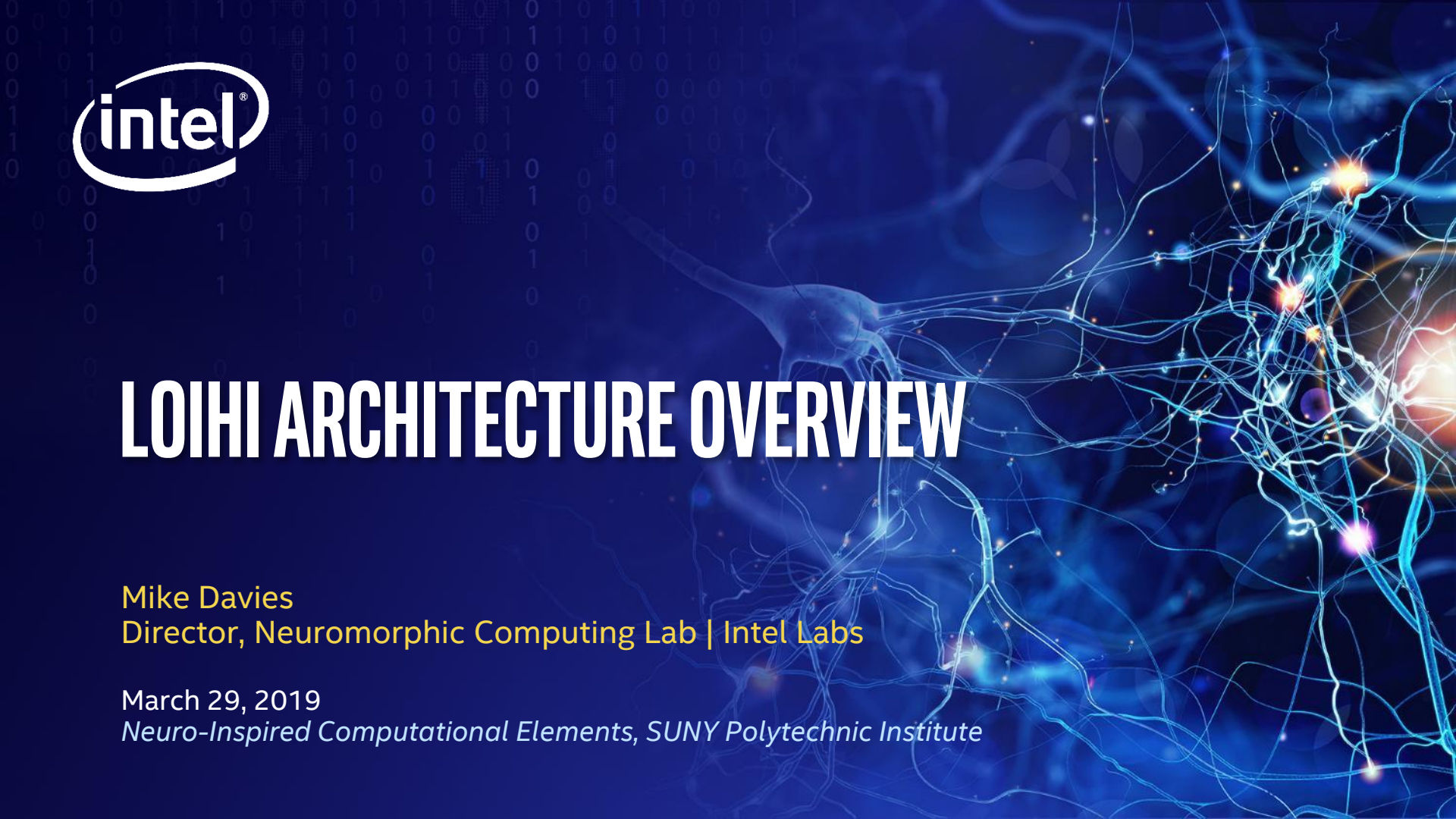


# LOIHI ARCHITECTURE OVERVIEW

Mike Davies
Director, Neuromorphic Computing Lab | Intel Labs

March 29, 2019
*Neuro-Inspired Computational Elements, SUNY Polytechnic Institute*

# LEGAL INFORMATION

This presentation contains the general insights and opinions of Intel Corporation ("Intel"). The information in this presentation is provided for information only and is not to be relied upon for any other purpose than educational. Intel makes no representations or warranties regarding the accuracy or completeness of the information in this presentation. Intel accepts no duty to update this presentation based on more current information. Intel is not liable for any damages, direct or indirect, consequential or otherwise, that may arise, directly or indirectly, from the use or misuse of the information in this presentation.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.

No computer system can be absolutely secure. No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document. Intel, the Intel logo, Movidius, Core, and Xeon are trademarks of Intel Corporation in the United States and other countries.

*Other names and brands may be claimed as the property of others

# The Engineering Perspective

- Nature has come up with something amazing. Let's copy it…
- Not so simple – very different design regimes
- Yet objectives and constraints are largely the same…
  - Energy minimization
  - Fast response time
  - Cheap to produce

Need to understand and apply the basic principles, *adapting for differences*

Status today:

| | Nature | Silicon | Ratio |
|---|---|---|---|
| Neuron density[1] | 100k/mm$^2$ | 5k/mm$^2$ | 20x |
| Synaptic area[1] | 0.001 um$^2$ | 0.4 um$^2$[2] | 400x |
| Synaptic Op Energy | ~2 fJ | ~4 pJ | 2000x |

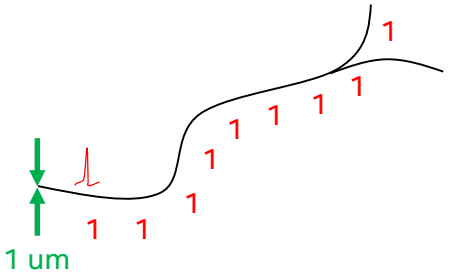
[1] Planar neocortex [2] ~5b SRAM

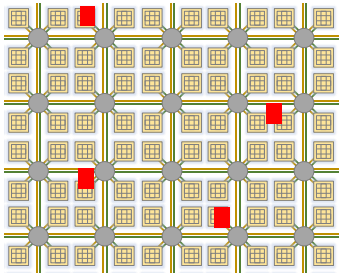
But…

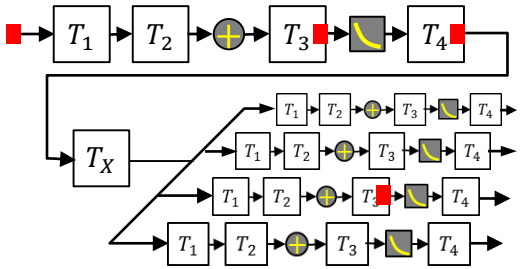
| | | | |
|---|---|---|---|
| Max firing rate | 100 Hz | 1 GHz | 10,000,000x |
| Synaptic error rate | 75% | 0% | ∞ |

