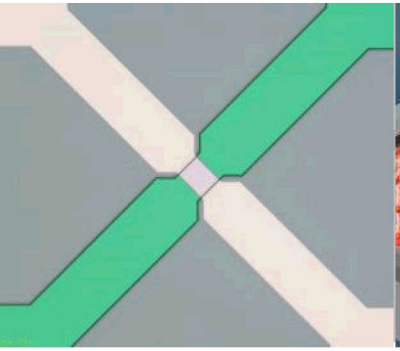
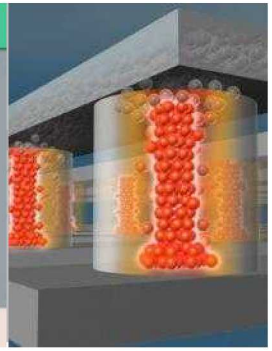
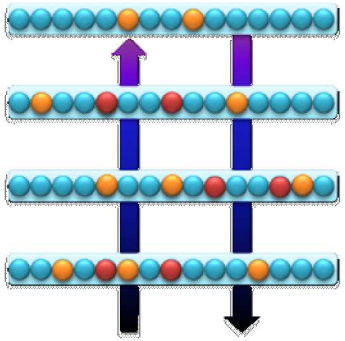
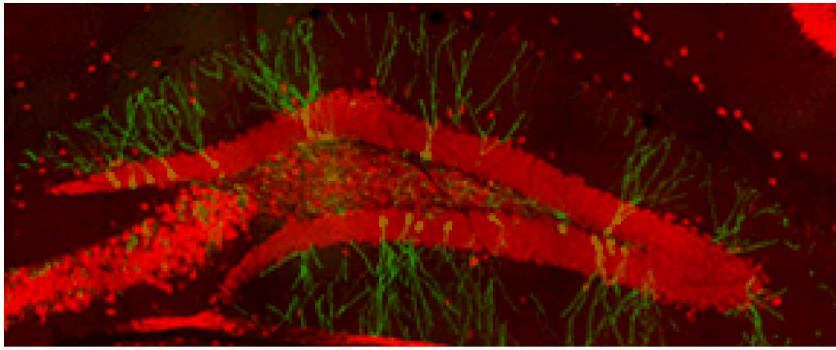
SAND2019-8753PE

Neuromorphic Computing Algorithms and Architecture Research at Sandia

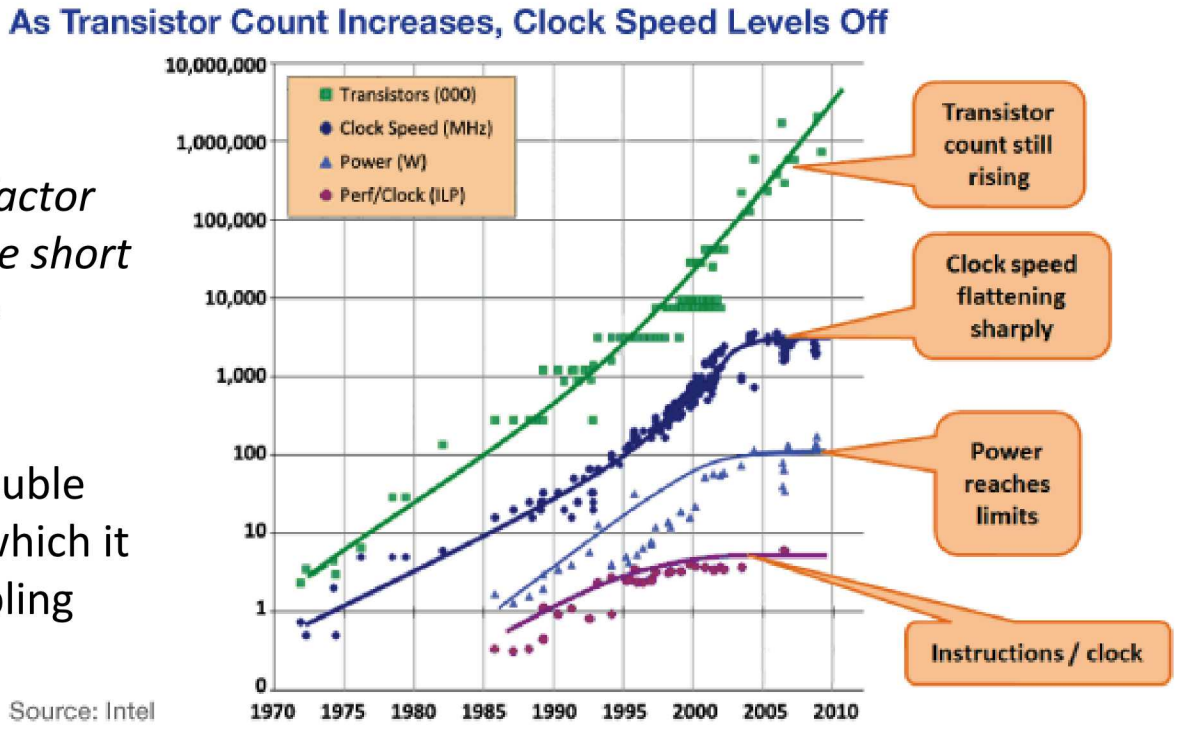
Aaron J. Hill[1], Jonathon W. Donaldson[1], Fredrick H. Rothganger, Craig M. Vineyard[1], David R. Follett[2], Pamela L. Follett[2,3], Michael R. Smith[1], Stephen J. Verzi[1], William Severa[1], Felix Wang[1], James B. Aimone[1], John H. Naegle[1], and Conrad D. James[1]

[1]Sandia National Laboratories, [2]Lewis Rhodes Labs, [3]Tufts University

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Motivation-Moore's Law

1965-1975: *The complexity for minimum component costs has increased at a rate of roughly a factor of two per year. Certainly over the short term this rate can be expected to continue, if not to increase*
Revision: 1975: Semiconductor complexity would continue to double annually until about 1980 after which it would decrease to a rate of doubling approximately every two years

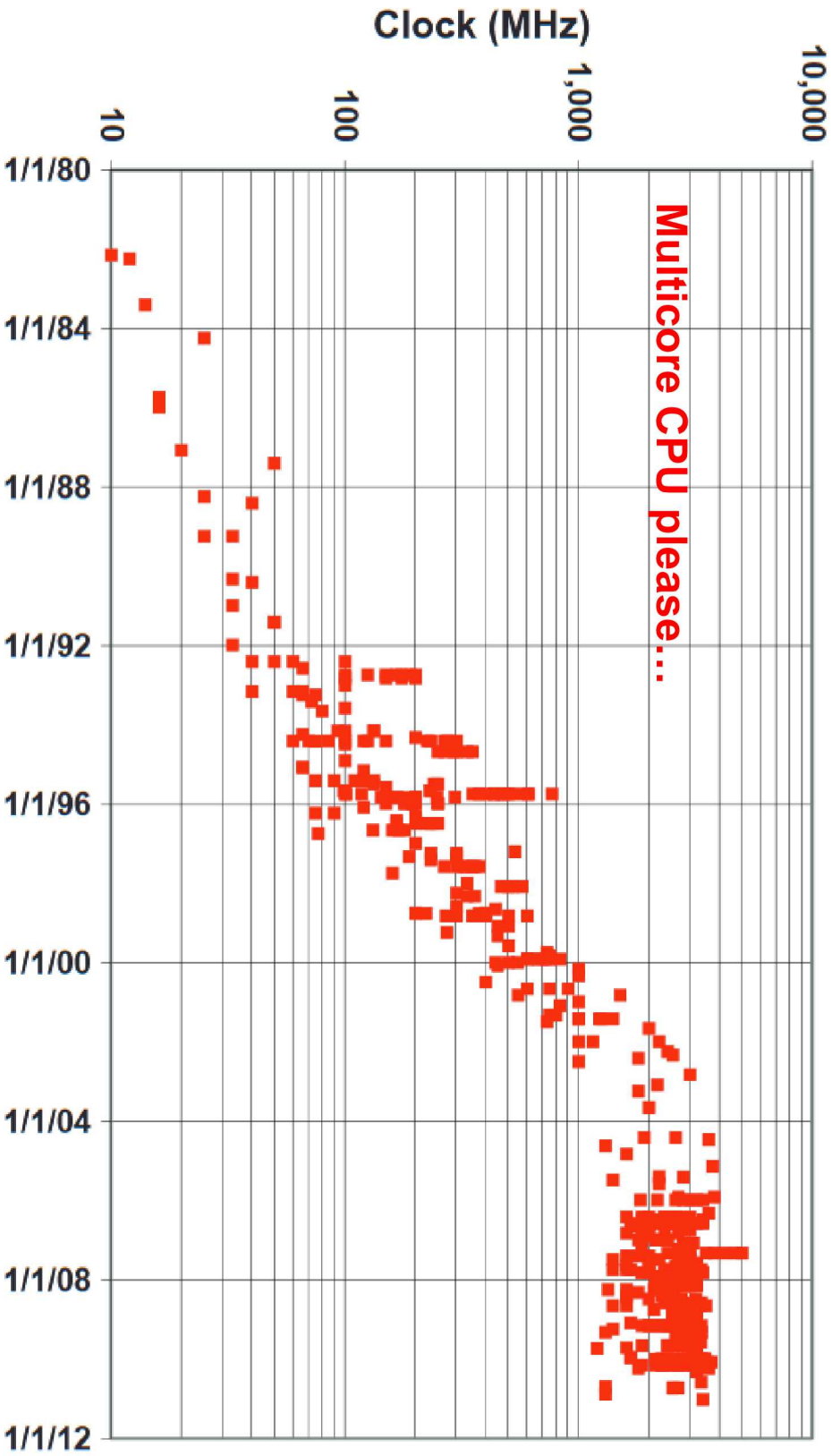


David House, an Intel colleague, perpetuated the misconception of Moore's Law with performance doubles every 18 months.

Self-fulfilling prophecy, an industry driver.

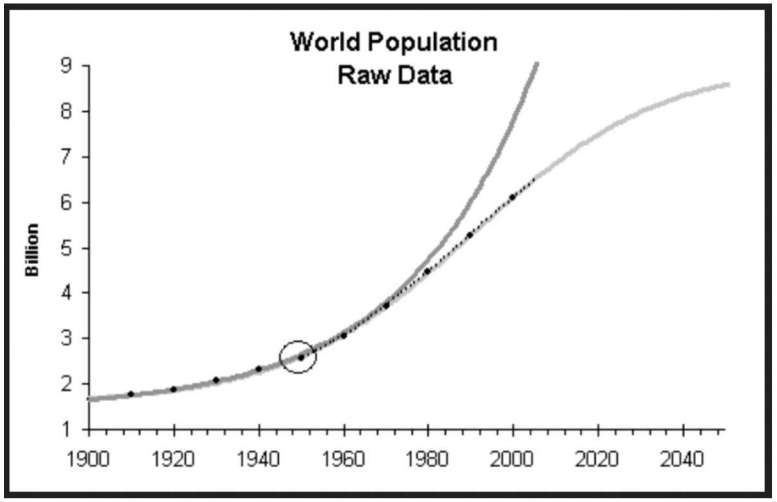
Dennard Scaling

- Why haven't clock speeds increased, even though transistors have continued to shrink?
- $P = \quad fCV^2 + VI_{leakage}$
- Capacitance falls with feature size
- as the size of the transistors shrink, and the voltage reduced, circuits can operate at higher frequencies at the same power
- As transistors get smaller, power density increases because these don't scale with size
- This created a "Power Wall" that has limited practical processor frequency to around 4 GHz since 2006 (65nm node)
- There is a general industry consensus that the laws of Dennard scaling broke down somewhere between 2005-2007

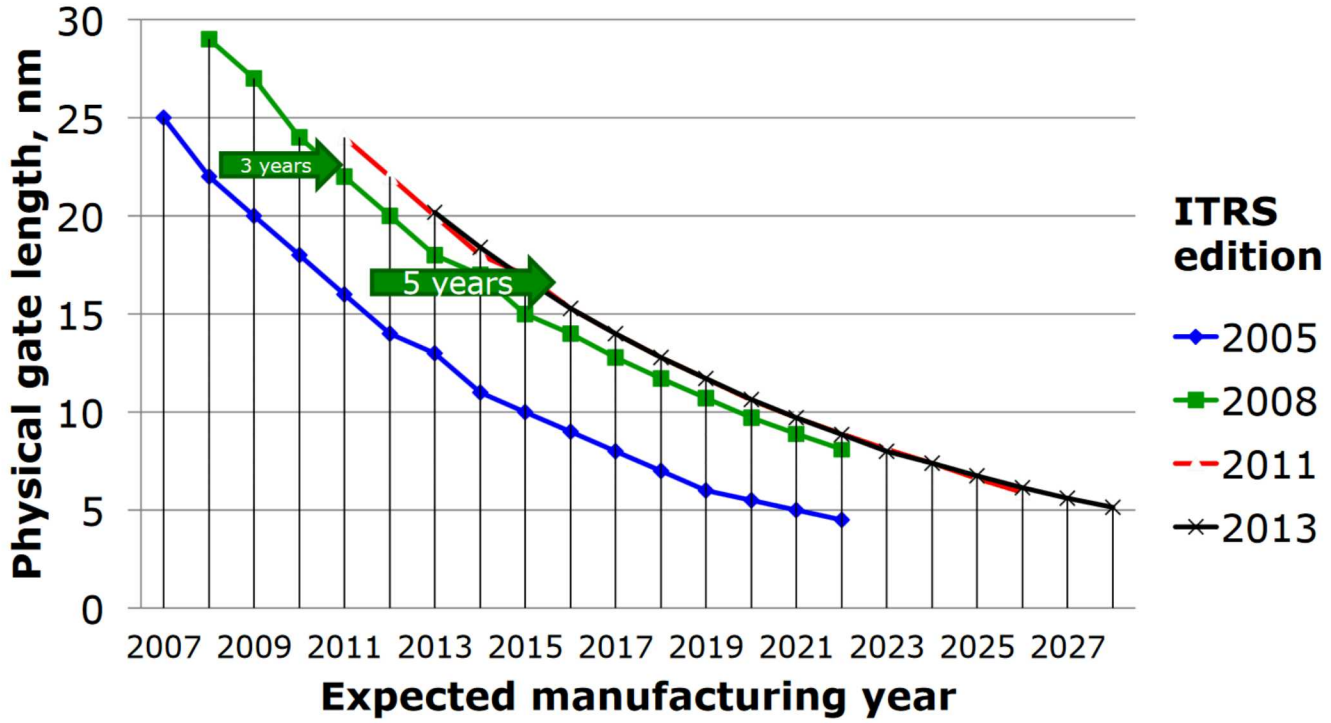

Clock (MHz)
10
100
1,000
10,000
1/1/80
1/1/84
1/1/88
1/1/92
1/1/96
1/1/00
1/1/04
1/1/08
1/1/12
Multicore CPU please…

The Singularity is Nigh!

