SAND2018-1862C

Neural-Inspired Technologies for Data Processing and Scientific Computing

Presented by: Craig M. Vineyard

Sapan Agarwal, Brad Aimone, Frances Chance, Ryan Dellana, Tim Draelos, Jonathon Donaldson, Meg Galiardi, Sam Green Aaron Hill, Conrad James, Matt Marinella, John Naegle, Ojas Parekh, Cindy Phillips, Tu-Thach Quach, Fred Rothganger, William Severa, Mike Smith, Steve Verzi, Felix Wang, Christy Warrender, Kevin Dixon, Reno Sanchez, John Wagner, John Feddema

Sandia National Laboratories

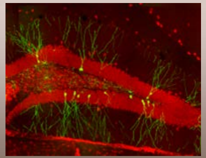
Exceptional service in the national interest

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.

Neural Computing at Sandia Labs Leverages a Large Research Foundation

Sandia National Laboratories

Neuroscience Theory



IARPA MICrONS Government Team for Test & Evaluation of Neural Models and Machine Learning

- Neuro-informatics
- Modeling and Simulation
- Neural Data Analytics
- Computational Neuroscience

Neural Computing Capabilities

HAANA Grand Challenge – Flagship LDRD across computing, materials, and cyber security centers

- Formal Neural Computing Theory
- Neural Inspired Architectures
- Neural Machine Learning Algorithms
- UQ / SA of Neural Algorithms and Neural Architectures

Neural-enabling Hardware

MESA Fabrication Facility provides materials and design research capabilities for next generation neural systems

- Micro-sensors
- Non-von Neumann architectures
- Memory technology
- Neuromorphic Computing Lab

Mission Impacts

Enabling Advanced Simulation and Computing

- Neural-inspired communication paradigms
- Adaptive memory management
- Numerical computing with neurons

ASC

Neuroscience Contributions

- Contribute to the science of understanding the brain
- BRAIN Initiative
- MICrONS

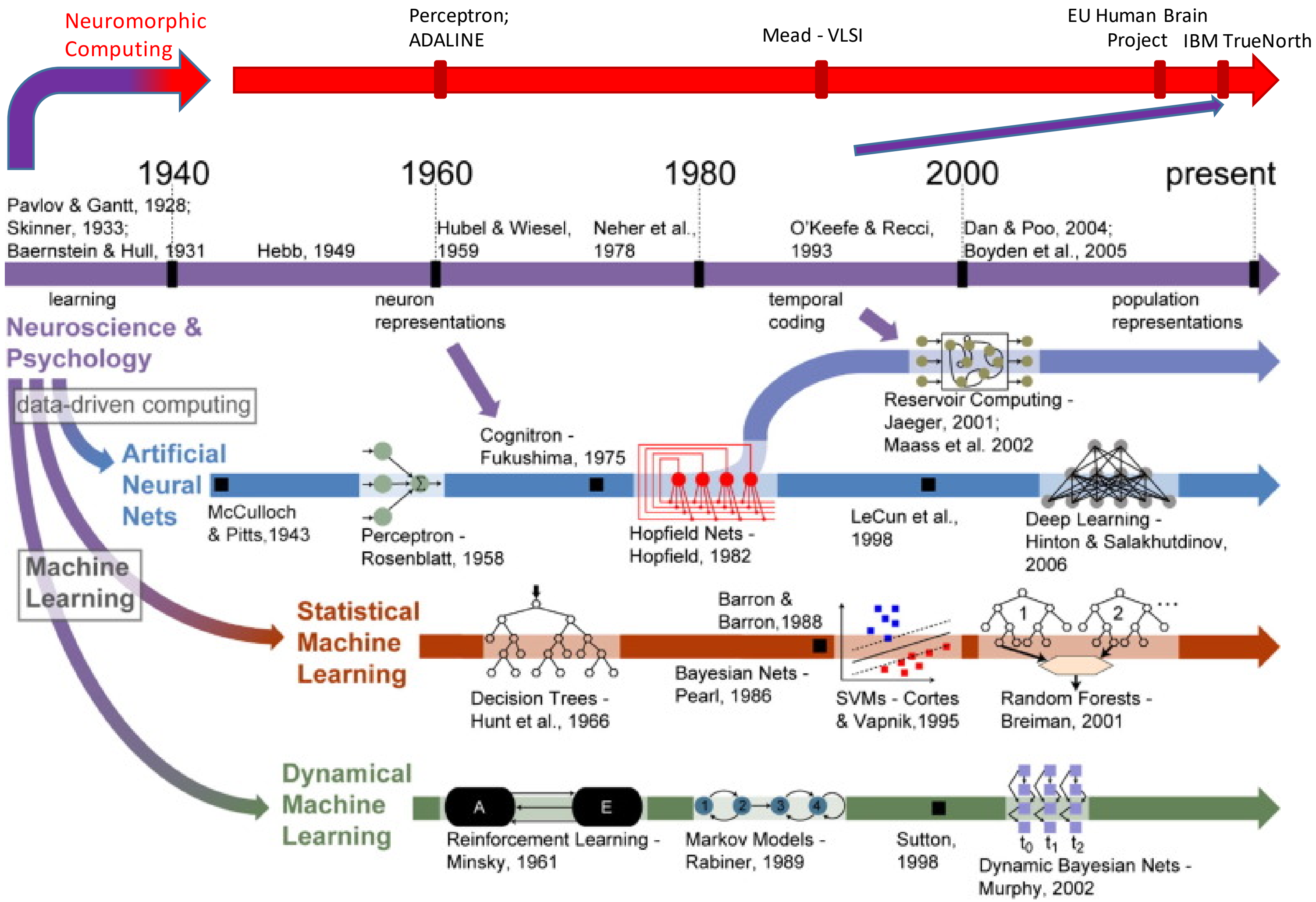
Deployable National Security Applications