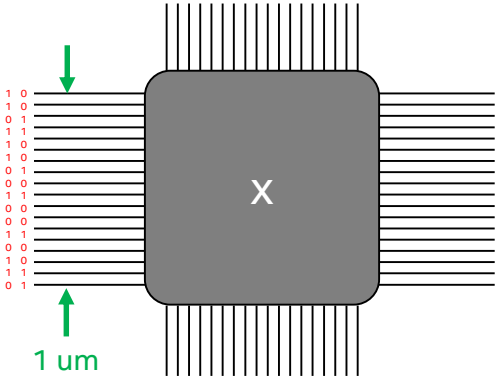
| Nature | Silicon |
|---|---|
| Autonomous self-assembly | Fabricated manufacturing |
| Per-instance variability desired | Variability causes brittle failures |
| Limited plasticity over lifetime | Must support rapid reprogramming |
| Nondeterministic operation | Deterministic operation desired |

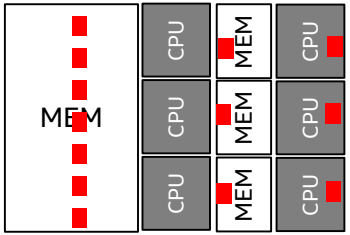
# Exploiting Sparsity with Spikes

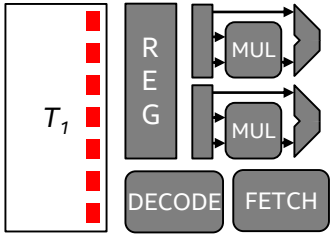




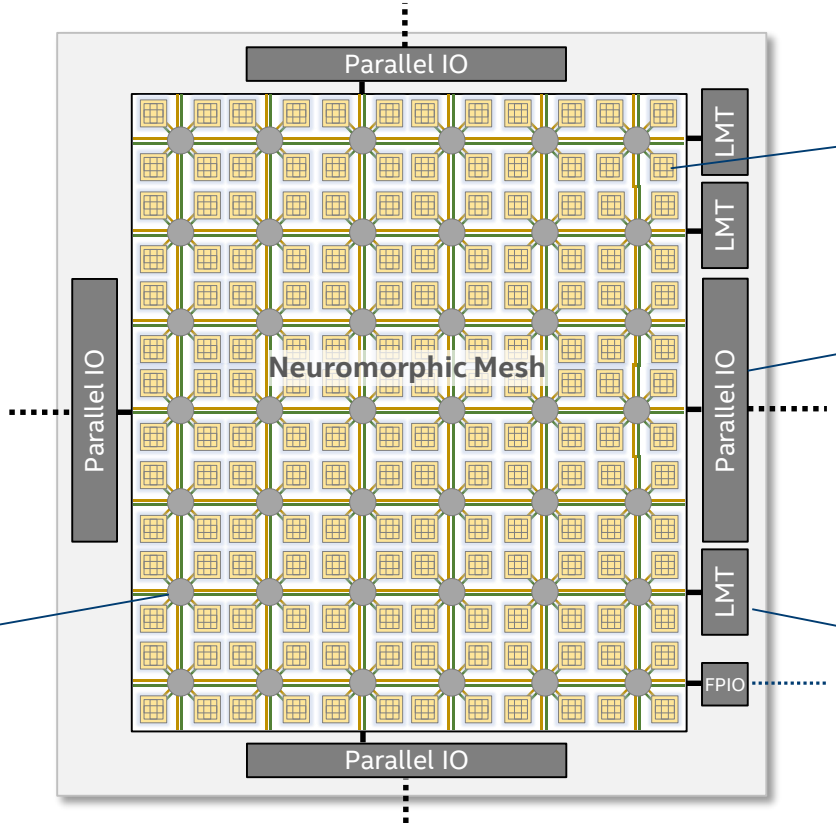



# Chip Architecture

| | |
|---|---|
| Technology: | 14nm |
| Die Area: | 60 $mm^2$ |
| Core area: | 0.41 $mm^2$ |
| NmC cores: | 128 cores |
| x86 cores: | 3 LMT cores |
| Max # neurons: | 128K neurons |
| Max # synapses: | 128M synapses |
| Transistors: | 2.07 billion |




**Neuromorphic core**
- LIF neuron model
- Programmable learning
- 128 KB synaptic memory
- Up to 1,024 neurons
- Asynchronous design

**Parallel off-chip interfaces**
- Two-phase asynchronous
- Single-ended signaling
- 100-200 MB/s BW

**Low-overhead NoC fabric**
- 8x16-core 2D mesh
- Scalable to 1000's cores
- Dimension order routed
- Two physical fabrics
- 8 GB/s per hop

**Embedded x86 processors**
- Efficient spike-based communication with neuromorphic cores
- Data encoding/decoding
- Network configuration
- Synchronous design

# Neuromorphic Core Architecture

All synaptic connections pooled
128KB shared memory

*Sparse*, *dense*, and *hierarchical*
Synaptic mapping representations

Synaptic delays

Synaptic eligibility traces

Flexible 3-tuple synaptic variables
(1-9b weight, 0-6b delay, 0-8b tag)

Graded “reward spikes”

**Flexible synaptic plasticity with microcode-programmable rules**

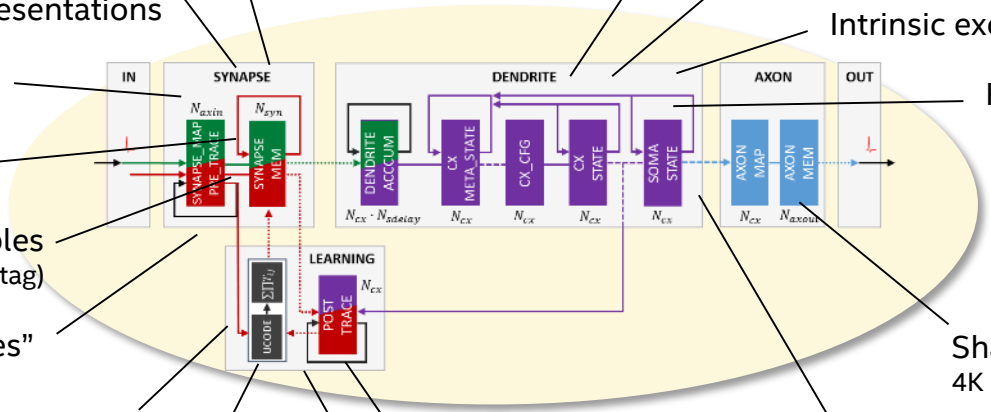
Sum-of-products rule semantics

Discrete time LIF neuron model (CUBA)