- In 2005, author Ray Kurzweil published *The Singularity Is Near: When Humans Transcend Biology*
- Based on “Moore’s Law” artificial computation will surpass the human brain by ~2045.
- In **"The Singularity Myth,"** physicist Theodore Modis illustrates that in countless real world examples, growth actually follows a logistic "S-curve.“, which initially appears to be exponential.



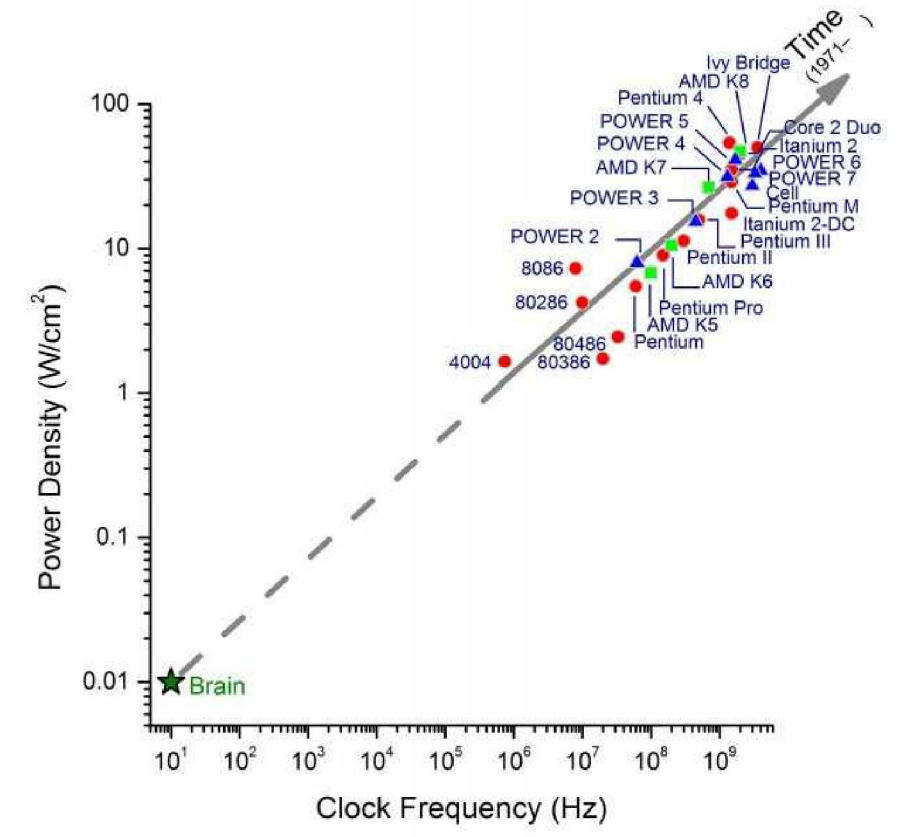
Semiconductor Roadmap Predictions



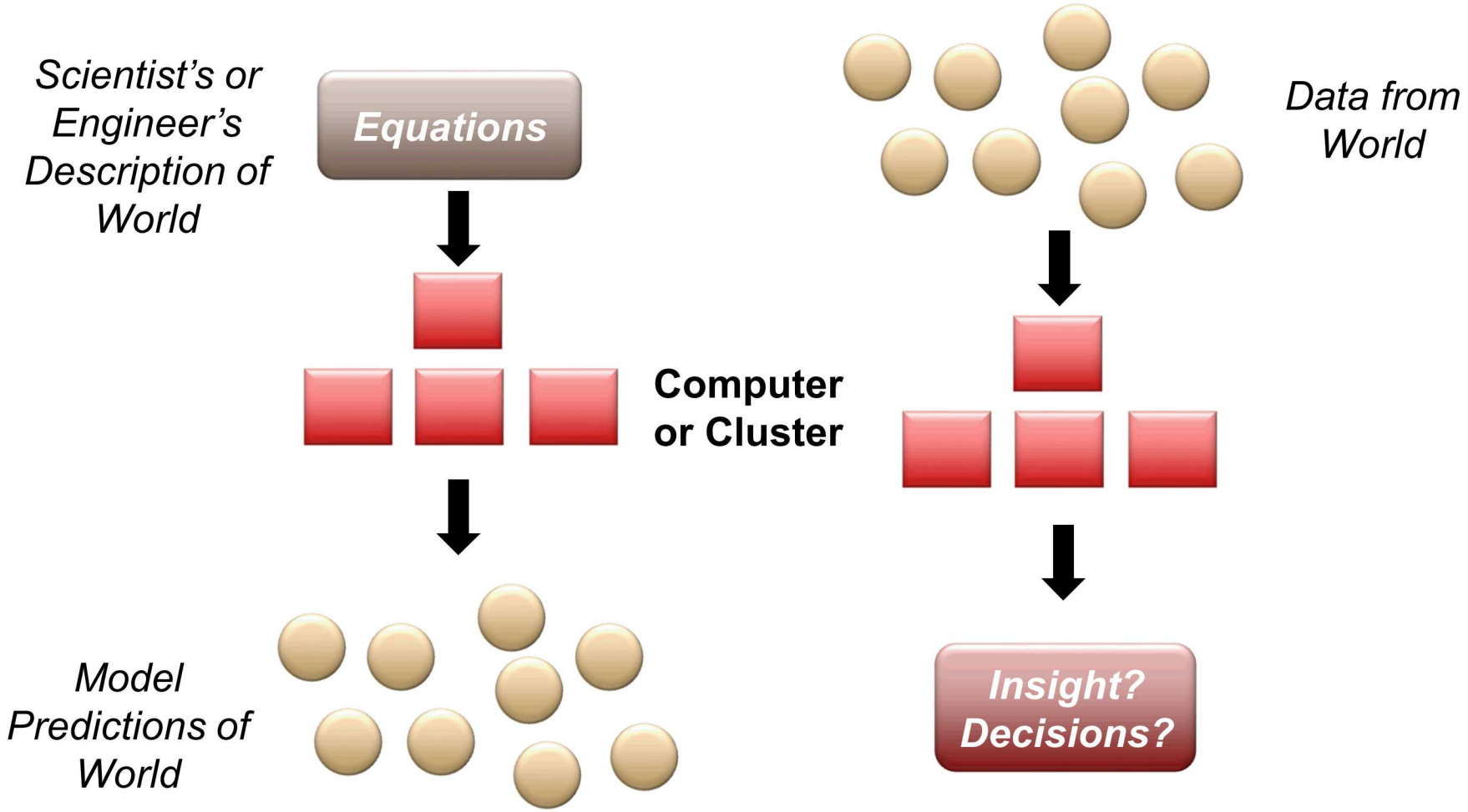
- ~2016 industry shifted from using semiconductor scaling as a driver to more of a focus on meeting the needs of **major computing applications. GPU please…**

Different Computing Paradigm

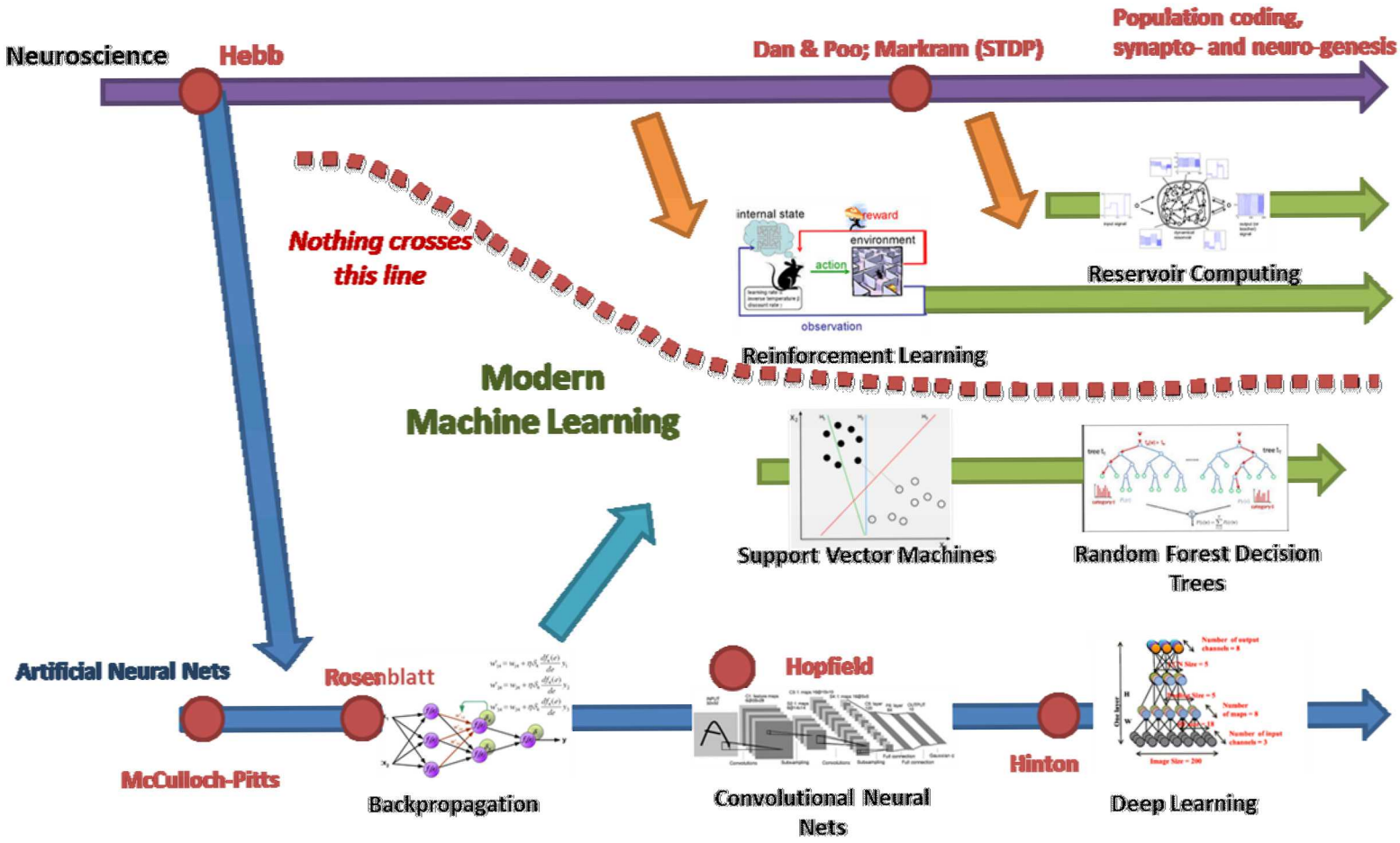
- Over last decade, different aspects of Moore's Law have stalled
- Last decade has seen emphasis on software, but new hardware paradigms are being sought
- Neural networks
  - Biological neural networks operate in parallel
  - Very low energy