- Cyber Defenses
- Embedded Pattern Recognition Systems
- Smart Sensor Technologies

U.S. DEPARTMENT OF ENERGY

Neuromorphic Computing

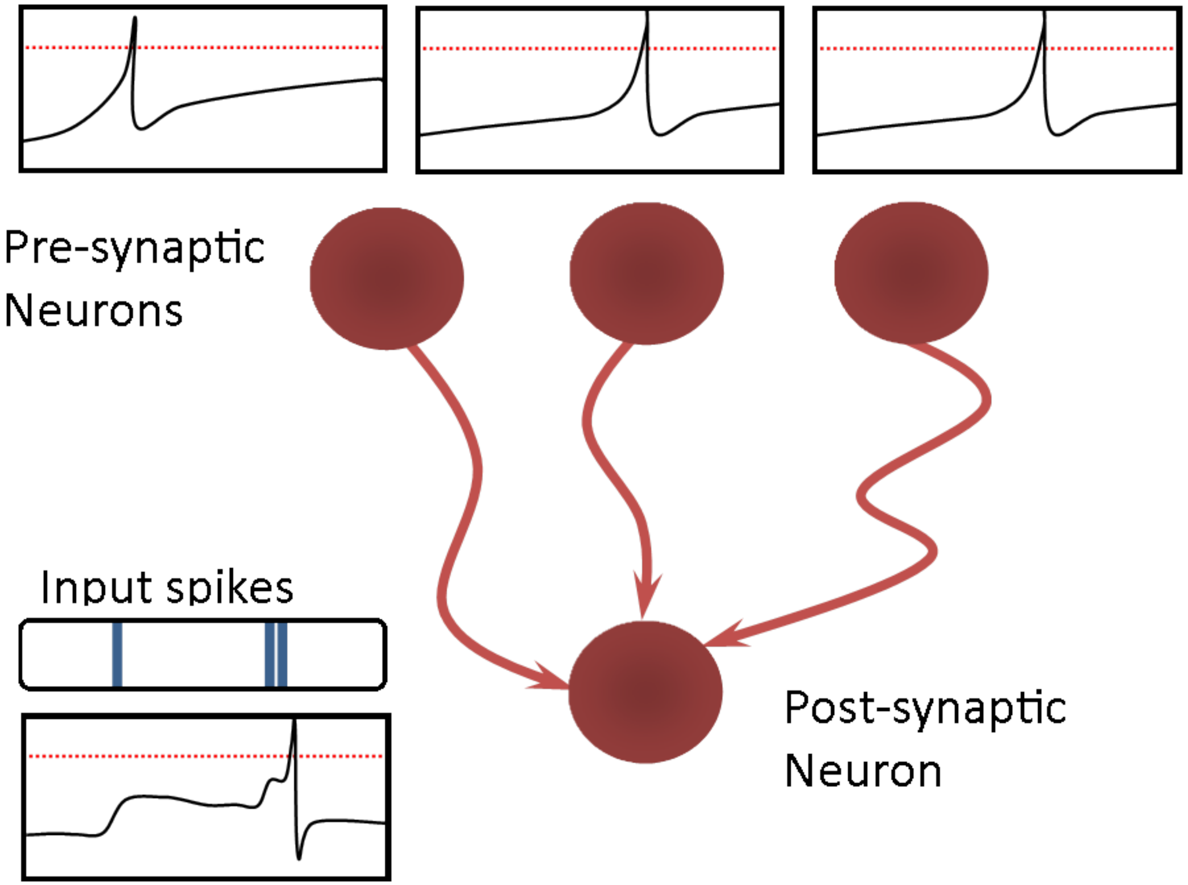
- What is neural-inspired, neuromorphic, brain-inspired computing?
  - Many terms
  - Fundamental notion of taking inspiration from how the brain performs computation
- With the advent of mathematical reductionist models going back to 1943 there have been many parallel efforts to likewise implement them in hardware
- HOWEVER, many of these efforts are simply accelerators of classic architectures
- Do NOT incorporate many neural principles since 1940s
- Rather took advantage of Moore’s Law & Dennard scaling to allow neural networks to deliver upon original promise