Multi-compartment dendritic trees
*up to 1K compartments*

Intrinsic excitability homeostasis

Random noise sources




Shared output routing table
4K axon routes

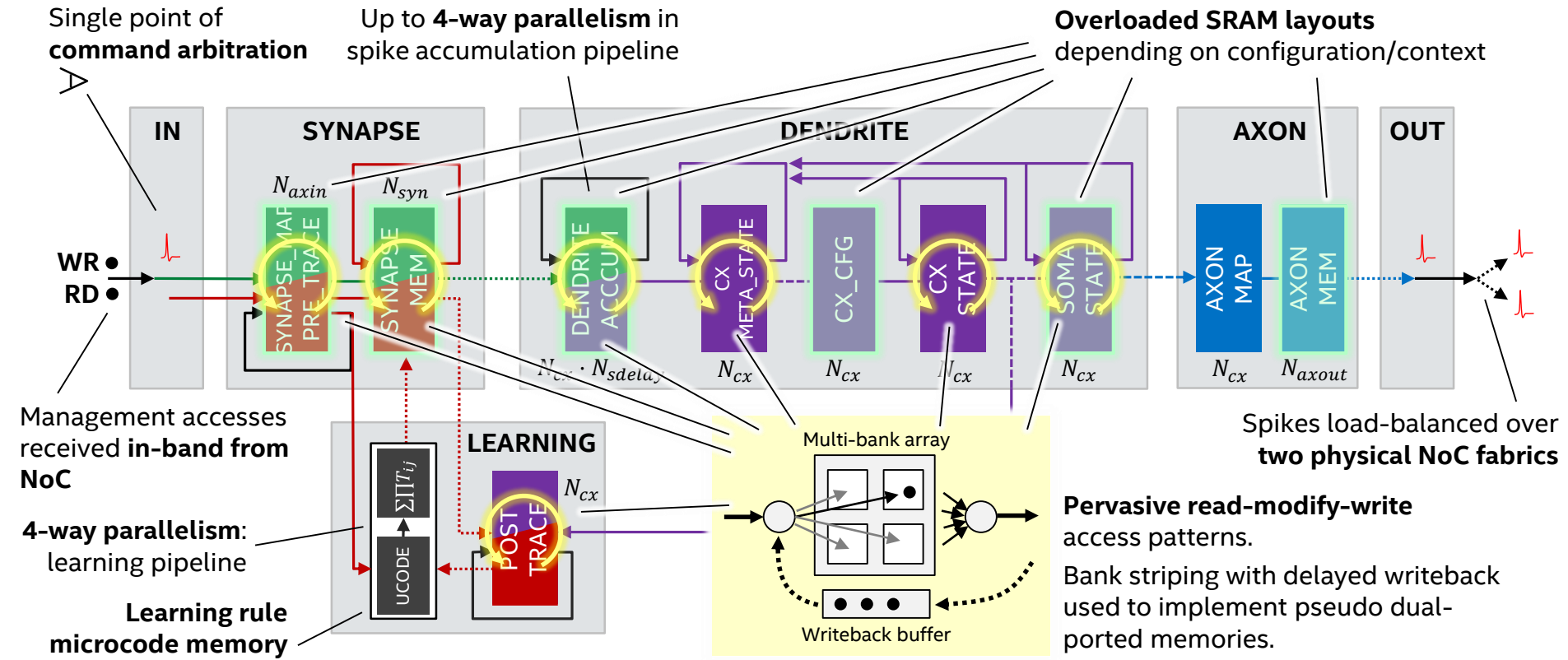
Axon delays

Refractory delays (+ random)

Filtered spike train traces

Plasticity rules target any synaptic variable

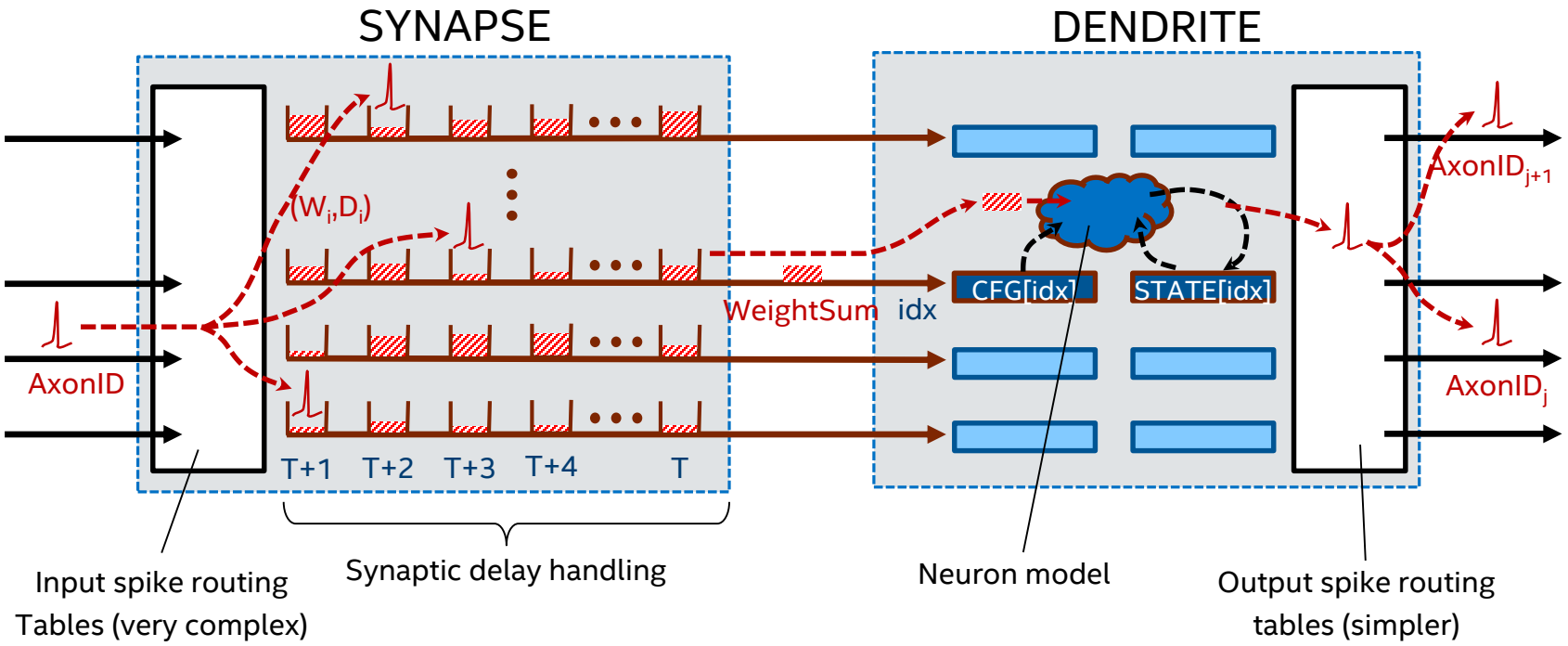
# Neuromorphic Core Microarchitecture




Widespread **variable iteration** and **unpredictable stalling** ⇒ **excellent match for async design**

