KING'S College LONDON

Brain Inspired Computing and Hardware Design

By

Group B

Canberk Karabayir

Juan Dagorret Martinez

Omar Behaj Soullami

Introduction:

LOIHI: A neuromorphic Manycore Processor with On-Chip Learning

Novel Features:

- Hierarchical Connectivity
- Dendritic Compartments
- Synaptic Delays
- Programmable synaptic learning rules

Outline

- Introduction
- Spiking Neural Networks
- Architecture
- Design Implementation
- Results
- Conclusion

Spiking Neural Unit:

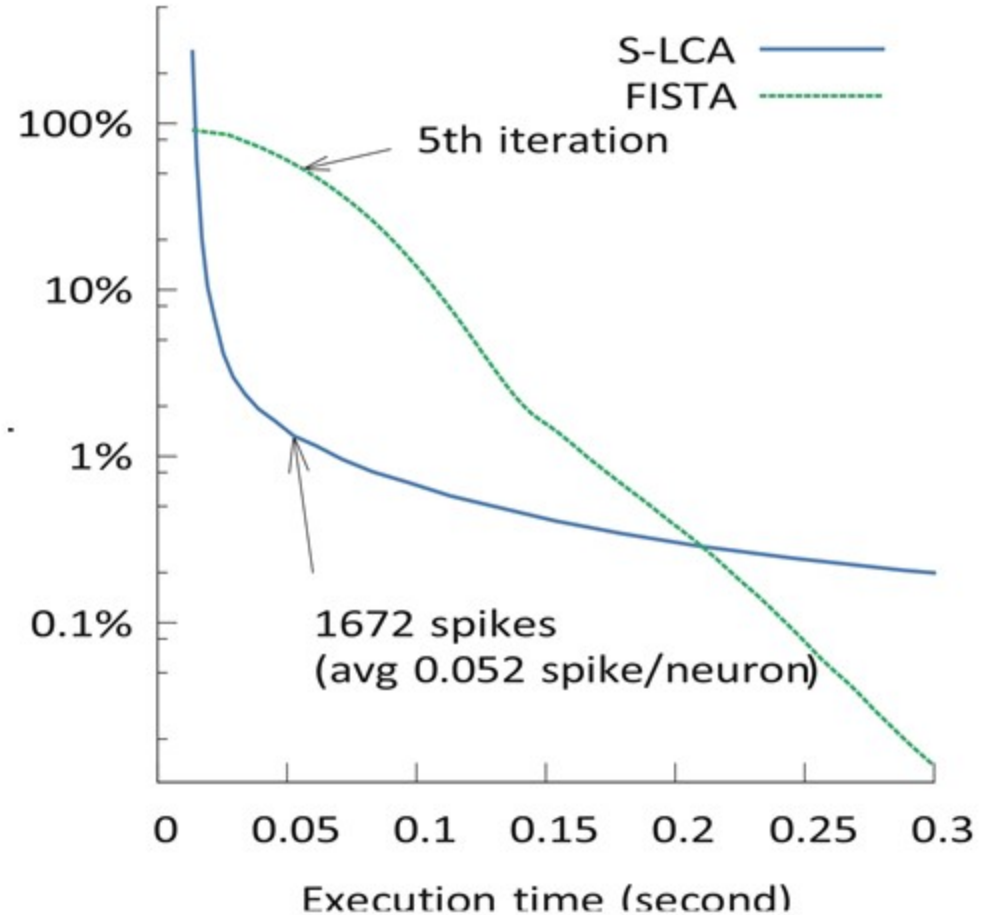
$$u_{i(t)} = \sum_{j \neq i} w_{i,\mathrm{j}} (\alpha_u * \sigma_j)(t) + b_i \qquad \text{Equation 1}$$

U(t)= Synaptic Response Current

$$\dot{v}_i(t) = -\frac{1}{\tau_v} v_i(t) + u_i(t) - \theta_i \sigma_i(t) \qquad \text{Equation 2}$$

V(t)= Membrane Potential

Computation with Spikes and Fine grained Parallelism:



S-LCA == SNN approach

FISTA == ANN approach

Learning with Local Information:

Pioneer work on decentralized learning algorithms:

- Oja's rule
- Widrow-Hoff rule
- complex unsupervised sparse dictionary learning using feedback
- event driven random back propagation

Local Features offered by LOIHI

- *Spike traces* corresponding to filtered presynaptic and postsynaptic spike trains with configurable time constants
- Multiple spike traces for a given spike train filtered with different time constants.
- Two additional state variables per synapse, besides the normal weight, in order to provide more flexibility for learning.
- Reward traces that correspond to special *reward spikes* carrying signed impulse values to represent reward or punishment signals for reinforcement learning.