James, et al., BICA 2017

Spiking Neurons

- Neurons are connected via synapses and communication is sent in single-state signals called spikes
- Spikes require time to propagate
- Time Dimension/Spikes are the main differentiator between Spiking Neural Networks and more basic Artificial Neural Networks
- Incoming spikes adjust an internal potential by some weight; if potential reaches a threshold, the neuron sends out spikes
- If potential is sub-threshold, it decays according to a leakage constant
- Leaky Integrate and Fire neurons roughly approximate biological neurons



Neuromorphic Processors

Analog

- Focus on Kirchhoff Law – enabled computation
  - Neurons sum current across weighted synapses
  - Neural nodes sum current over weighted memristors
- Substantial energy and time savings
  - Non-trivial costs of precision
  - Practical issues limit size and integration with digital logic
- Ideal scenario
  - Train weights in situ
  - Compatible with linear algorithms

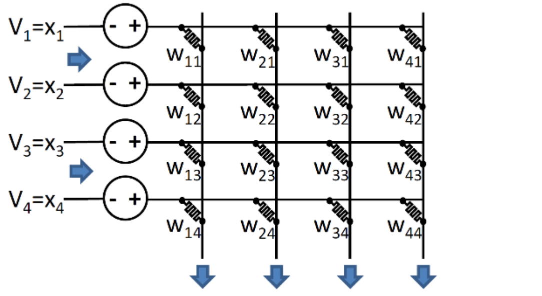
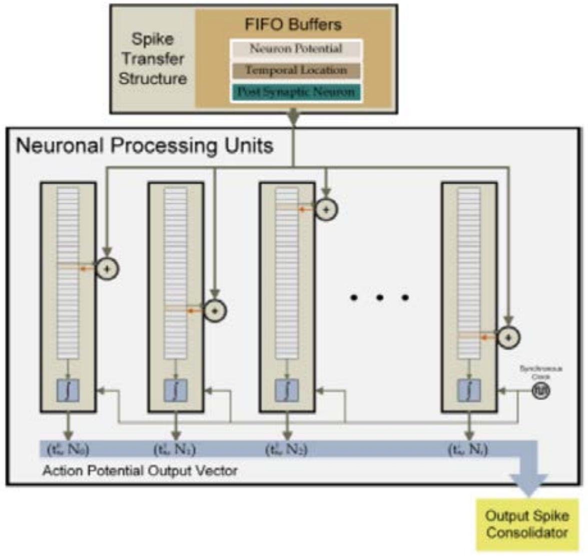


Fig 1: Analog RRAMs can be used to reduce the energy of a vector matrix multiply. The conductance of each RRAM represents a weight. Analog input values are represented by the input voltages or input pulse lengths. This allows all the read operations, multiplication operations and sum operations to occur in a single step. A conventional architecture must perform these operations sequentially for each weight resulting in a higher energy and delay.