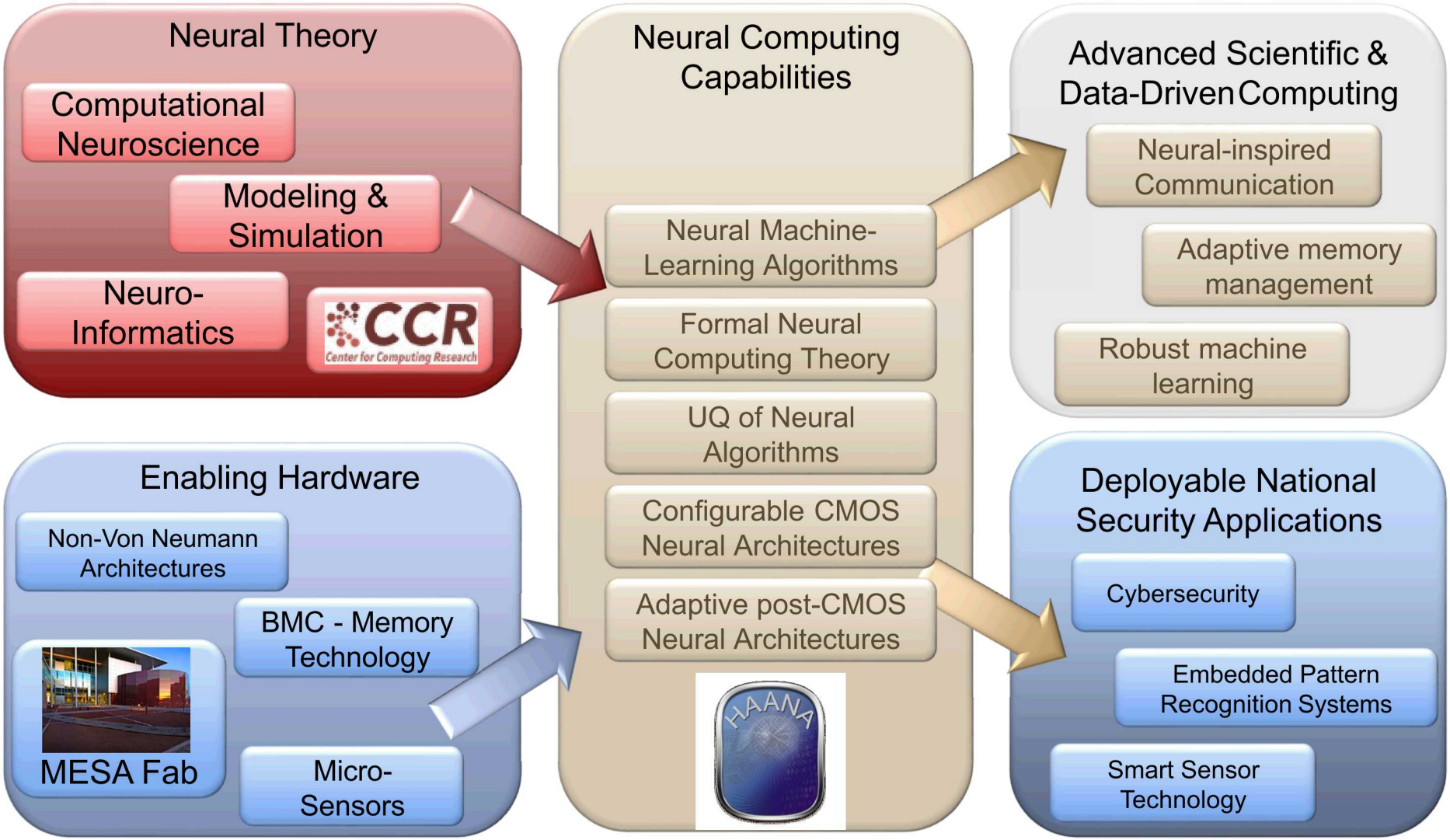
Brain Inspired Computing



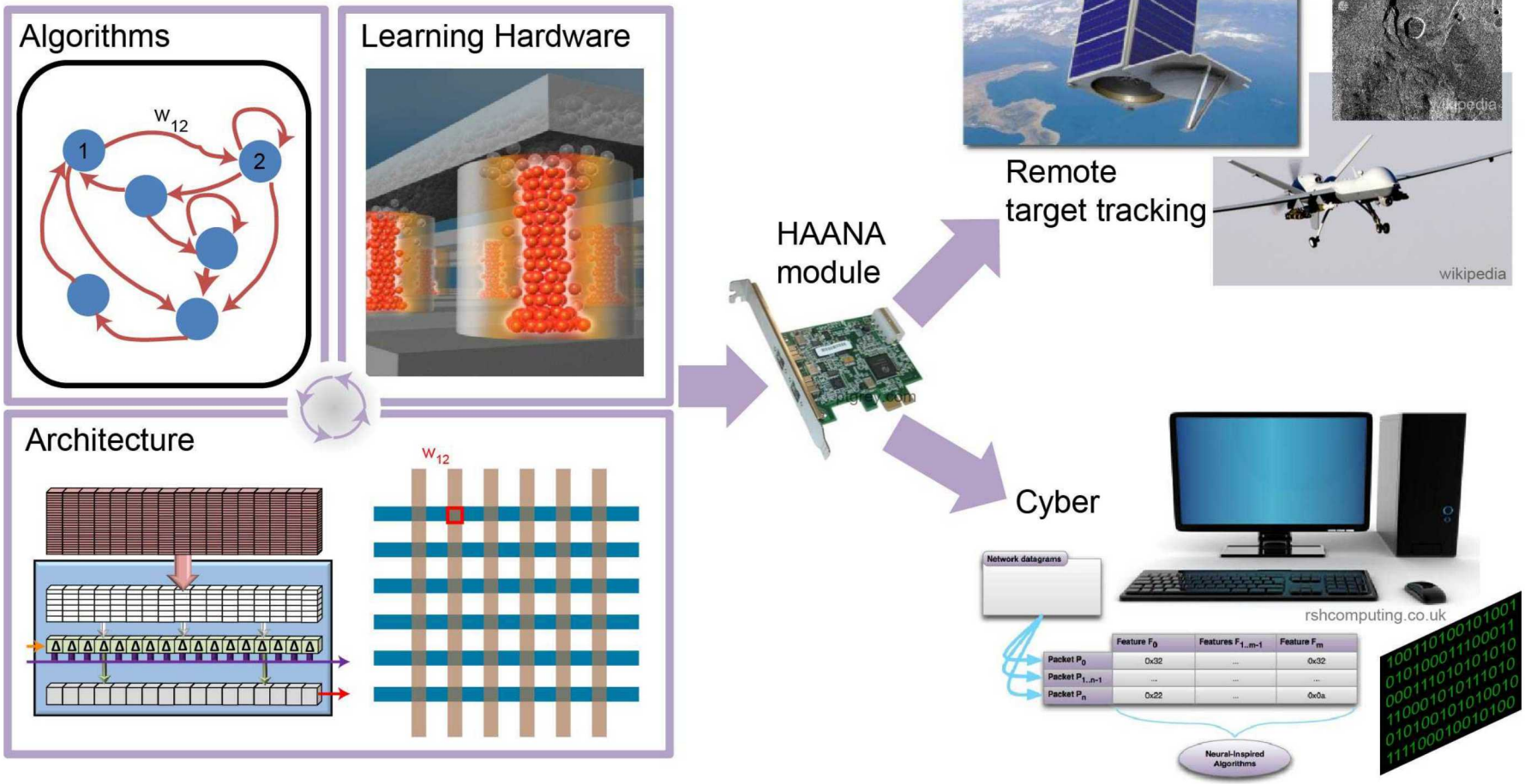
Neural computation and artificial neural networks are not the same



Neural computing at Sandia Labs leverages a large research foundation

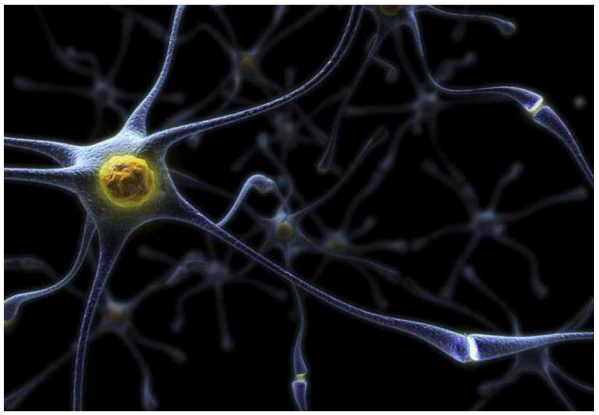


Hardware Acceleration of Adaptive Neural Algorithms (HAANA)

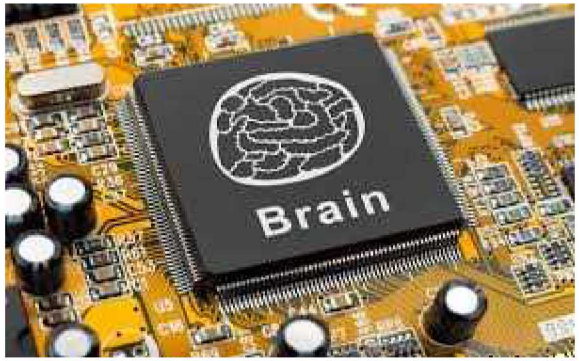


Challenges for Neuromorphic Hardware

- Inspirations from neurobiology
  - Low computational power (femtojoules)
  - Large population of neurons (10^{10})
  - Real time learning
  - High degree of fan-in (10^4)
  - Temporally coded information

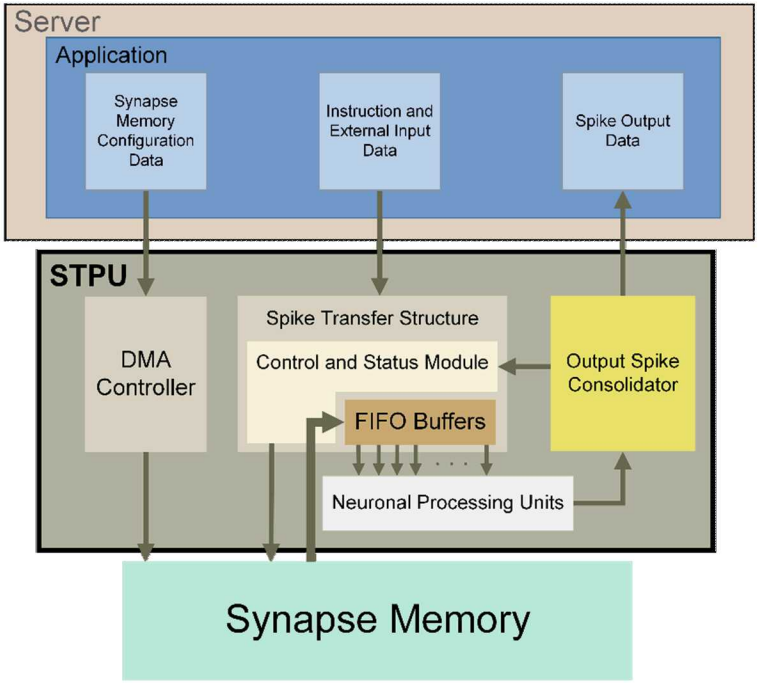


- Manifestations of neurobiological-inspired hardware tend to deviate from the original inspirations due to constraints in
  - Device physics
  - Fabrication technologies
  - Readily available hardware
- It's difficult to get everything in one package



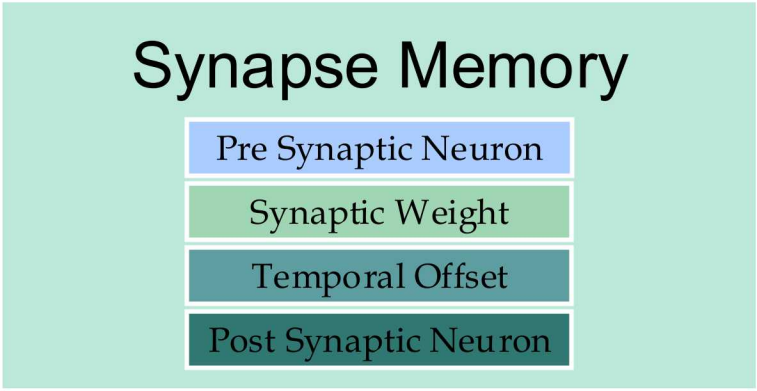
The Spiking Temporal Processing Unit (STPU) Overview

- A spiking neuromorphic research architecture
  - Scalable and highly parallel
  - Prioritizes temporal and spatial complexity of spiking neural systems
  - Supports high fidelity spike timing dynamics
- Simple Leaky Integrate and Fire (LIF) neuron model with 3 parameters
  - Spiking threshold
  - Minimum neuron potential
  - Leak rate
- Temporal buffer for synaptic delays
- Supports arbitrarily connected networks with configurable weights and delays per synapse