Other Computational Primitives:

- Stochastic noise.
- Configurable and adaptable synaptic, axon, and refractory delays.
- Configurable dendritic tree processing.
- Neuron threshold adaptation in support of intrinsic excitability homeostasis
- Scaling and saturation of synaptic weights in support of “permanence” levels that exceed the range of weights used during inference.

ARCHITECTURE

Overview

Loihi has a manycore mesh with 128 neuromorphic cores, 3 embedded x86 processor cores, and off-chip communication interfaces that hierarchically extend the mesh in four planar directions to other chips. All communication is transported between cores in the form of packetized messages.

Messages may be hierarchically encapsulated for offchip communication over a second-level network.

Their state variables are updated in every algorithmic time-step. When the activation exceeds a threshold, it generates a spike message that is sent to fan-out compartments in the destination cores.

• Sparse network compression. • Core-to-core multicast. • Variable synaptic formats. • Population-based hierarchical connectivity.

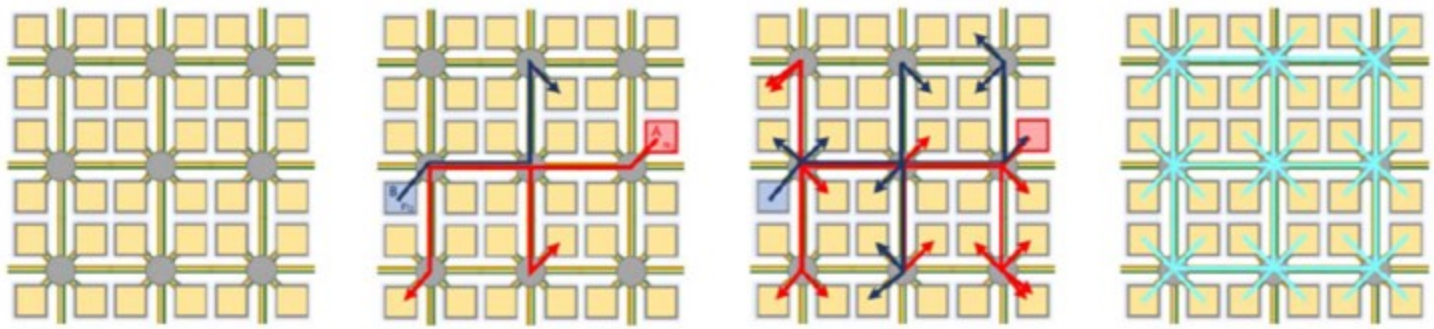
Loihi is the first SNN chip that supports any of the above features. Each core has a programmable learning engine that can evolve synaptic state variables as a function of historical spike activity. Their learning rules are microcode programmable and support a selection of input terms and output synaptic target variables.

ARCHITECTURE (CONTINUED)

All logic in the chip is digital, functionally deterministic, and implemented in an asynchronous bundled data design style. This allows spikes to be generated, routed, and consumed in an event-driven manner with maximal activity gating during idle periods. This implementation style is well suited for SNNs that fundamentally feature a high degree of sparseness in their activity across both space and time.

MESH OPERATION

Figure 2 shows the operation of the neuromorphic mesh as it executes an SNN model. All cores begin at algorithmic time-step t. Each core independently iterates over its set of neuron compartments, and any neurons that enter a firing state generate spike messages that the NoC distributes to all cores that contain their synaptic fan-outs.



NETWORK CONNECTIVITY ARCHITECTURE

In its most abstract formulation, the neural network mapped to the Loihi architecture is a directed multigraph structure G = (N,S), where N is the set of neurons in the network and S is a set of synapses (edges) connecting pairs of neurons. In general, Loihi will autonomously modify the synaptic variables (wgt,dly,tag) according to programmed learning rules.

This connectivity architecture can support arbitrary multigraph networks subject to the cores' resource constraints:

1. The total number of neurons assigned to any core may not exceed 1,024 (Ncx).
2. The total synaptic fan-in state mapped to any core must not exceed 128 KB (Nsyn × 64b, subject to compression and list alignment considerations).
3. The total number of core-to-core fan-out edges mapped to any given core must not exceed 4,096 (Naxout). This corresponds to the number of output-side routing slots highlighted in yellow in Figure 3.
4. The total number of distribution lists, associated by axon_id, in any core must not exceed 4,096 (Naxin). This is the number of input-side axon_id routing slots highlighted in red in Figure 3.