Agarwal et al., E3S 2015

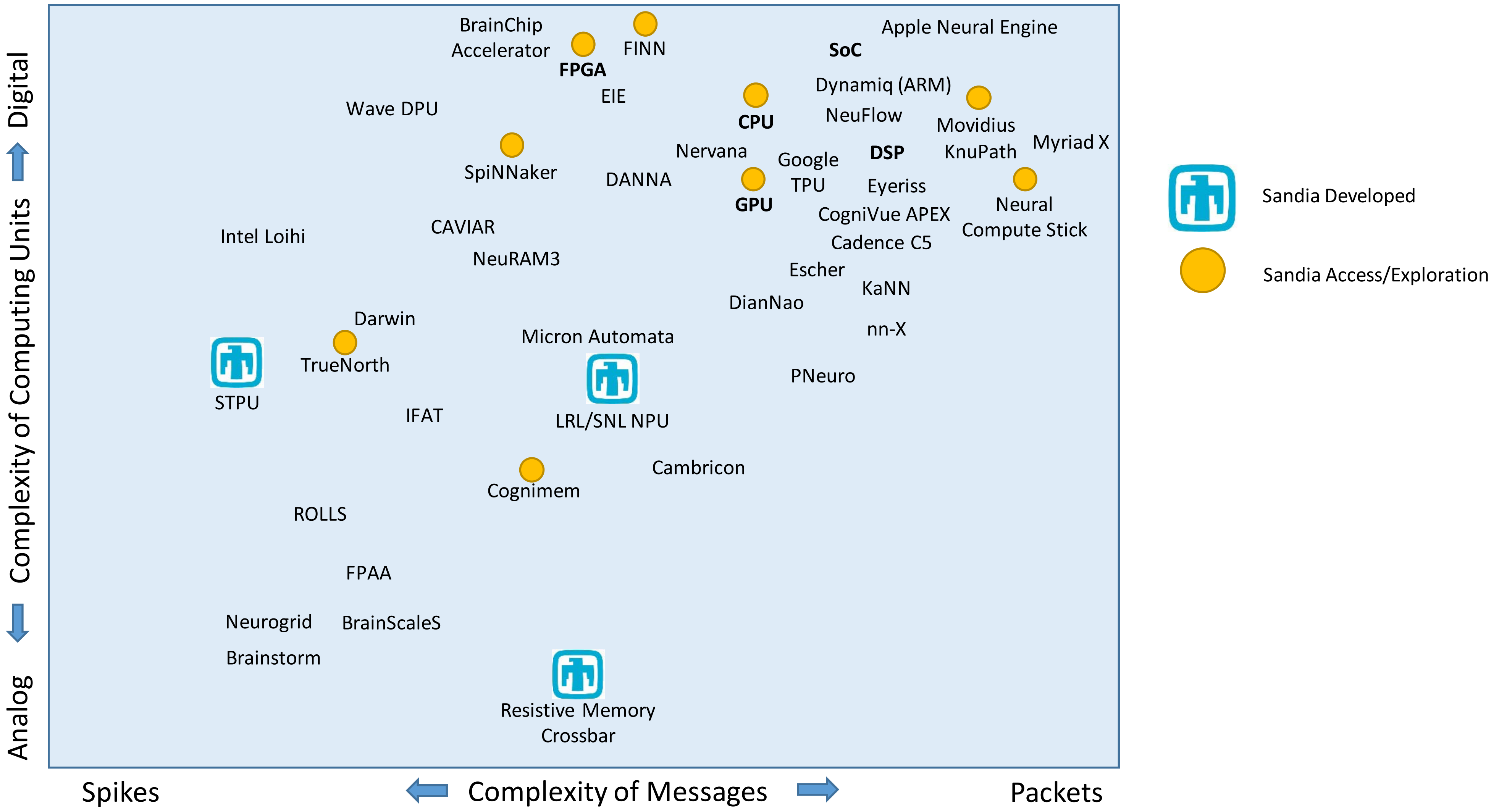
Digital

- Rely on event-driven "spiking" for communication
  - Communication only needed for '1's', not otherwise
  - Equivalent to large threshold gate networks + time dimension
- Substantial energy savings
  - Information in time dimension; limiting time savings
- Compatible and scalable using conventional technology
- Ideal scenario
  - Algorithms can be reframed in discrete spiking form
  - Learning algorithms are reformulated for spiking approaches