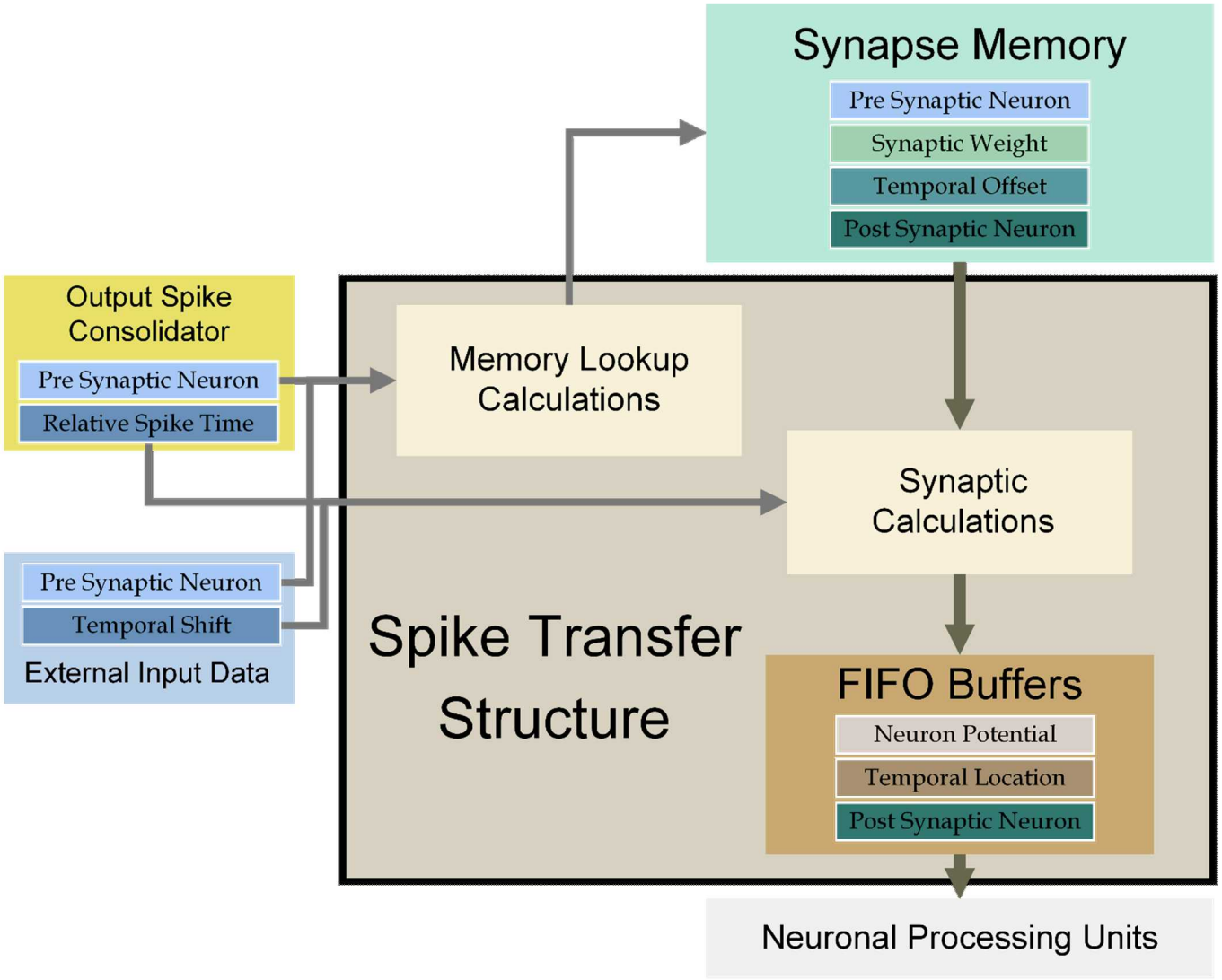


Synapse Memory

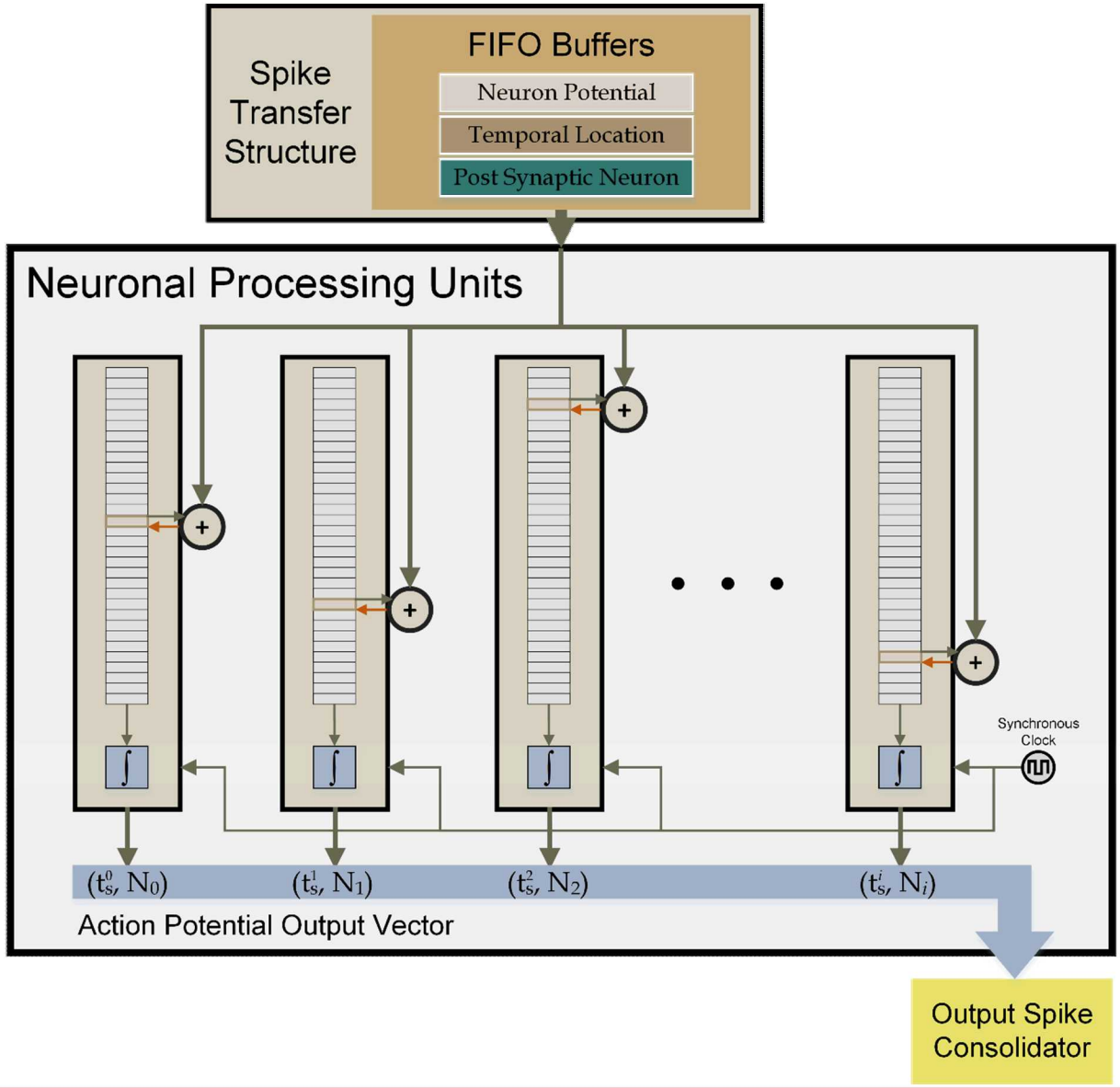
- Large off chip memory
- Contains all synaptic information
  - Pre-synaptic neuron designator
  - Synaptic weight
  - Synaptic delay (temporal offset)
  - Post-synaptic neuron designator
- Enough memory to support fully connected topologies
- Unique weights and delays per synapse
- Online/Real-time updating allowed



Spike Transfer Structure



Neuronal Processing Unit



Neuronal Computation Dynamics

Synaptic Input Update

$$\hat{R}_{d,j}(t) = R_{d,j}(t) + \sum_k i_k(t) W_{d,j,k}(t) \tag{1}$$

Temporal Integration

$$\begin{aligned} R_{\bar{d},j}(t+1) &= \hat{R}_{\bar{d}+1,j}(t) \\ R_{D-1,j}(t+1) &= 0 \\ \bar{A}_j &= A_j(t) + \hat{R}_{0,j}(t) \end{aligned} \tag{2}$$

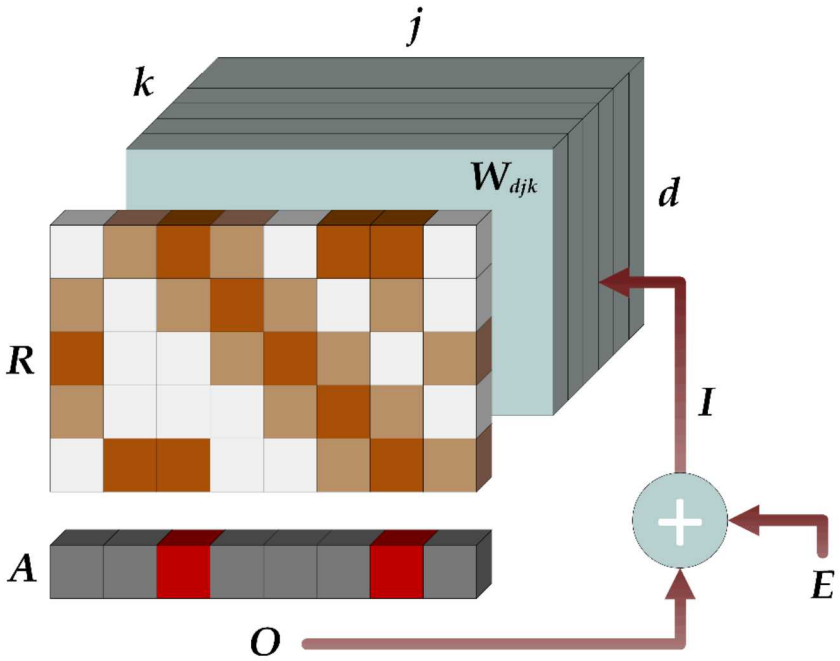
where $\bar{d} \in \{0, \cdots, D-2\}$.

Threshold, Fire, Reset

$$A_j(t+1) = \begin{cases} \Lambda(\bar{A}_j) & \text{if } \bar{A}_j < T_j \\ 0 & \text{if } \bar{A}_j \geq T_j \rightarrow \text{SPIKE} \end{cases} \tag{3}$$

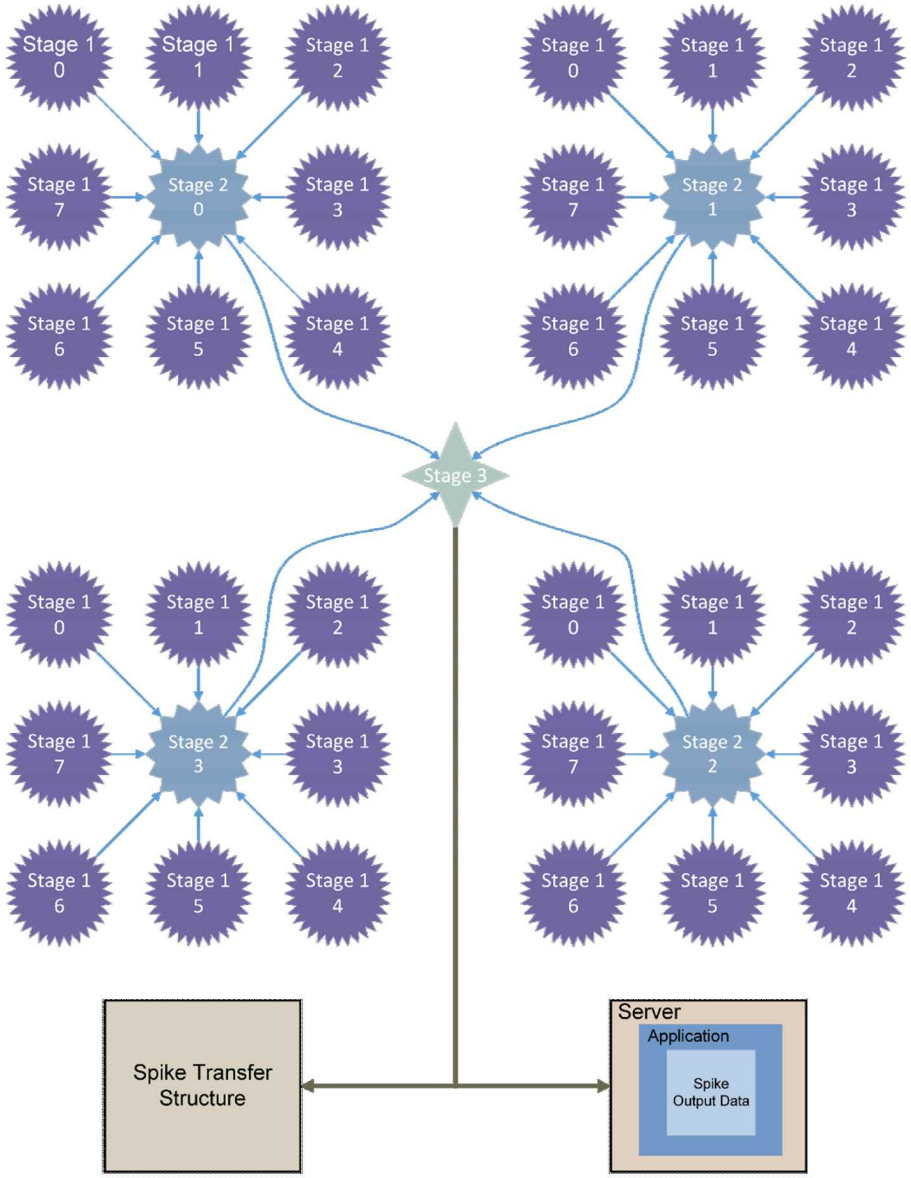
$$\Lambda(\bar{A}_j) = \bar{A}_j(1-\lambda)$$

where $A_j(0) = 0$. Note that $A_j(t+1)$ is the value of the integration register after a complete LIFop has been performed, T_j is the threshold to fire value of the j^{th} neuron, and λ is the leakage value.



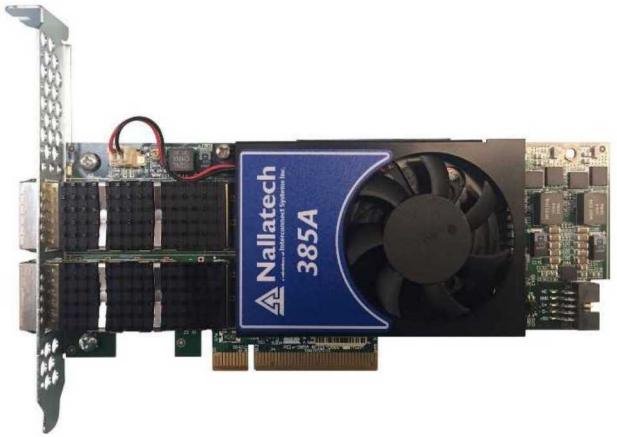
Output Spike Consolidation

- Three-stage consolidator
- Pipelined efficiency
- Each stage operates in parallel
- Number of stage 1, 2, and 3 consolidators is a parameter
- Number of stages is a parameter



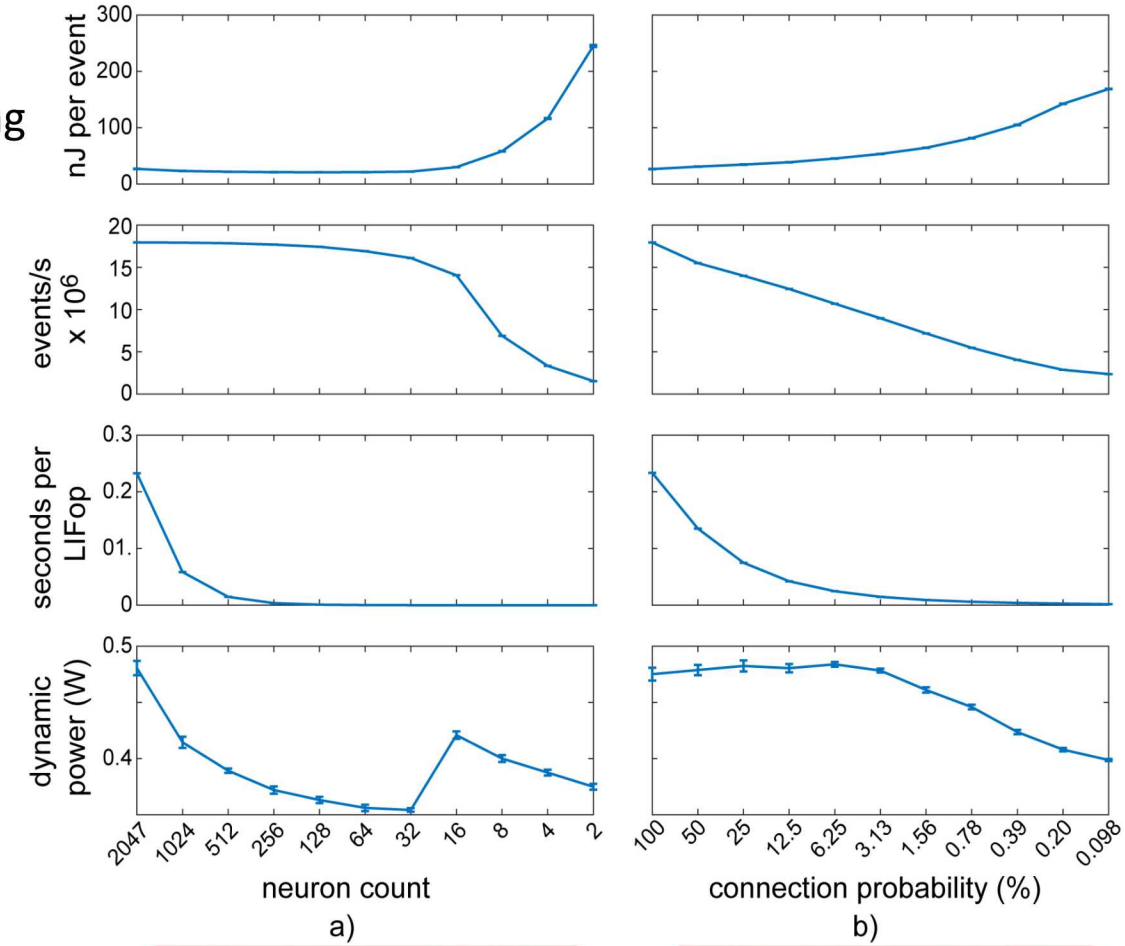
Hardware Development Environment

- Current implementation on a Nallatech 385A
  - 8-lane PCIe 3.0 card
  - Intel® Arria® 10 GX 1150 FPGA
  - Two 4GB banks of DDR3L SDRAM @ 2133 MT/s
- 2048 Instantiated neurons (4096 max)
- 32 deep temporal buffer (64 max)
- 16 parallel computation paths
- 16MB of Synapse Memory
  - 18 bits per synaptic weight
  - 6 bits per synaptic delay
  - 8 bits for post-synaptic neuron
  - 2048×2048 total synapses



