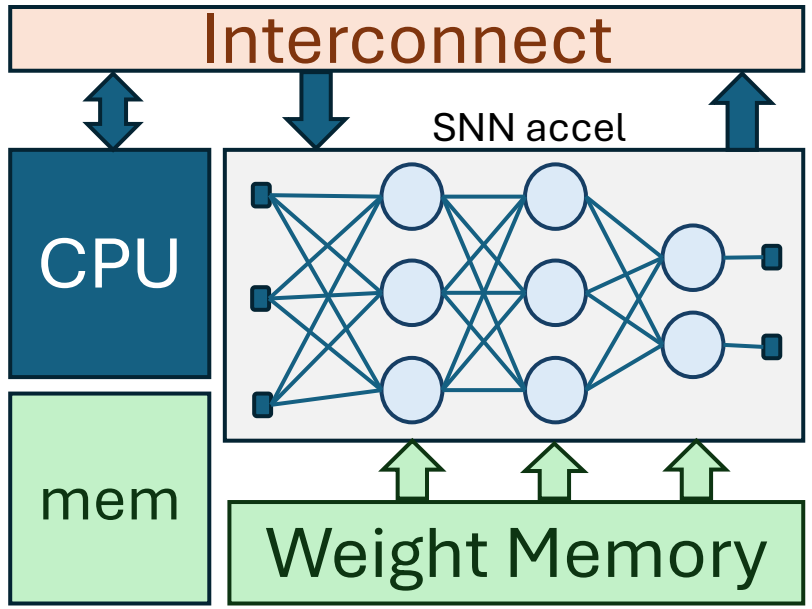
In memory SNN accel



Other approaches

- Analog/Mixed signal design
- SNNs with silicon photonics or other emerging devices