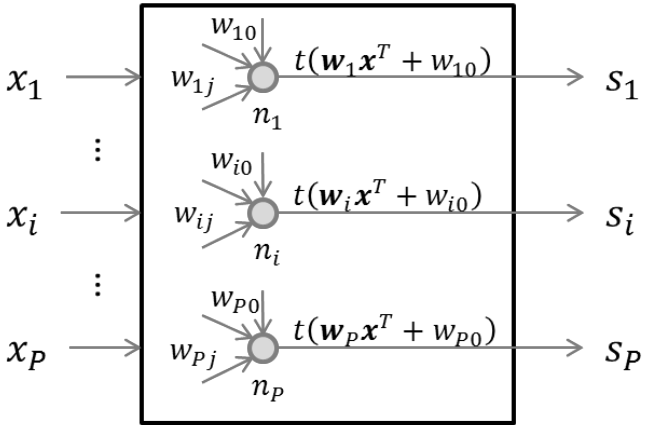


Architectural Landscape

Landscape of emerging neuromorphic architectures (non-exhaustive)

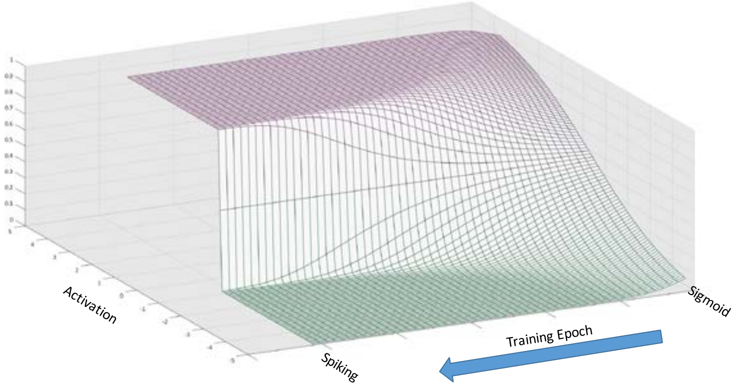


Neuromorphic Computing Algorithms



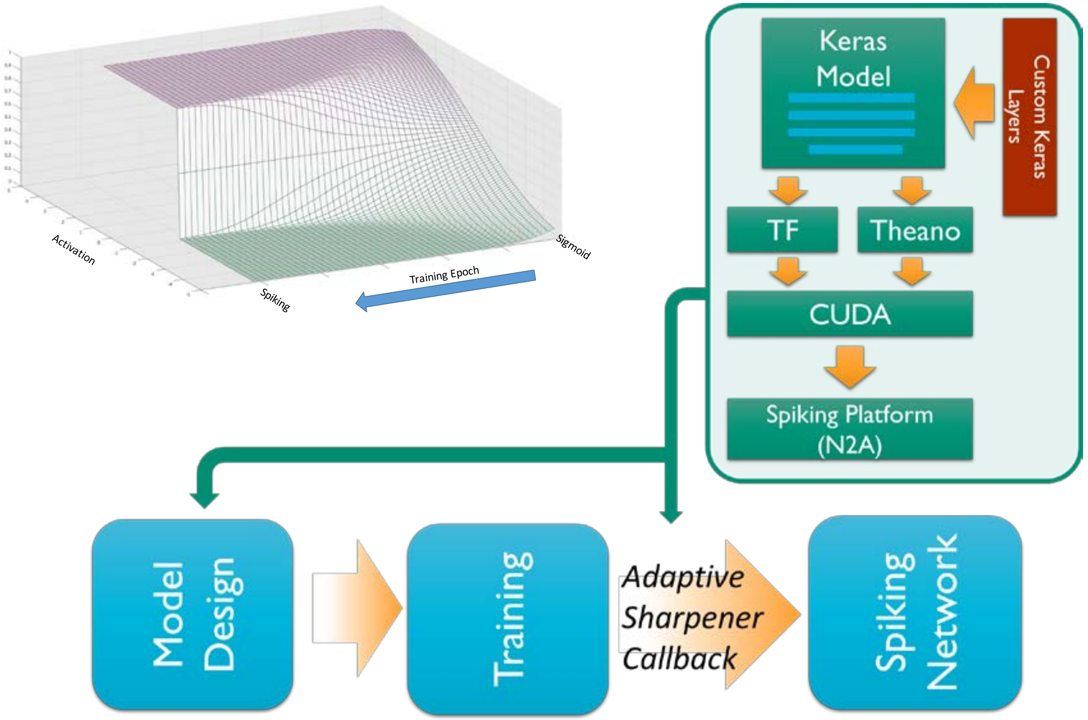
Verzi et al., IJCNN 2017

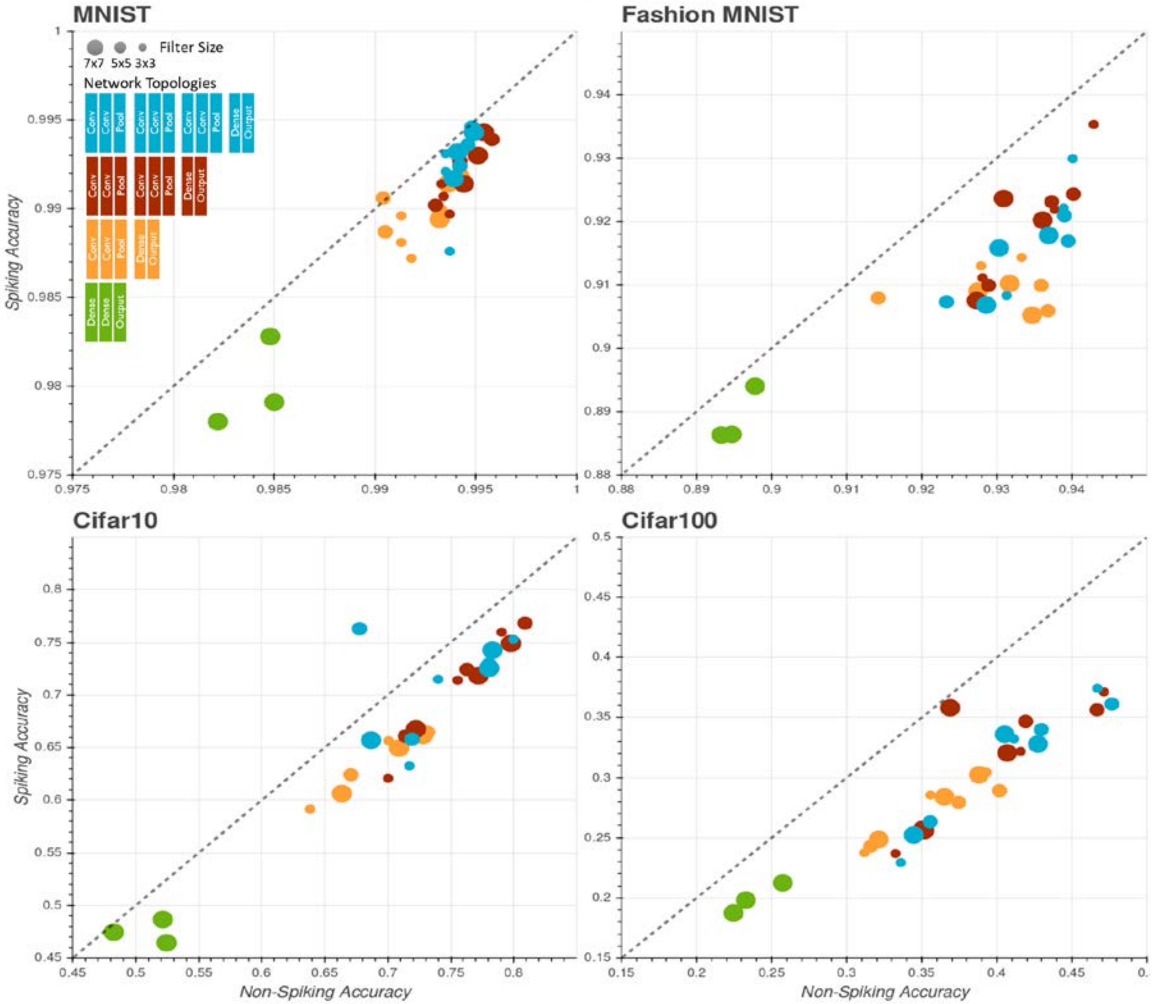
- May require different algorithmic approaches
- May require different encodings
  - Example: Rate coding vs. temporal coding
    Non-spiking vs. spiking
  - Fundamentally changing how computation and representation are done
- Compile/Link standard does not yet exist
- Requires new metrics for benchmarking