Loihi supports a hierarchical network model. Convolutional artificial neural networks (ConvNets), in which a single kernel of weights is repeatedly applied to different patches of input pixels, is an example class of network that greatly benefits from hierarchy.

LEARNING ENGINE

Pairwise STDP is simple, event-driven, and highly amenable to hardware implementation. For a given synapse connecting presynaptic neuron j to postsynaptic neuron i, an implementation needs only maintain the most recent spike times for the two neurons.

This is at odds with the algorithmic impetus for more complex network routing functions $R : j \rightarrow Y$, where $i \in Y$. The more complex R becomes, the more expensive, it becomes to implement an inverse lookup $R{-}1$ in hardware. Some implementations have explored solutions to this problem, but these approaches are not scalable. For Loihi, we adopt a less event-driven epoch-based synaptic modification architecture to support arbitrarily complex R and to extend the architecture to more advanced learning rules.

A number of architectural challenges arise in the pursuit of supporting more advanced learning rules. These rules are at the frontier of algorithm research and, therefore, require a high degree of configurability. Second, the rules involve multiple synaptic variables, not just weights. Finally, advanced learning rules rely on temporal correlations in spiking activity over a range of timescales, which means more than just the most recent spike times must be maintained. These challenges motivate the central features of Loihi's learning architecture.

Design Implementation

Core Microarchitecture

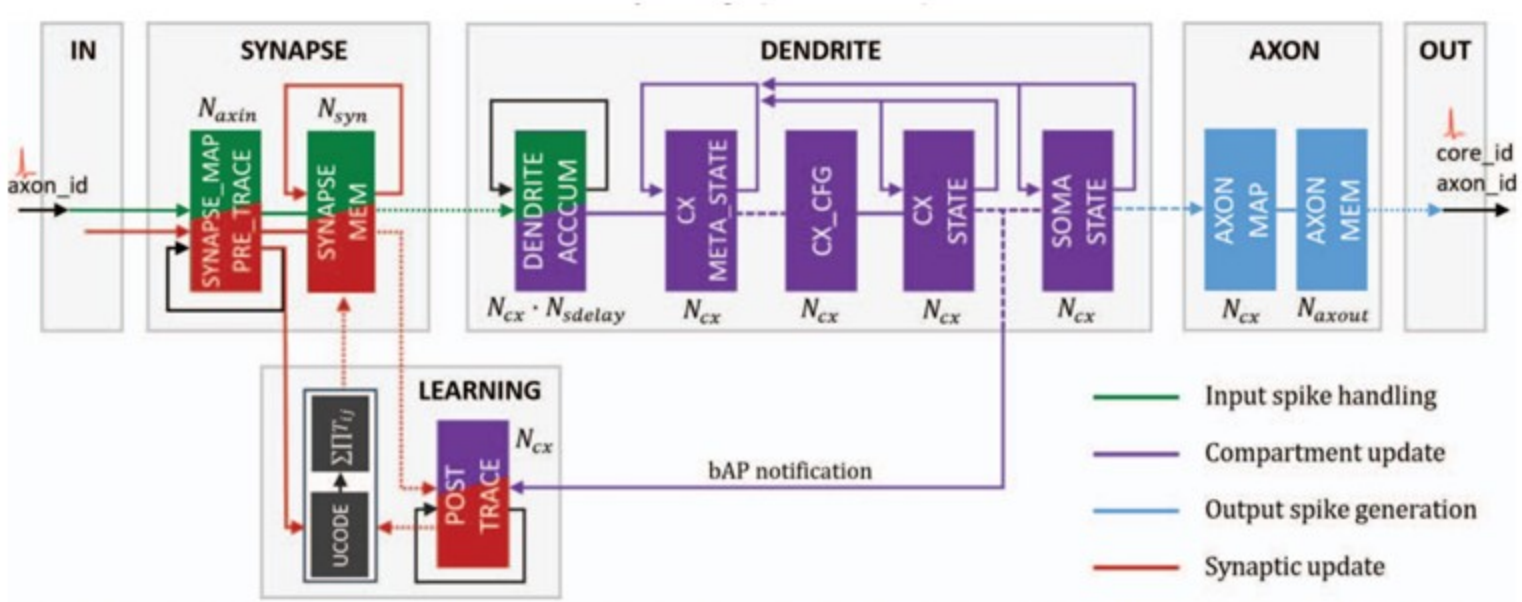


Fig. 4: Core Top-Level Microarchitecture. The SYNAPSE unit processes all incoming spikes and reads out the associated synaptic weights from the memory. The DENDRITE unit updates the state variables u and v of all neurons in the core. The AXON unit generates spike messages for all fanout cores of each firing neuron. The LEARNING unit updates synaptic weights using the programmed learning rules at epoch boundaries.