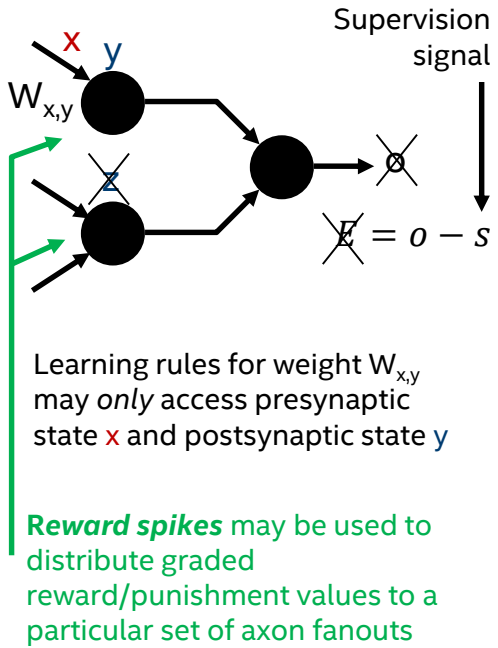
# Basic Core Operation (Non-Learning)

(Time multiplexing illustrated unrolled in space)

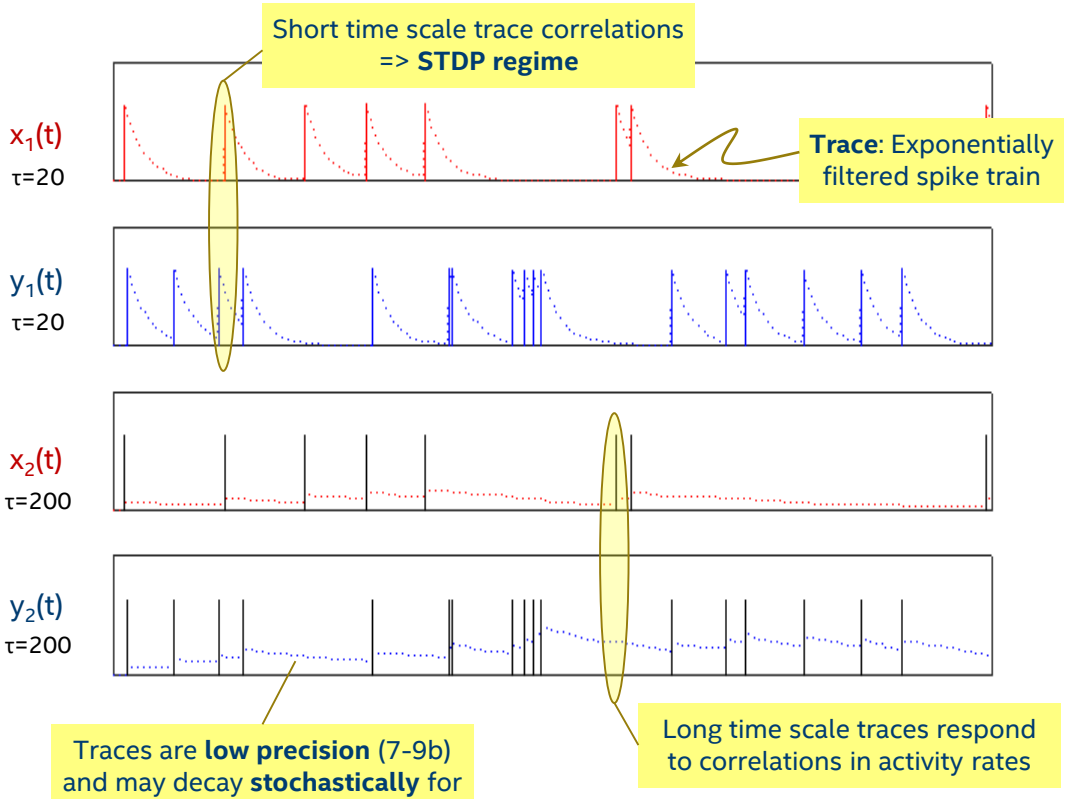


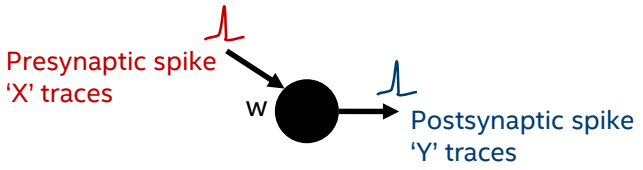
# Learning with Synaptic Plasticity

- **Local learning rules** – essential property for efficient scalability
- Rules derived by **optimizing an emergent statistical objective**
- Plasticity on **wide range of time scales** for
  - ✓ Immediate supervised (labelled) learning
  - ✓ Unsupervised self-organization
  - ✓ Working memory
  - ✓ Reinforcement-based delayed feedback



# Loihi's Trace-Based Programmable Learning




Weight, Delay, and Tag learning rules programmed as **sum-of-product equations**

$$w' = w + \sum_{i=1}^{N_P} S_i \prod_{j=1}^{n_i} (V_{i,j} + C_{i,j})$$

<u>Synaptic Variables</u>
Wgt, Delay, Tag
(variable precision)

<u>Variable Dependencies</u>
$X_0$, $Y_0$, $X_1$, $Y_1$, $X_2$, $Y_2$, $R_1$
Wgt, Delay, Tag, etc.

# Learning Rule Examples

**Pairwise STDP:**

$$W(t+1) = W(t) - A_{-}x_0(t)y_1(t) + A_{+}x_1(t)y_0(t)$$

**Triplet STDP with heterosynaptic decay:**

$$W(t+1) = W(t) - A_{-}x_0(t)y_1(t) + A_{+}x_1(t)y_0(t)y_2(t) - B \cdot W(t) \cdot y_3(t)$$

**Delay STDP:**

$$D(t+1) = D(t) - A_{-}x_0(t)(127 - y_1(t)) + A_{+}(127 - x_1(t))y_0(t)$$

# Two-variable Learning Rule Examples

**Distal Reward with Synaptic Tags:**

$$T(t+1) = T(t) - A_{-}x_0(t)y_1(t) + A_{+}x_1(t)y_0(t) - B \cdot T(t)$$

$$W(t+1) = W(t) + C \cdot r_1(t) \cdot T(t)$$

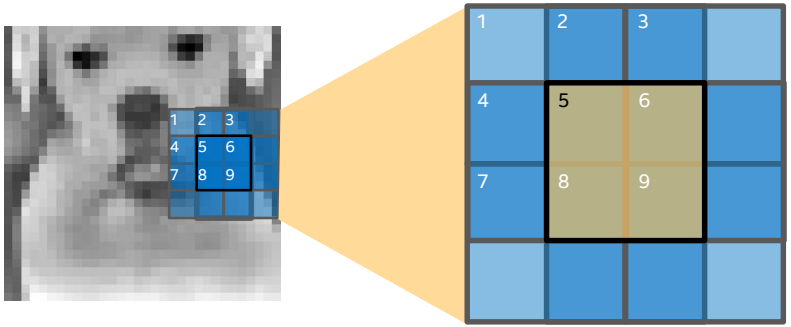
**STDP with dynamic weight consolidation:**