Power Measurements on Nallatech 385

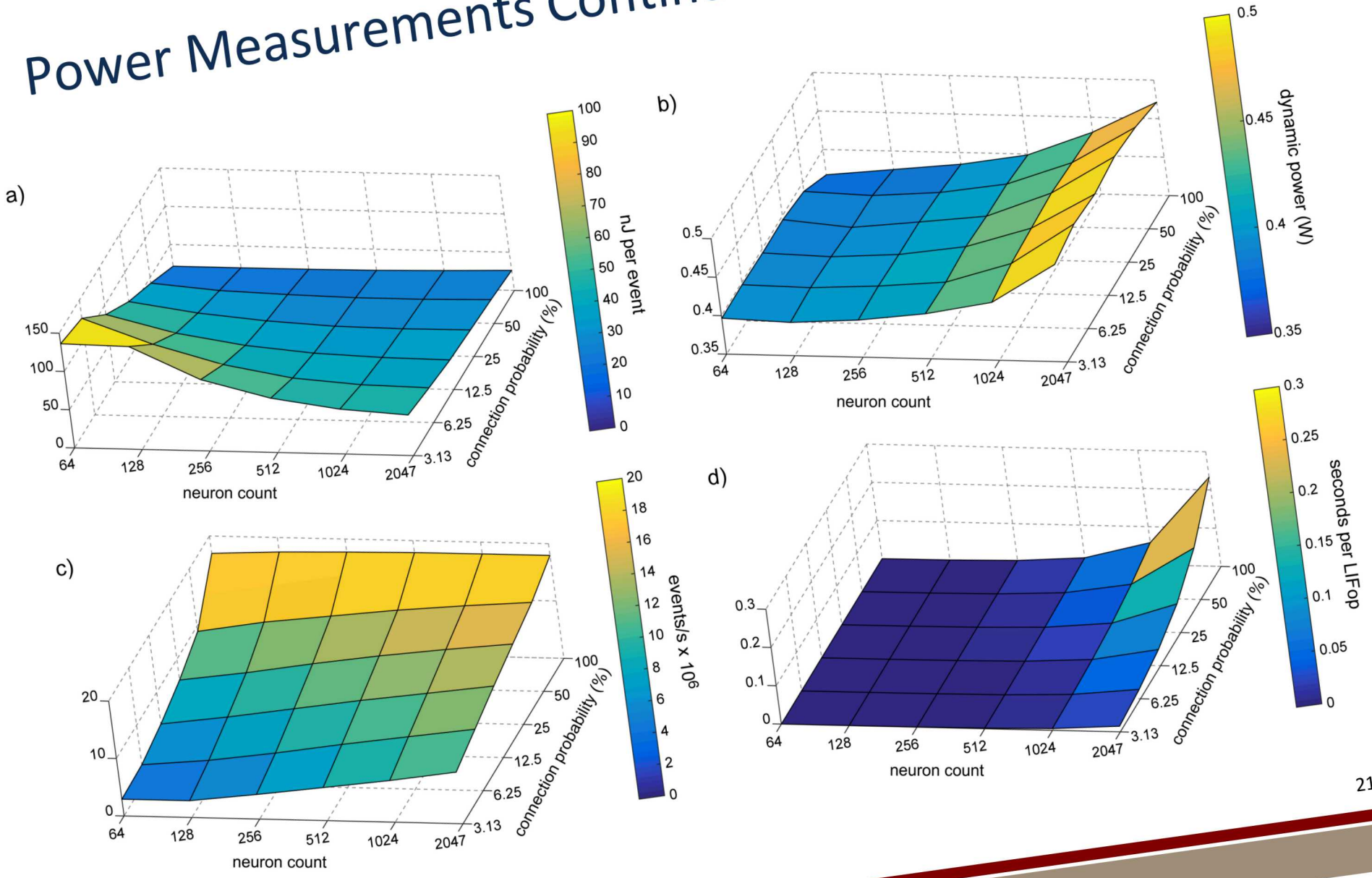
- Testing Setup
  - DC Power Analyzer measuring voltage and current simultaneously
  - Only dynamic power was considered
  - Every neuron spikes every cycle
  - Total synaptic event count is constant in all experiments
- (a) is a fully connected network, neuron count decreases
- (b) neuron count is constant and connectivity decreases



~ 21nJ per event for 2048 to 32 neurons

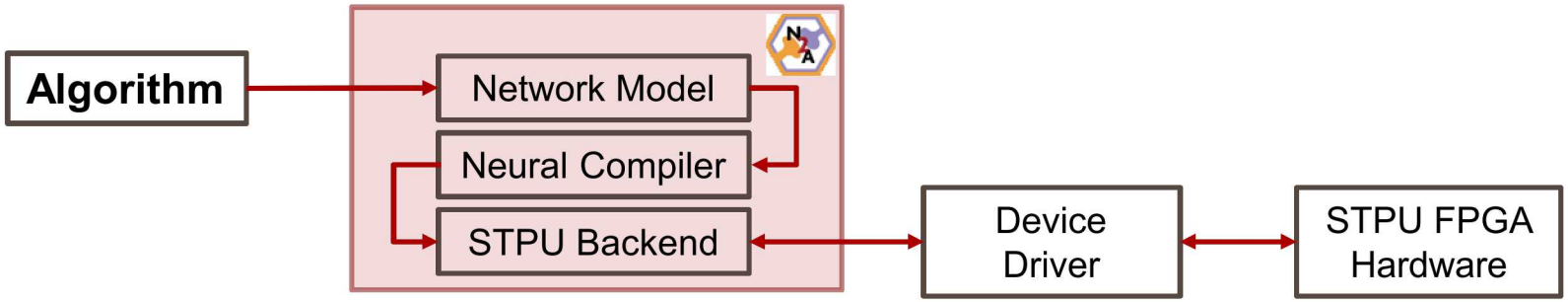
Performance favors dense connectivity

Power Measurements Continued



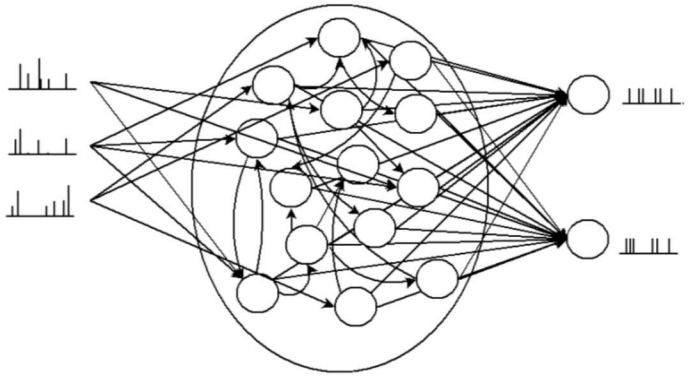
Neurons to Algorithms

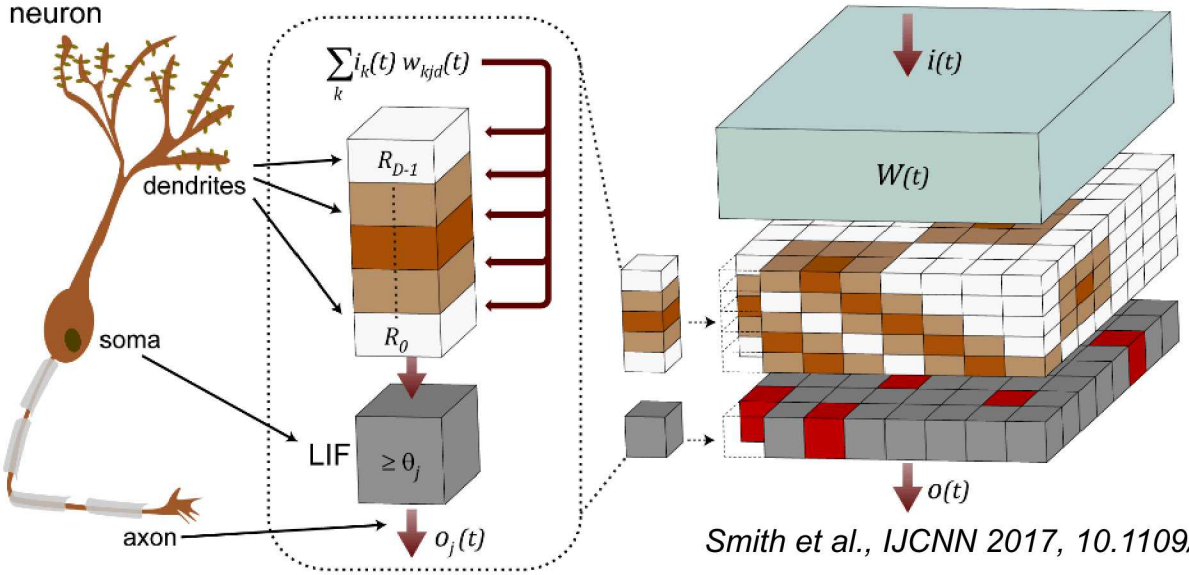
- A language for modeling neural systems
- Object-oriented modeling – seamless scaling
  - Declarative language, not procedural
  - Neural components (parts) are bundles of equations and parameters
  - Parts are inheritable and extendable
- Backend designed for the STPU
  - Translates N2A models to configuration information for the FPGA
  - Utilizes a user space library to send the configuration data
  - Executes the algorithm, loads input data, and receives output data



Liquid State Machine (LSM)

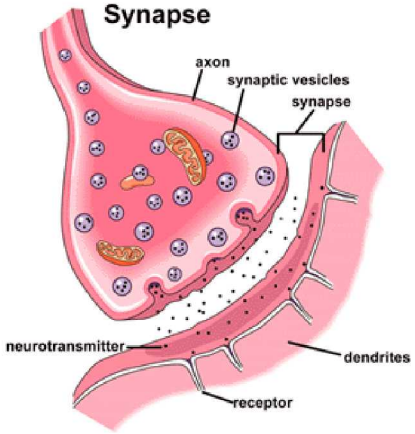
- Algorithm for learning from sequential (or temporal) data
- Randomly connected spiking neurons encode complex temporal dynamics
- Temporal dynamics map well to the STPU architecture

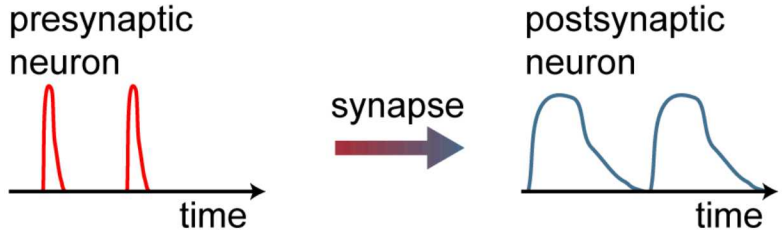




Smith et al., IJCNN 2017, 10.1109/IJCNN.2017.7966150

Synaptic Response Functions



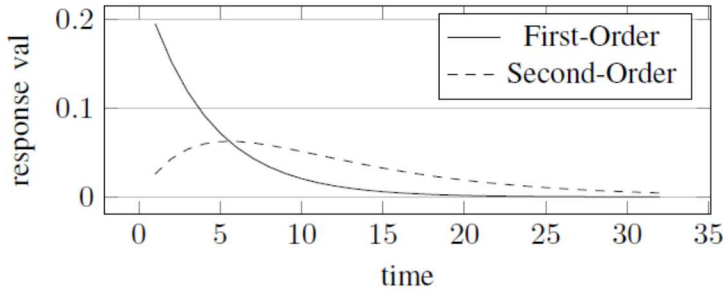


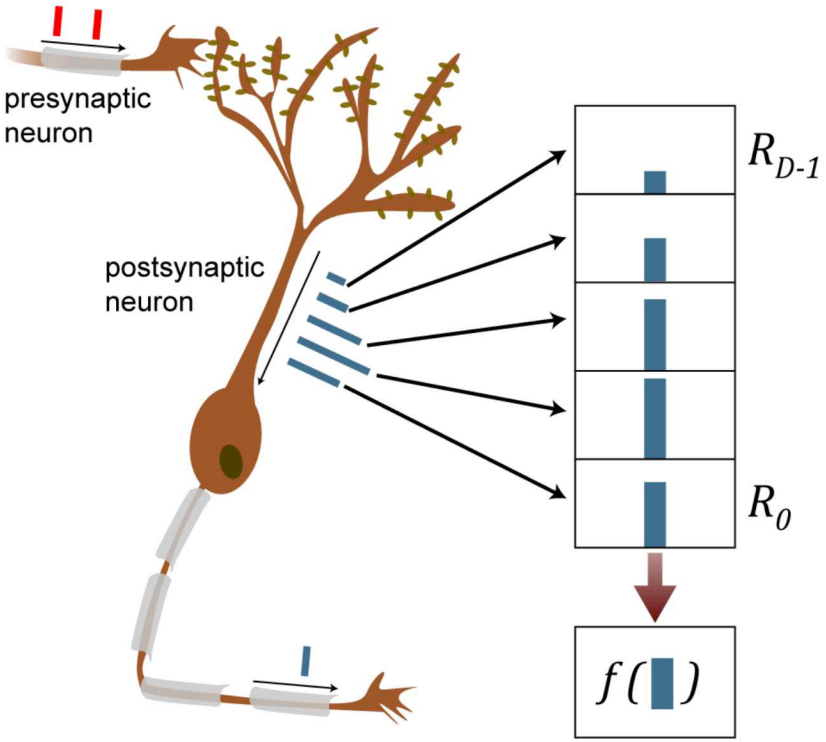
First-order response

$$\frac{1}{\tau_m} e^{-\frac{t-t_{ij}-d_i}{\tau^s}} \cdot H(t - t_{kd} - \Delta_{kd})$$