Scaling challenges for SNNs in FPGA

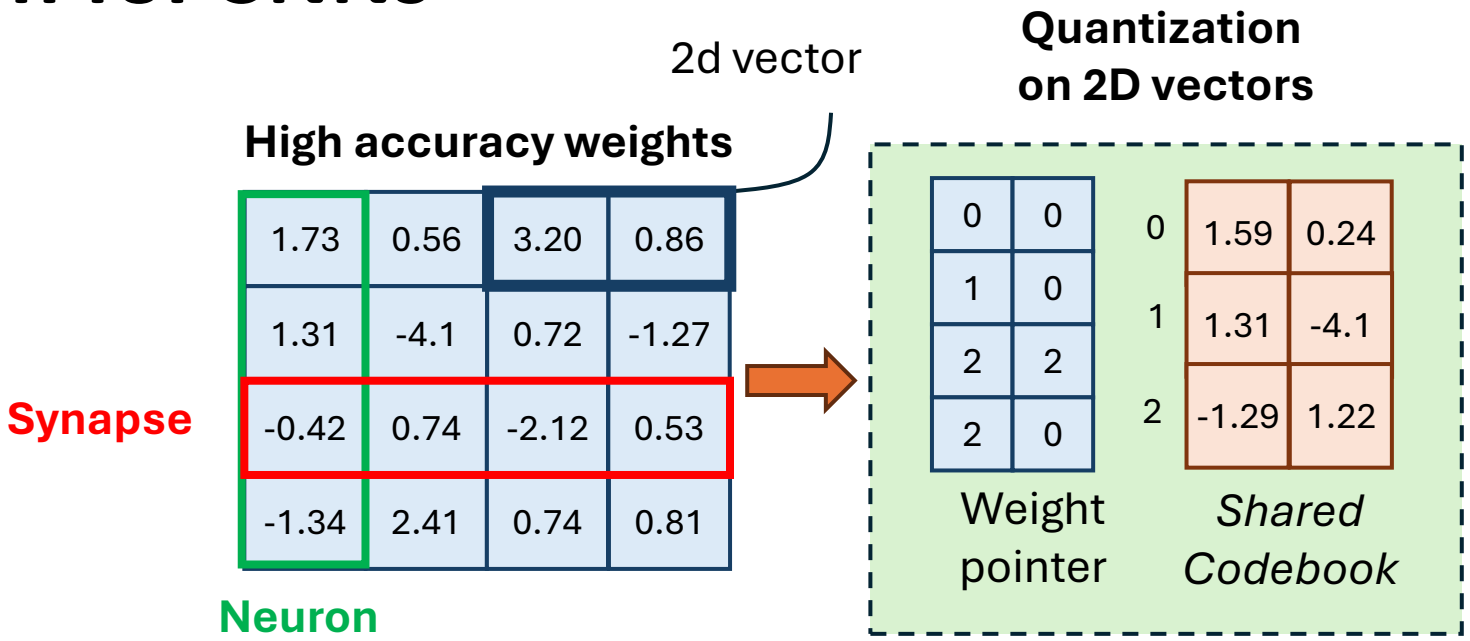


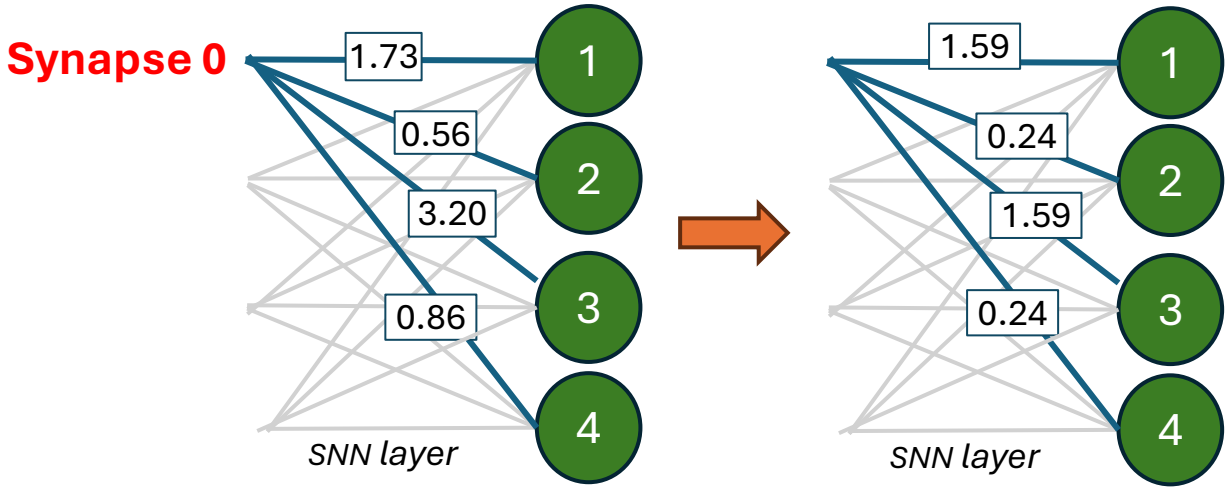
- Programmable SNN processors sacrifice efficiency for flexibility.
  - Runtime programmability incurs substantial memory, routing, and control overheads.
- We shift flexibility to FPGA design time
  - Model-specific hardware specialization replaces generic runtime mechanisms while retaining reconfigurability.
- But specialization alone is not enough. FPGA SNNs face four fundamental scalability walls:
  - Memory
  - Routing (fan-out)
  - Synchronization (event-driven vs. clocked execution)
  - Reliability

Vector quantization for SNNs

Proposed Approach

- Apply Vector Quantization to groups of synaptic weights
- Replace weight matrices with
  - Pointer memory
  - Shared codebook
- Interleaved execution scheme