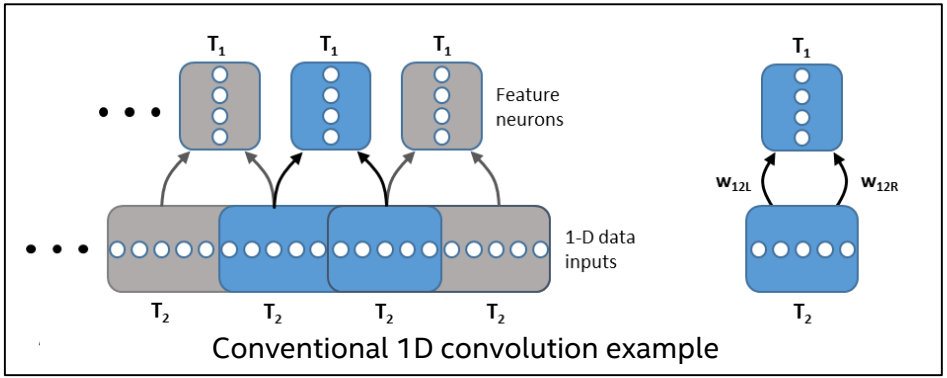
$$W(t+1) = W(t) - A_{-}x_0(t)y_1(t) + A_{+}x_1(t)y_0(t)y_2(t) - B_1(W-T)y_3(t)y_0(t)$$

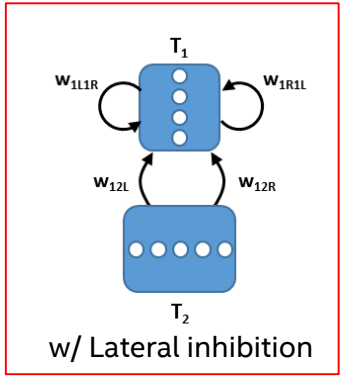
$$T(t+1) = T(t) + \frac{1}{\tau_{cons}}(W-T) - B_2 T(w_\theta - T)(w_{max} - T)$$