Second-order response

$$\frac{1}{\tau_1^s - \tau_2^s}\left(e^{-\frac{t-t_{kd}-\Delta_{kd}}{\tau_1^s}} - e^{-\frac{t-t_{kd}-\Delta_{kd}}{\tau_2^s}}\right) \cdot H(t - t_{kd} - \Delta_{kd})$$





Smith et al., IJCNN 2017, 10.1109/IJCNN.2017.7966150

Effects of Parameter Selection and Synaptic Response Function

Linear Model	3 X 3 X 15 $\theta_j = 15$	5 X 5 X 5 $\theta_j = 11$	4 X 5 X 10 $\theta_j = 15$	2 X 2 X 20 $\theta_j = 10$
Linear SVM	0.906	0.900	0.900	0.914
LDA	0.921	0.922	0.922	**0.946**
Ridge Regress	0.745	0.717	0.717	0.897
Logistic Regress	0.431	0.254	0.254	0.815

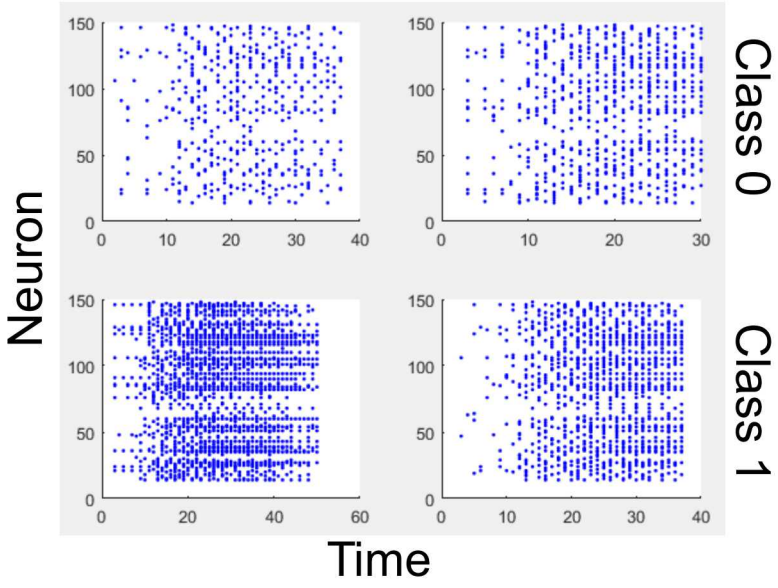
Synaptic Response	Train Sep	Train Rate	Test Sep	Test Rate	SVM
Dirac Delta	0.129	0.931	0.139	0.931	0.650
First-Order	0.251	0.845	0.277	0.845	0.797
Second-Order	**0.263**	**0.261**	**0.290**	**0.255**	**0.868**
First-Order 30	**0.352**	0.689	**0.389**	0.688	0.811
First-Order 40	0.293	0.314	0.337	0.314	0.817
First-Order 50	0.129	0.138	0.134	0.138	0.725

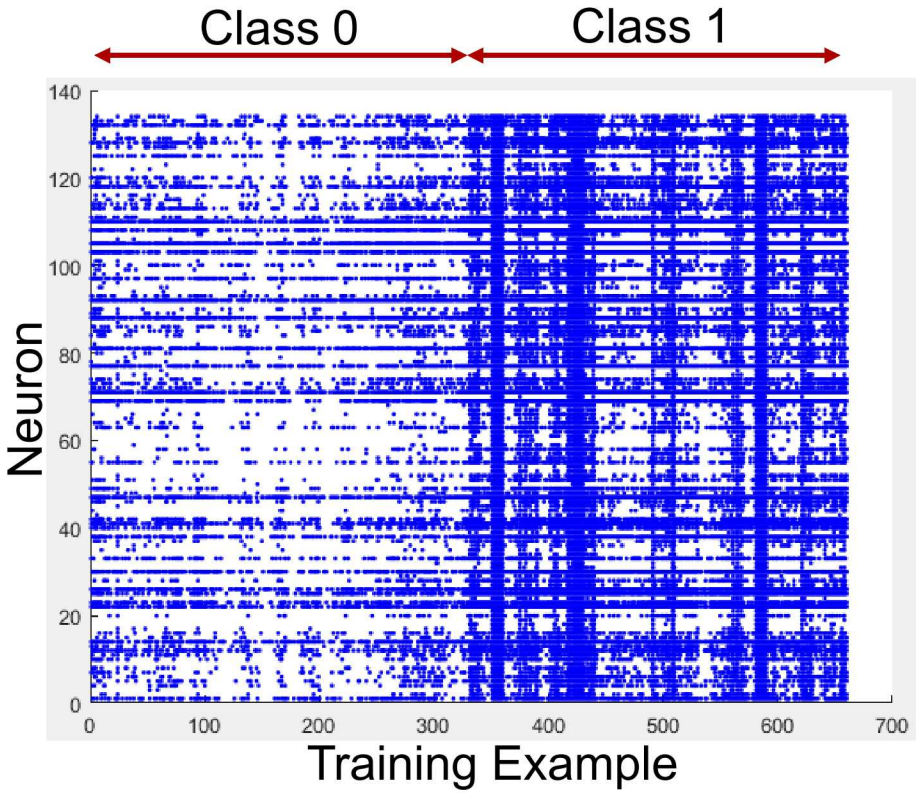
Red indicates the best values for default parameters

Blue indicates values that improved over second-order

STPU Results of LSM

- 220 test examples, 110 zeros and 110 ones.
  - 87.3% accuracy on zeros
  - 84.6% accuracy on ones
- Distinct separation in the data between 0s and 1s.

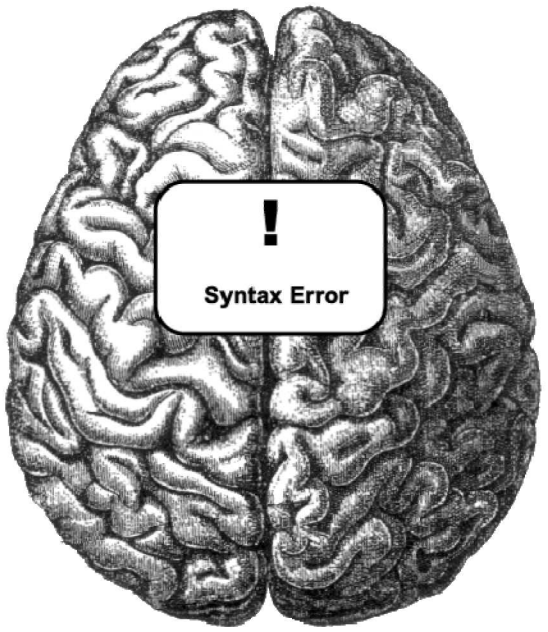


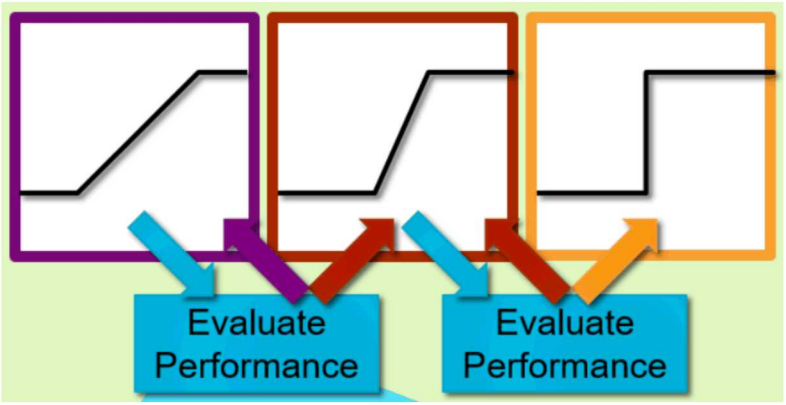


Final liquid state of all 660 training examples. 330 examples of class 0 and 330 examples of class one.

Porting Algorithms to Neuromorphic Platforms

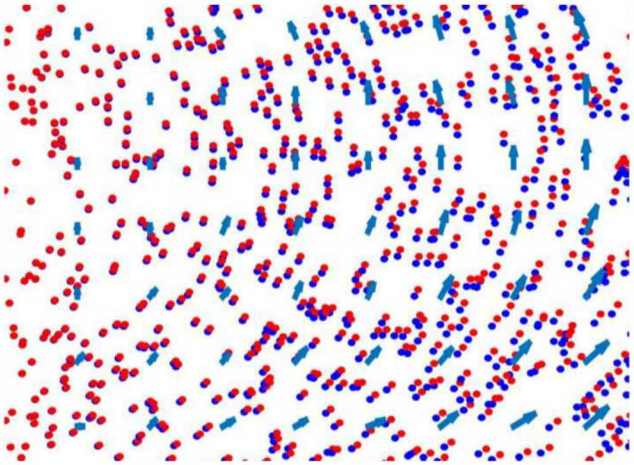
- Classical algorithms are tried-and-tested
- Neuromorphic platforms must meet and exceed classical results
- Neuromorphic has been cornered into learning based algorithms only
- View neurons as highly parallel and simple processors
  - Min, Max, Sorting, Optimization, and Filtering
  - Matrix multiplication
  - Cross-correlation with application to Particle Image Velocimetry
  - Random Walk with application to the diffusion equation
- Whetstone: A general ANN to SNN conversion tool
  - A process for training binary, threshold-activation SNNs using existing deep learning methods
  - Conversion introduces minimal loss in accuracy.





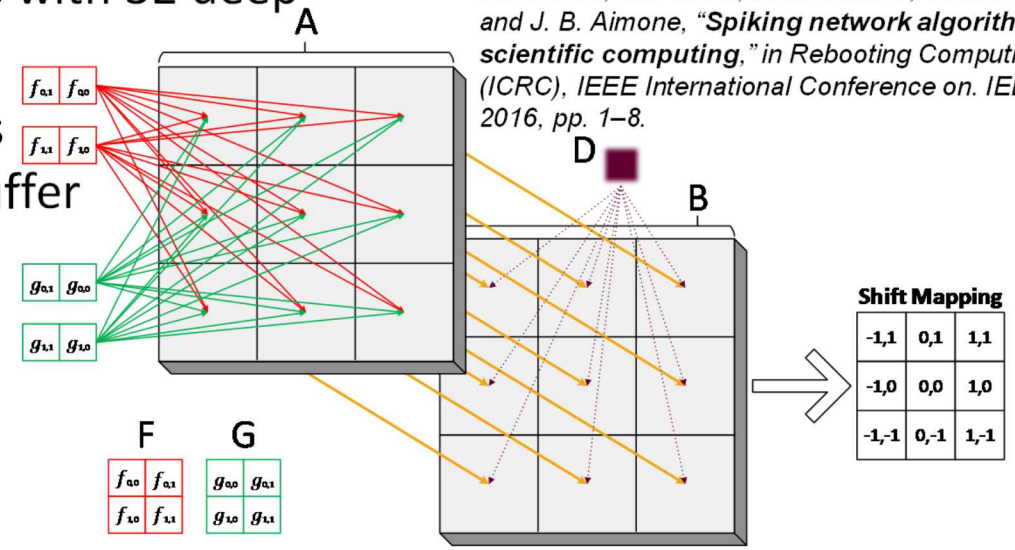
Particle Image Velocimetry

- Method for using particles in imagery to determine the local velocity flow
- Underlining kernel: Cross-Correlation
  - Compute Argmax $(f \star g)(n) = \sum_m f(n)g(m+n)$
- With the temporally coded algorithm
  - $13 \times 13 \rightarrow 1767$ neurons with 32 deep temporal buffer
  - $13 \times 13 \rightarrow 1516$ neurons with 64 deep temporal buffer
- 2×2 case illustrated



Particle Image Velocimetry

*W. Severa, O. Parekh, K. D. Carlson, C. D. James, and J. B. Aimone, "**Spiking network algorithms for scientific computing**," in Rebooting Computing (ICRC), IEEE International Conference on. IEEE, 2016, pp. 1–8.*



And more…

PIV Base Video 2

Circular Flow (Clockwise)

PIV Results Video 2

Circular Flow (Clockwise)

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Performance Results

Case: Seventy five (75) 640x480 image frames with 32x32 input tiles

- 300 tiles per image, 74 image compares, 22,200 algorithm executions, 1 execution requires 4994 ticks.

Mode	Chips	Inst.	Theoretical*	Actual*/ Overhead◊	Overclocked / Overhead◊
Serial	1	1	30.8 hrs.	31.8 hrs. / 88.7 hrs.	2.8 hrs.† / 59.5 hrs.
Parallel	1	5	6.2 hrs.	6.4 hrs. / 33.9 hrs.	0.6 hrs.† / 27.6 hrs.
Parallel	16	89	20.8 min.	21.0 min. / 4.7 hrs.	4.4 min.‡ / 4.6 hrs.
Parallel	16	110	16.8 min.	– / –	– / –

*1 tick = 1ms | †1 tick = 5µs | ‡1 tick = 200µs | ◊Includes I/O

Reported data is based on a small sample average and extrapolated.

Spike Optimization

- Implement fundamental mathematical operations through temporal coding
  - SpikingSort
  - SpikeMin, SpikeMax, SpikingMedian
  - SpikeOpt(median)
- Implemented on the STPU hardware

Algorithm	T_P	P	cost $= T_P \times P$
SpikingSort	$O(k)$	$O(N)$	$O(kN)$
SpikeMax	$O(k)$	$O(N)$	$O(kN)$
SpikingMedian	$O(k)$	$O(N)$	$O(kN)$
SpikeOpt (median), worst case	$O(N/2)$	$O(N)$	$O(N^2)$
SpikeOpt (median), $\|X\|$ is constant	$O(1)$	$O(N)$	$O(N)$

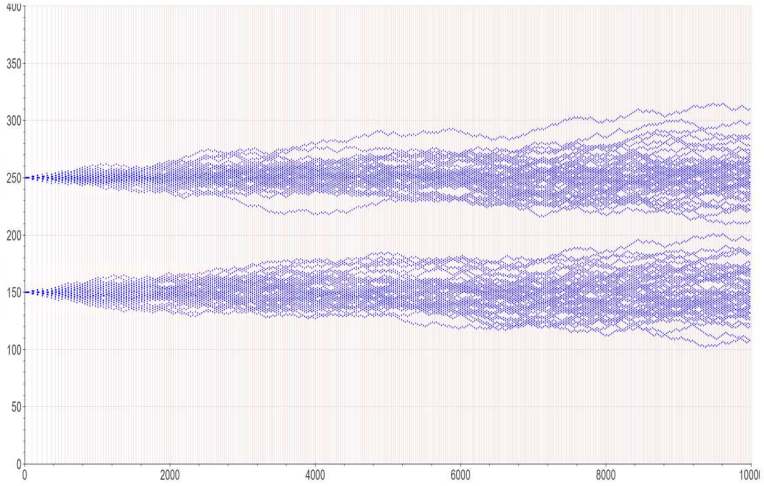
Spike Optimization

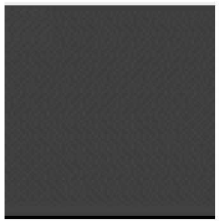
Verzi, Stephen J., et al. "**Optimization-based computation with spiking neurons**," *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 2017.