Asynchronous Design Methodology

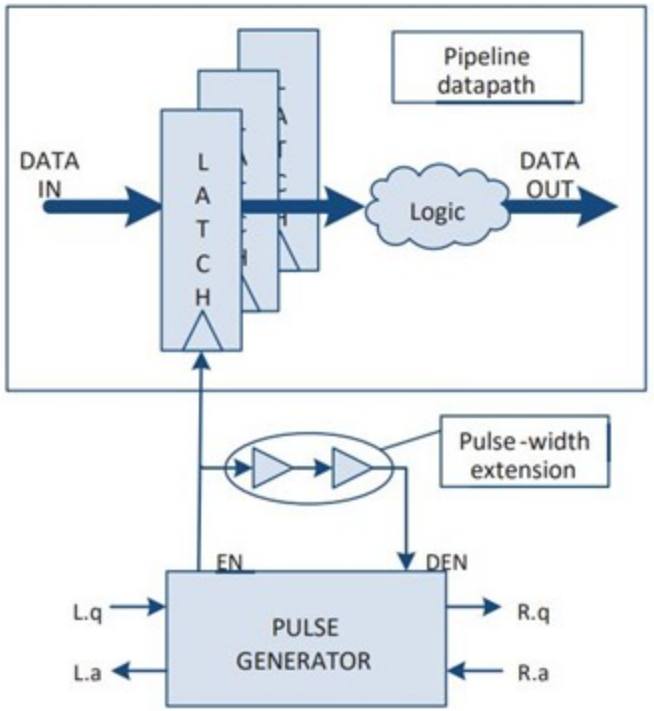


Fig. 5: Bundled data pipeline stage

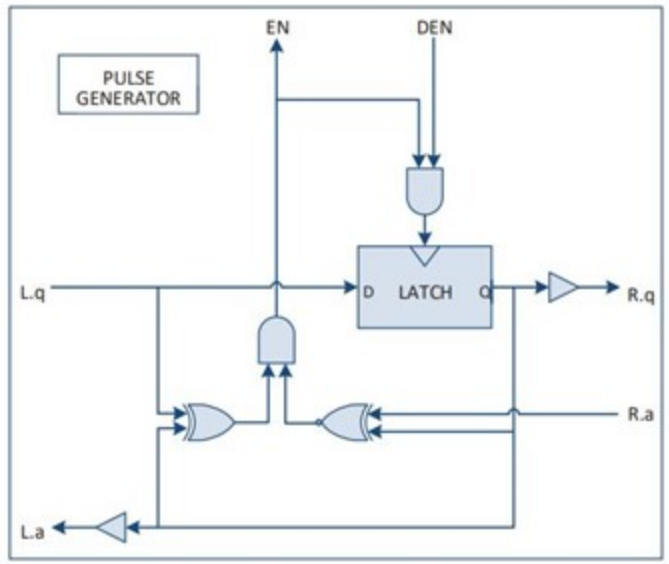


Fig. 6: Bundled data pulse generator circuit

Results:

(a) Original

(b) Reconstruction

No. Unknowns	400	1,700	32,256
No. nonzeros in solutions	≈10	≈30	≈420
Energy	2.58x	8.08x	48.74x
Delay	0.27x	2.76x	118.18x
EDP	0.7x	22.33x	5760x

Other experiments:

- A single-layer classifier using a supervised variant of STDP as the learning method.
- Solving the shortest path problem of a weighted graph.
- Solving a one-dimensional, non-Markovian sequential decision making problem.

Conclusion:

High-level points covered in our presentation:

- What was the question being studied?
- How did the authors try to answer it?
- Did they succeed in answering it?
- What are the implications of this work?

Reference:

This presentation is based on the research paper below. The figures and graphs used in this presentation are all from the below research paper.

“M. Davies et al., Loihi: A Neuromorphic Manycore Processor with On-Chip Learning, IEEE Micro 2018.”