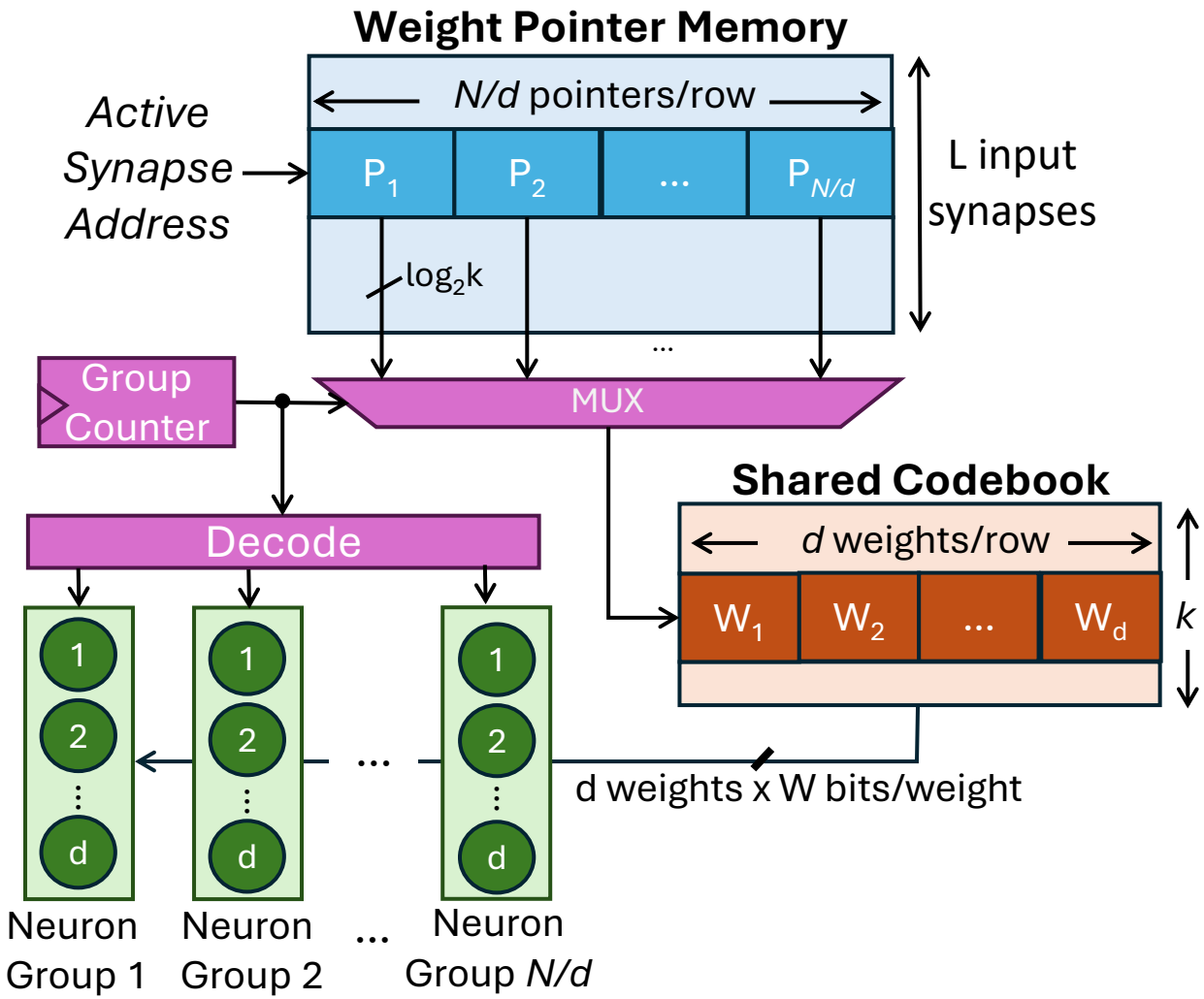
Vector quantization for SNNs

Proposed Approach

- Apply Vector Quantization to groups of synaptic weights
- Replace weight matrices with
  - Pointer memory
  - Shared codebook
- Interleaved execution scheme

Benefits

- 52–61% BRAM reduction
- Minimal impact on inference accuracy
- Improved scalability for larger FPGA SNNs



Fault Tolerance & Reliability

Memory optimizations

- Volatile memories are susceptible to radiation-induced soft errors
- Reduced memory footprint lowers the number of vulnerable cells

Runtime monitoring

- Algorithm-based fault-tolerant (ABFT) online checkers identify computational errors
- Spike Anomaly Watchdogs detect biologically implausible firing patterns in real time