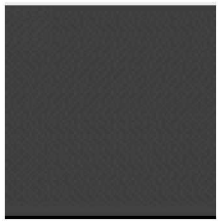
STPU Results of Temporal Coding

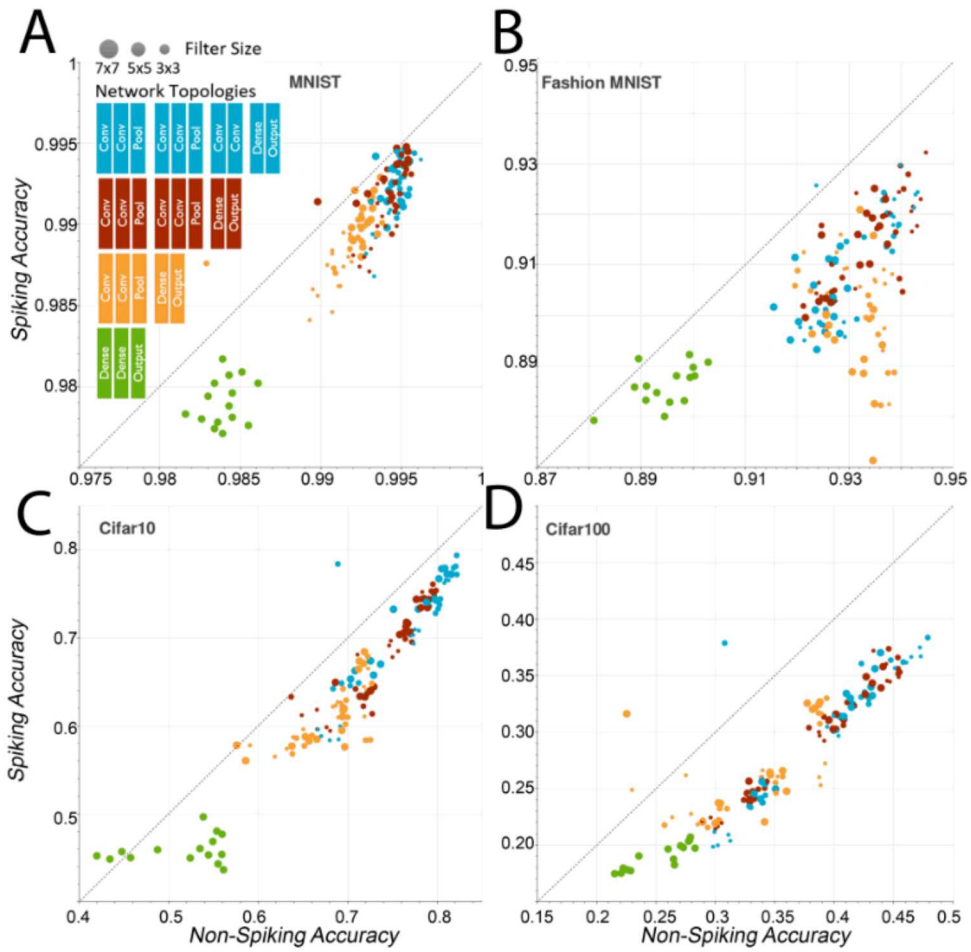
Aggregation of spikes weighted by their temporal code value

Example Algorithms on Neuromorphic





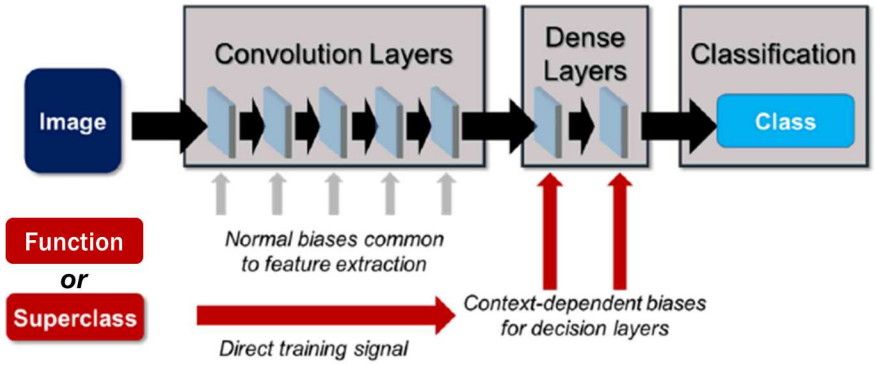


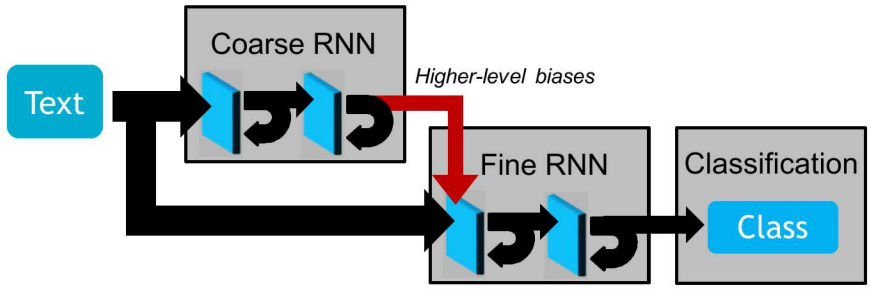


W. Severa, C. M. Vineyard, R. Dellana, S. J. Verzi, and J. B. Aimone, "Training deep neural networks for binary communication with the whetstone method," Nature: Machine Intelligence, In Press.

Context-Sensitive Deep Learning

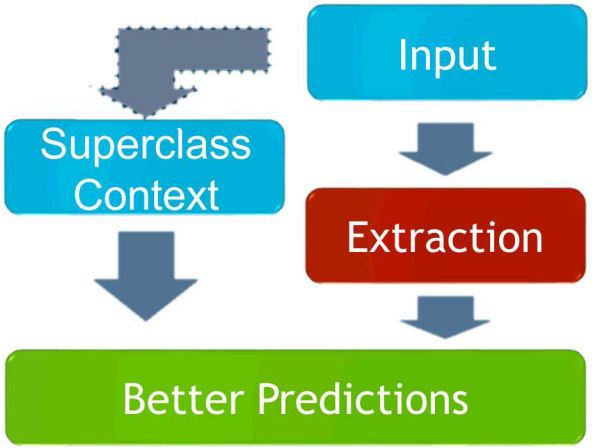
- Provide a network with the flexibility to perform different tasks without reprogramming
- Neuromodulation: The idea that diffuse, network-wide inputs can adjust behavior
  - Contextual information is fed into network through a parallel pathway
  - Context neuromodulation provides a biasing effect on downstream neurons
- Current capabilities:
  - **Superclass exclusion**: lower-level characteristics that are dependent on higher-level abstractions
  - **Context-dependent function**: ability of a singular network to incorporate multiple behaviors



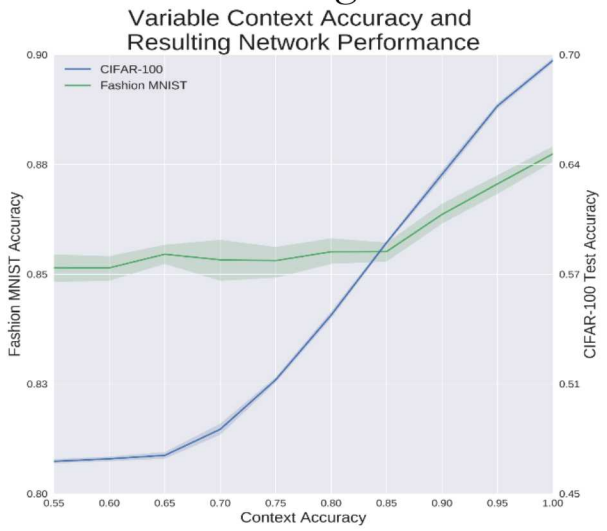


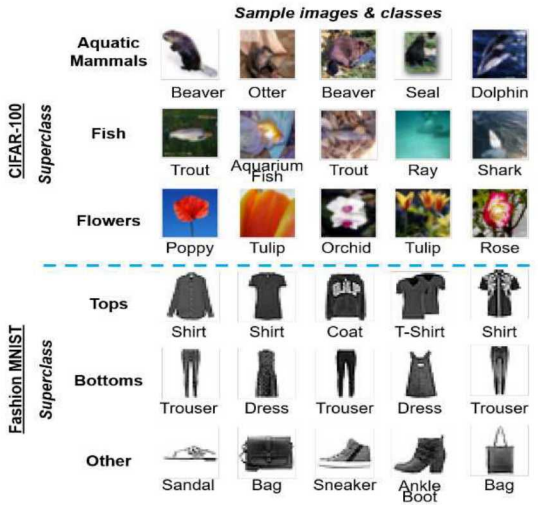
Superclass Exclusion

Lower-level characteristics that are dependent on higher-level abstractions



Images: CIFAR100 and FashionMNIST

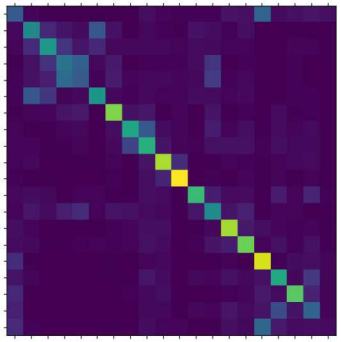




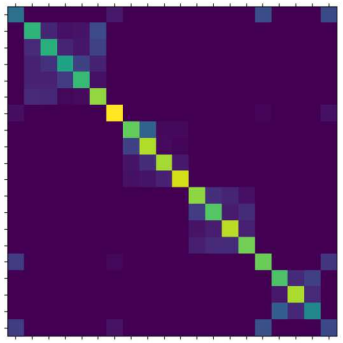
Text: 20 Newsgroups

Metric\Context	Control	Direct	Learned
Accuracy (Top 1)	.505	.689	.549
Accuracy (Top 3)	.757	.935	.779
F1-score	.482	.668	.536

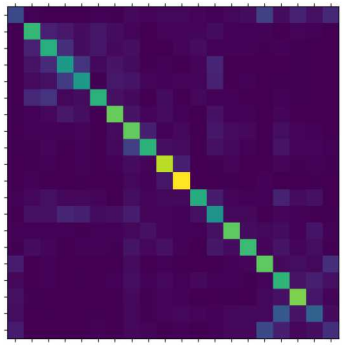
Control (No Context)



Direct (Provided Context)

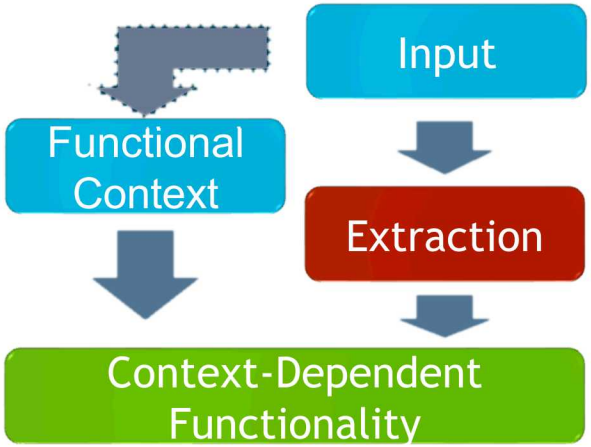


Learned (Inferred Context)



Context-dependent function

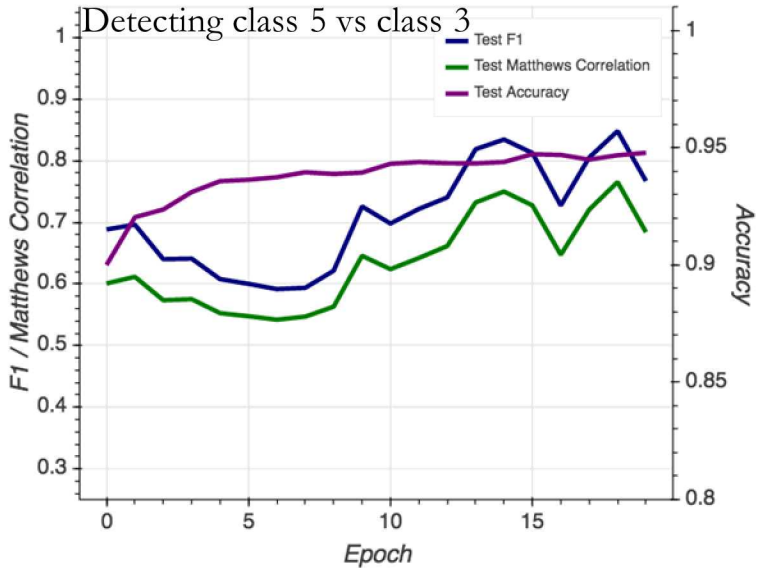
Ability of a singular network to incorporate multiple behaviors



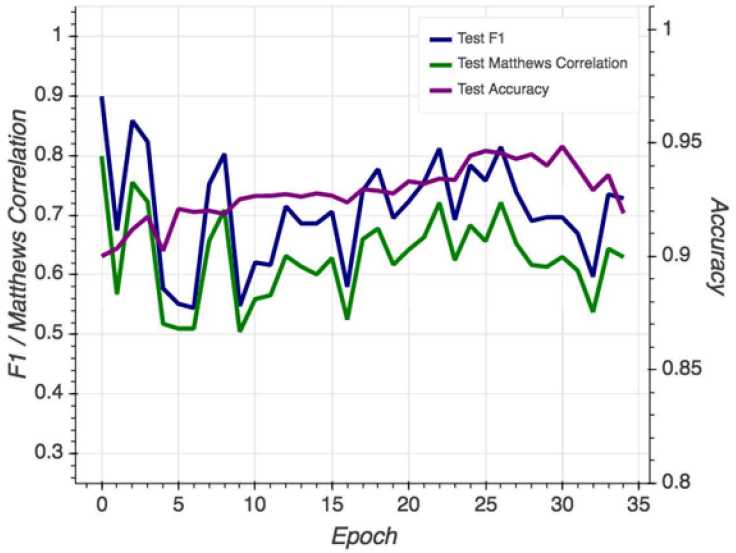
Detecting four separate classes dependent on context