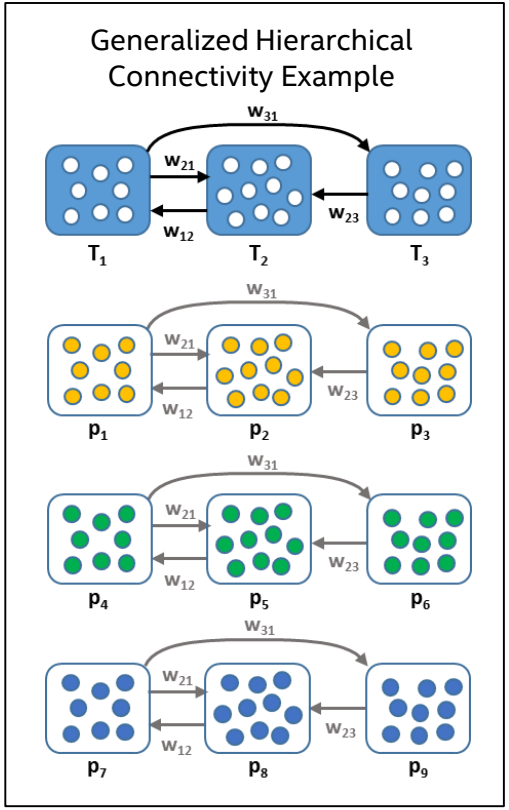
# Hierarchical Connectivity

## Generalization of the “convolution” in ConvNets

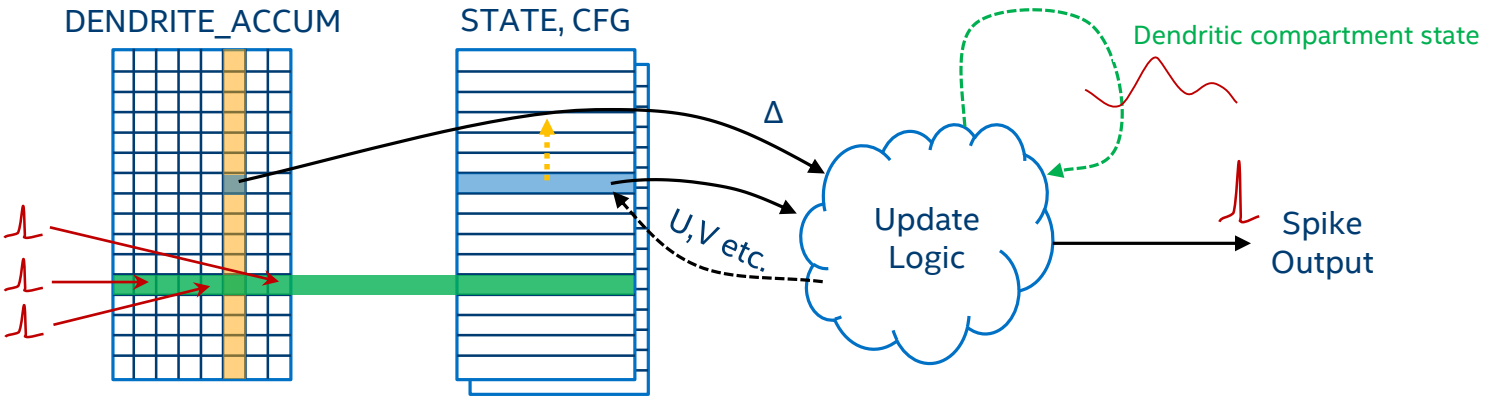


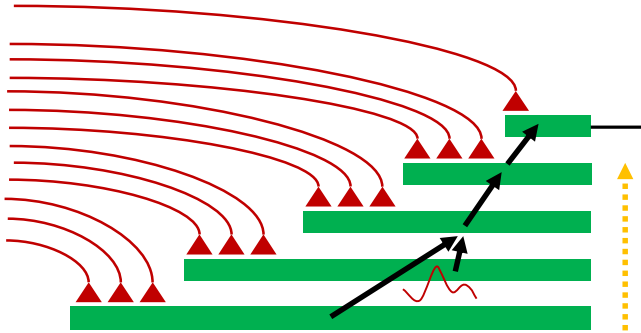



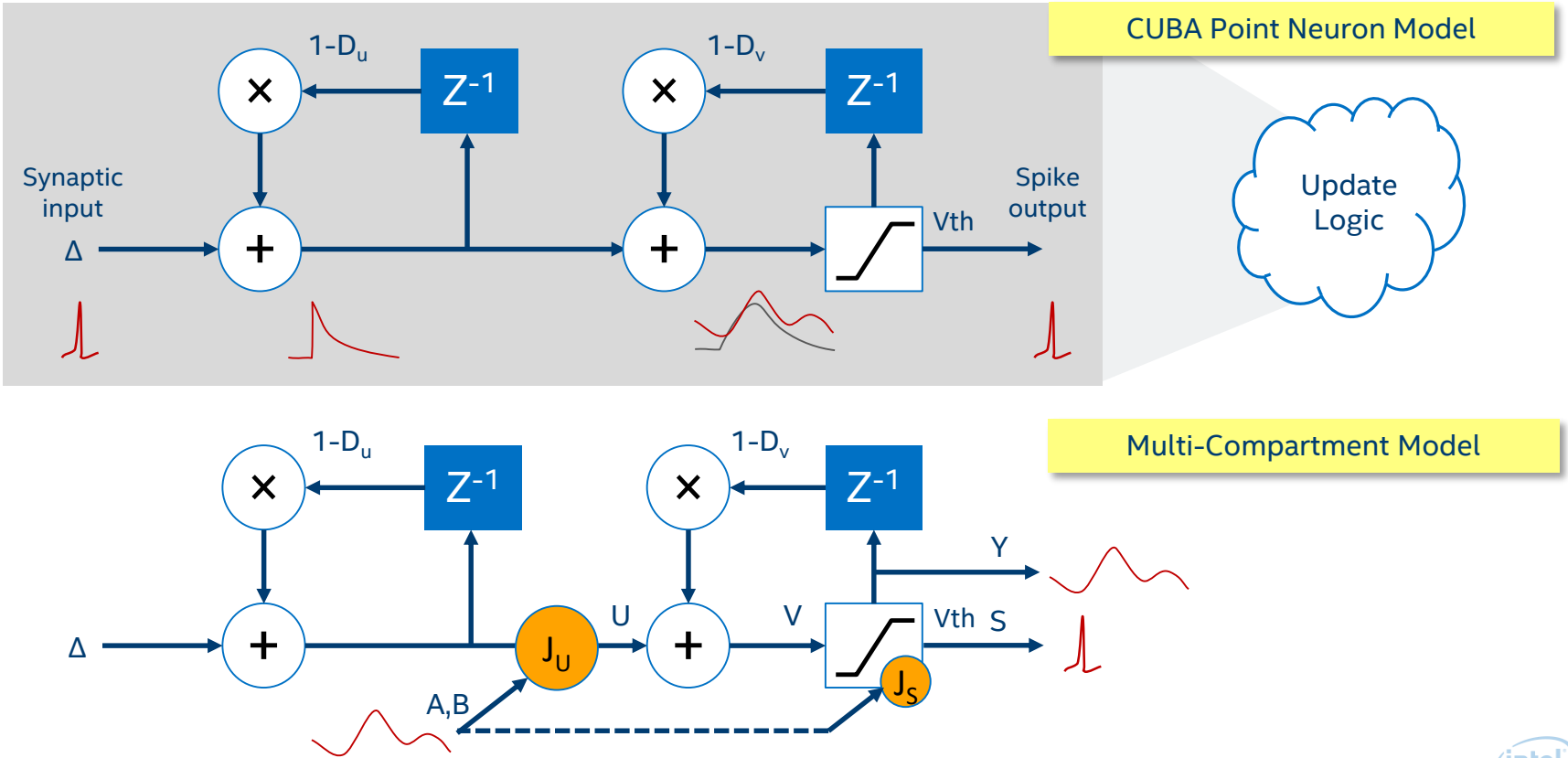
# Multi-Compartment Neurons



- Any DENDRITE index may be configured as either a compartment in the dendritic tree or a soma.
- Tree structure implemented by propagating and combining real-valued current/voltage state iteratively.

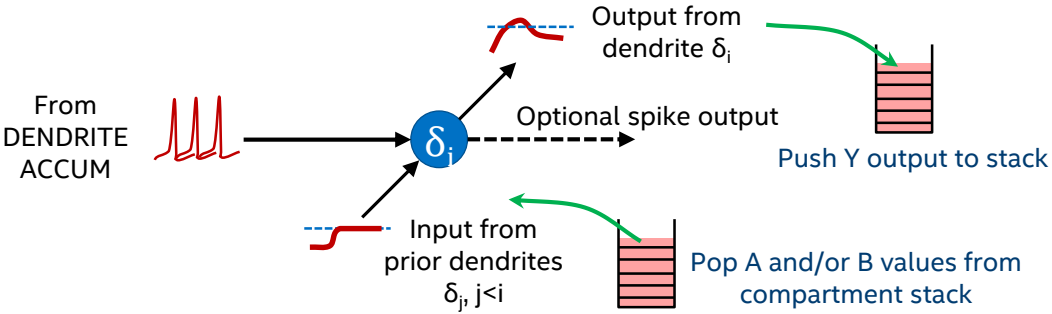
# Dendritic Compartment Unit Model

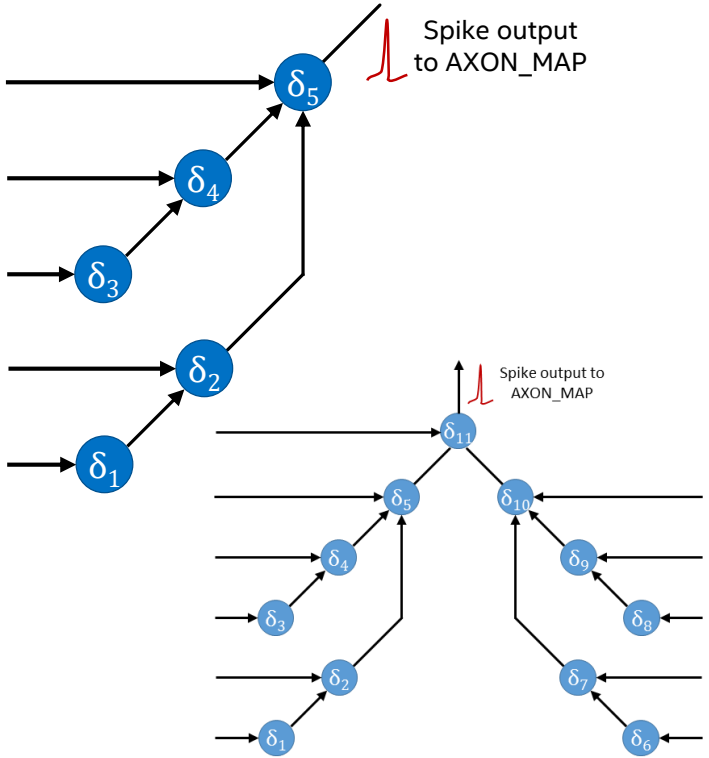


# Dendritic Compartments: Structural Model

**Compartment join operations**

```
0: (NOP)
1: (ADD_U)    U' = U+A+B
2: (MAX_U)    U' = max(U,A,B)
3: (MIN_U)    U' = min(U,A,B)
4: (PASS_U)   U' = A.S ? U+B : 0
5: (BLOCK_U)  U' = A.S ? 0 : U + B
6: (OR_S)     S' = A.S | B.S | S
7: (AND_S)    S' = A.S & B.S & S
```



# Min/Max Threshold Homeostasis

Loihi supports *intrinsic excitability homeostasis* (aka threshold adaptation)

**Dynamics:**

$$\Delta V_{th}(t) = \begin{cases} \beta(a(t) - a_{min}), & if\ a(t) < a_{min} \\ \beta(a(t) - a_{max}), & if\ a(t) > a_{max} \end{cases}$$
$$V_{th}(t) = V_{th}(t - T_{epoch}) + \Delta V_{th}(t)$$

(in terms of neuron's *activity trace* a(t))

Evaluated periodically every *Dendritic epoch*.
(Usually set the same as the learning epoch)

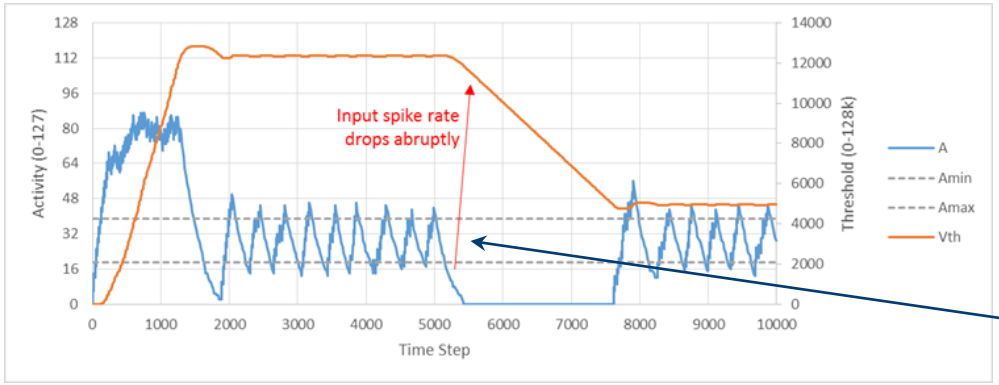
**Parameters:**

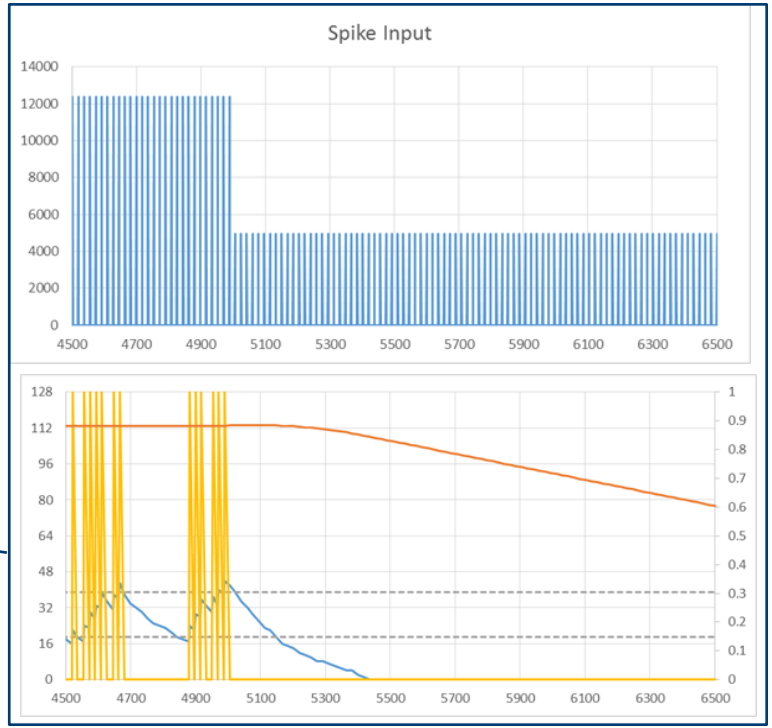
| Parameter | Bits | Definition |
|---|---|---|
| $a_{max}$ | 7 | Maximum activity level above which Vth will be raised. |
| $a_{min}$ | 7 | Minimum activity level, below which Vth will be lowered. |
| $\beta$ | 4 | Scaling constant relating activity trace differences to threshold changes. |

# Example Homeostasis Dynamics

**Neuron with abrupt input rate change**

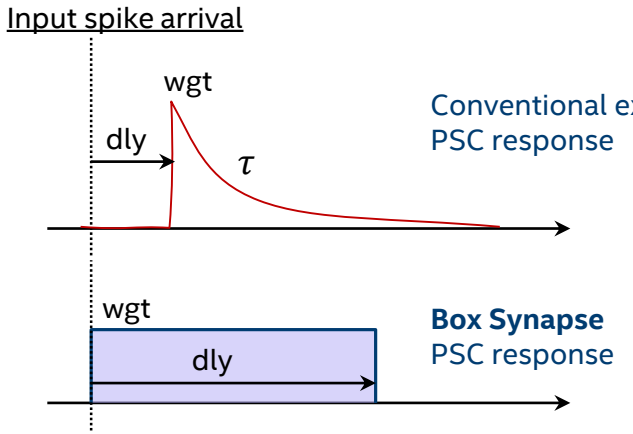
Synaptic input drops abruptly at t=5000.