Context	Accuracy
3 (Cat)	.9448
5 (Dog)	.9603
7 (Horse)	.9752
9 (Truck)	.9108

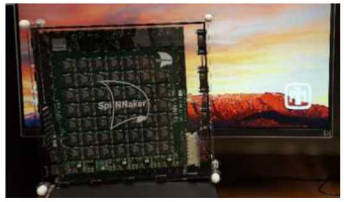
class 5

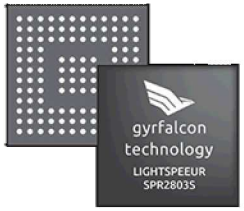


class 3

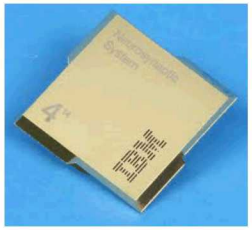


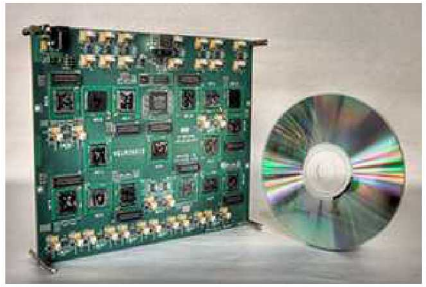
Neuromorphic Hardware

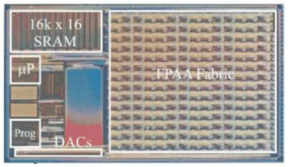


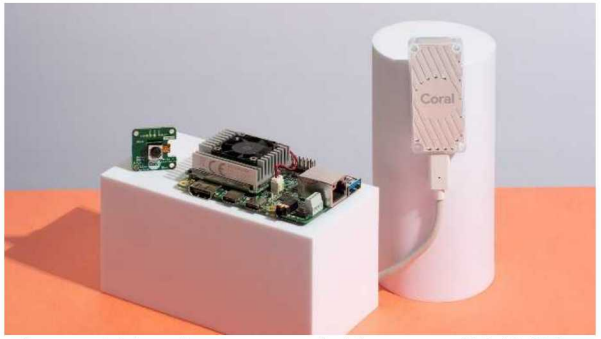


https://www.gyrfalcontech.ai/solutions/2803s/

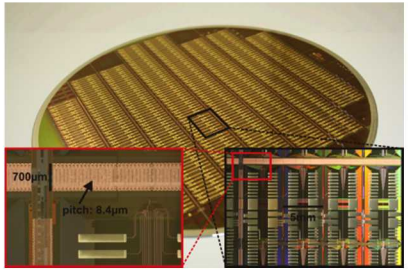




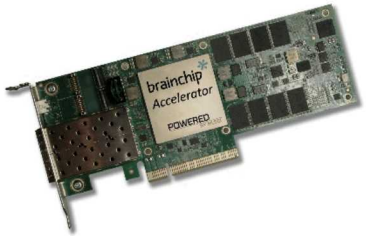




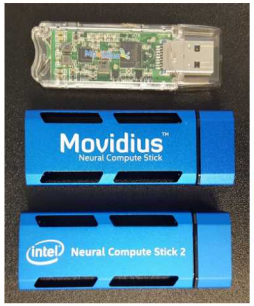
https://developers.googleblog.com/2019/03/introducing-coral-our-platform-for.html



http://www.artificialbrains.com/brainscales

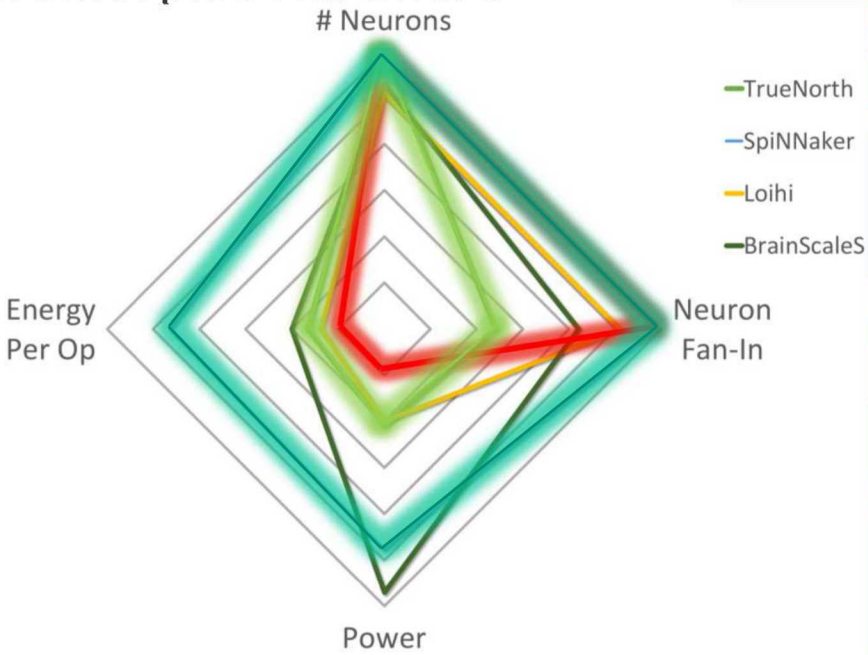


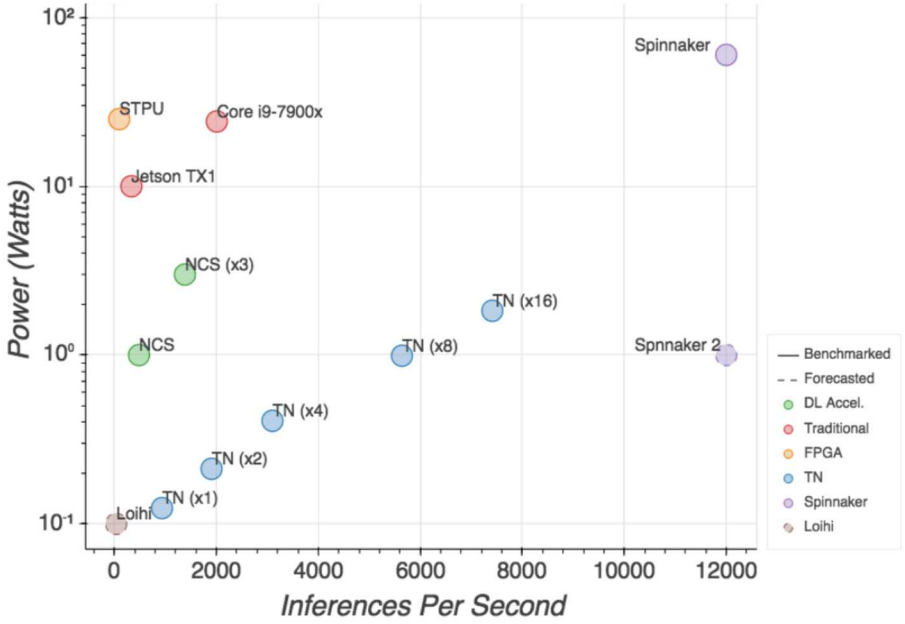
https://www.brainchipinc.com/products/brainchip-accelerator

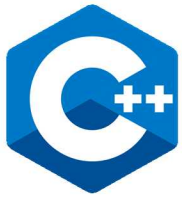


Programming and Performance of Neuromorphic Hardware

- There are many different emerging neuromorphic architectures
  - Design tradeoffs focus upon different features making them better suited for different applications
  - Architectural differences result in performance differences for different tasks
- Bottom figure shows benchmark results across a suite of architectures on an inferencing task comparing throughput with power consumption
- Seeing great promise in terms of performance per watt from emerging neuromorphic architectures
- Such approaches are an enabler for performing AI tasks in SWaP constrained environments

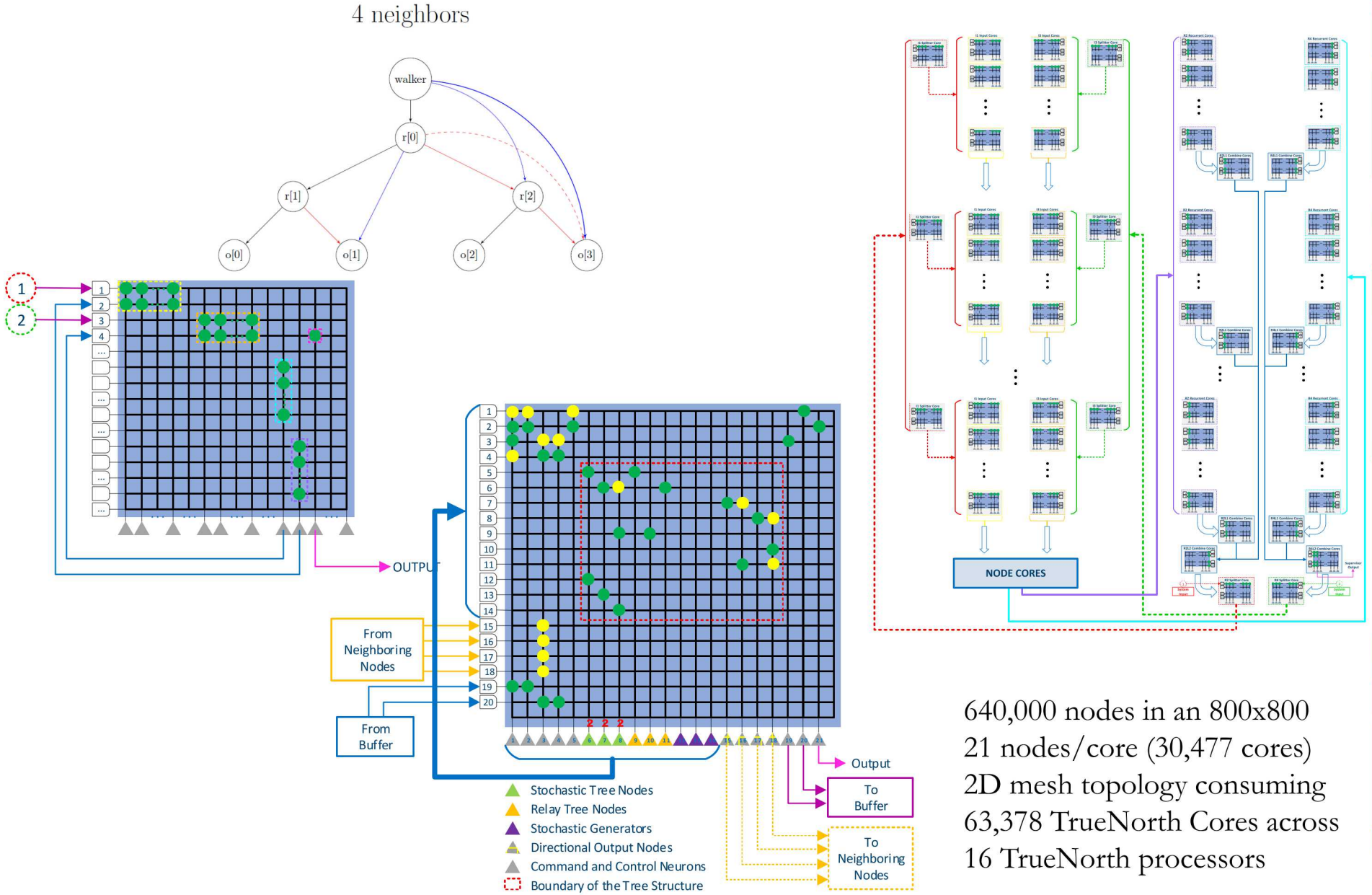




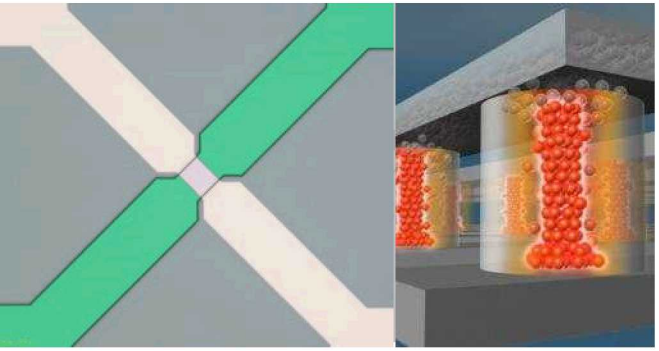


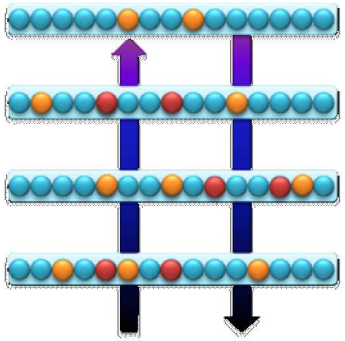
MATLAB

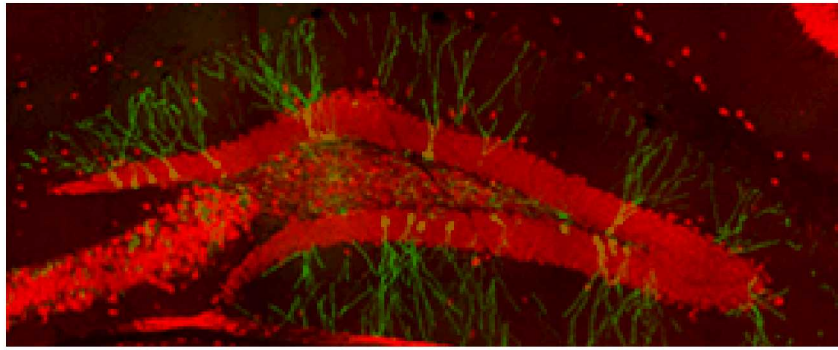
Programming and Performance of Neuromorphic Hardware



640,000 nodes in an 800x800
21 nodes/core (30,477 cores)
2D mesh topology consuming
63,378 TrueNorth Cores across
16 TrueNorth processors







Thank you for your time

Questions?

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.