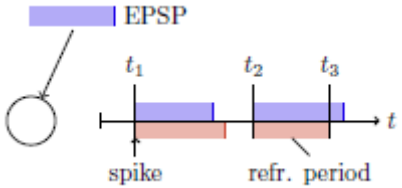



# Other Synaptic Features

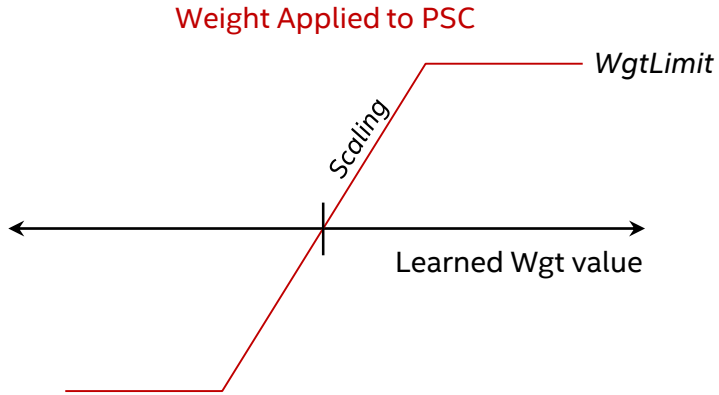
## Box Synapses




Example usage for CSP with stochastic SNNs, e.g.
*Jonke, Habenschuss , Maass 2016*



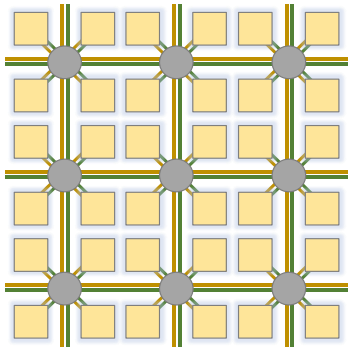

## Weight Scaling




$$\text{EffectiveWgt} = \text{sgn}(\text{Wgt}) \cdot \min(WgtLimit, |\text{Wgt}| \cdot 2^{Scaling})$$

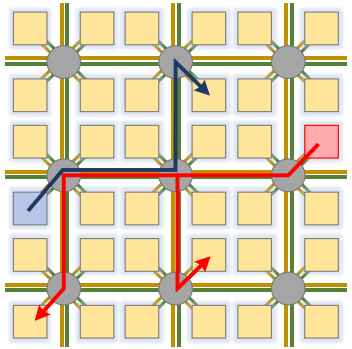
When enabled, may think of the learned weight as a *permanence* value.

# Mesh Operation: Fine-Grained Synchronization



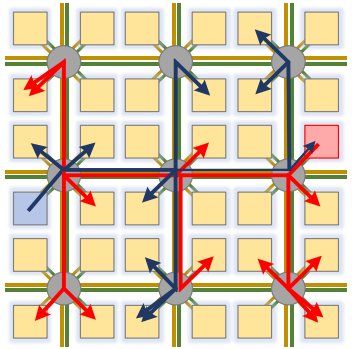
Time step T begins.

Cores update dynamic neuron state and evaluate firing thresholds

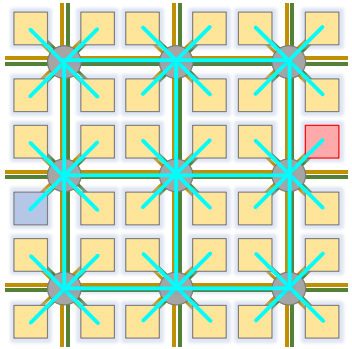


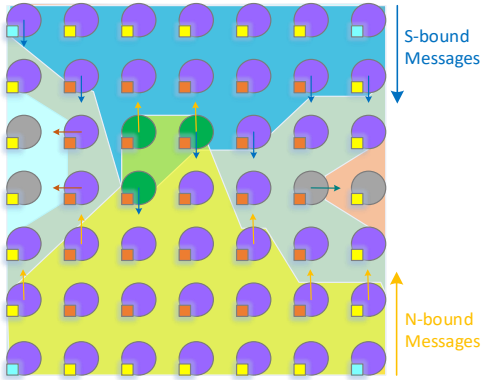
Above-threshold neurons send spike messages to fanout cores

(Two neuron firings shown.)



All neurons that fire in time T route their spike messages to all destination cores.



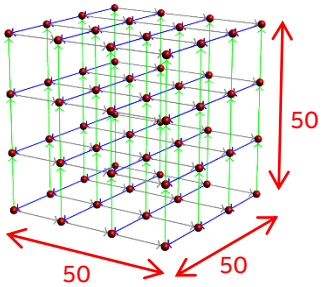


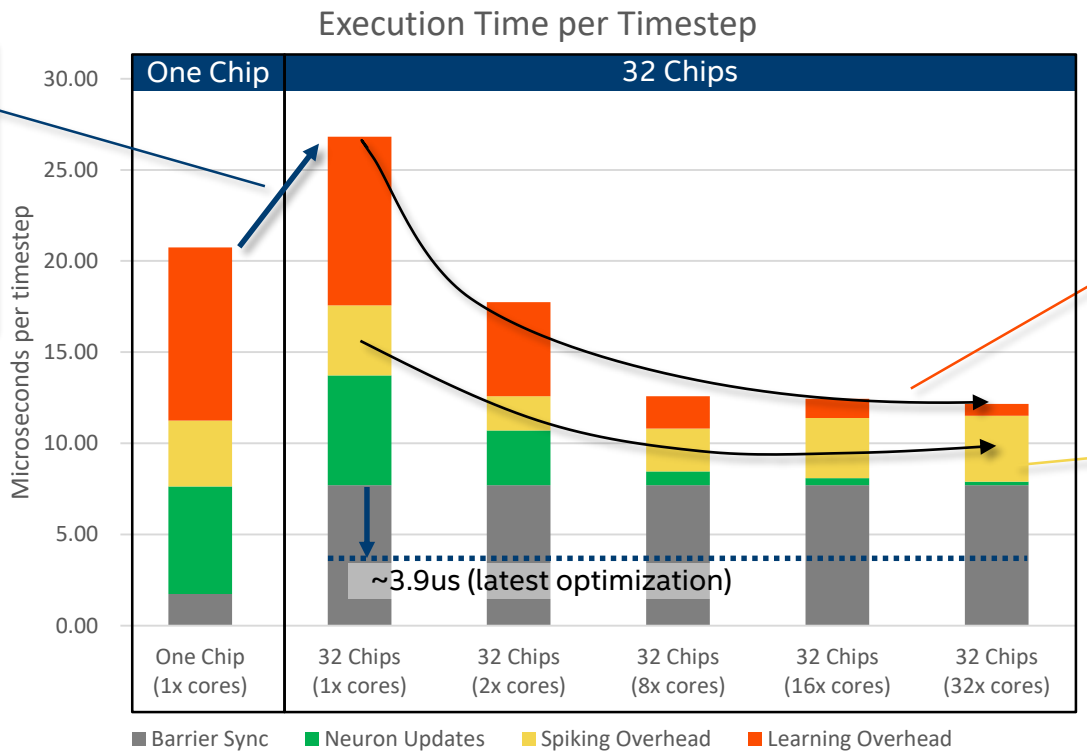

# Exploring Mesh Scaling to 32 Chips

## Graph Search on Nahuku (32-chip Loihi System)

Fixed 128-way core parallelism. Slowdown due to **increased barrier sync time** over 32 chips vs 1 chip

50x50x50 3D lattice




Increasing **core parallelism** with **fixed chip count**

Learning overhead **decreases** with increasing core parallelism

Spike overhead decreases, then **increases** with increasing core parallelism

Performance results are based on testing as of March 2019 and may not reflect all publicly available security updates. No product can be absolutely secure.


intel®