Whetstone

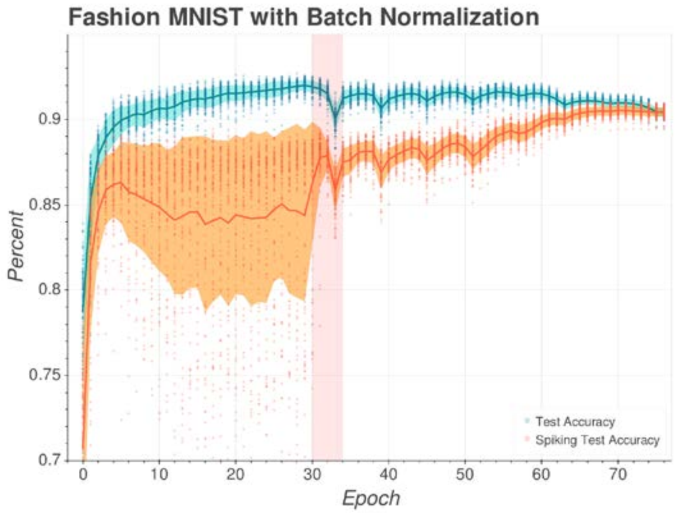
- An accessible, platform-independent method for training spiking DNNs for neuromorphic processors

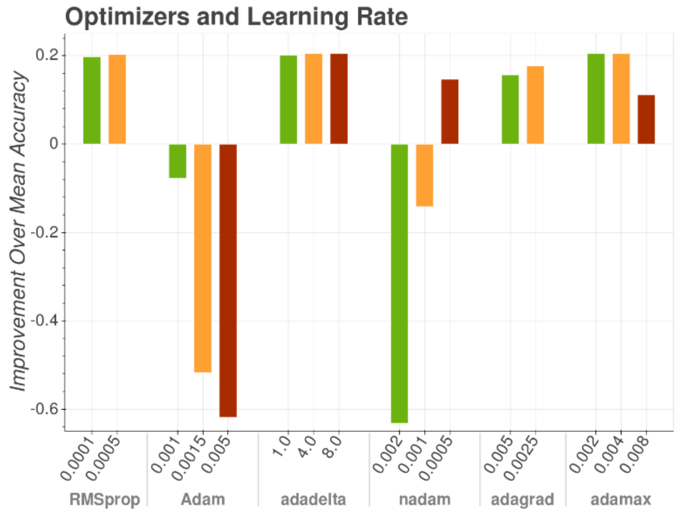




Whetstone

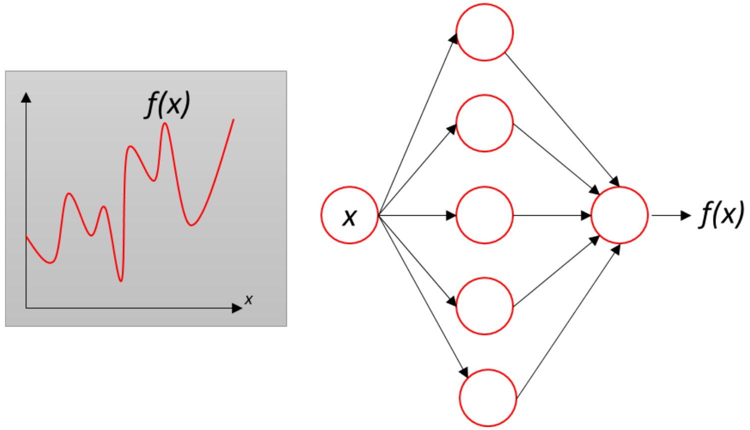
- Modifications for the network topology are limited to the activation function and output layer
- Many standard, effective techniques translate immediately to the spiking neural network: Dropout, Max Pooling, Batch Normalization
- Batch normalization greatly improves convergence to spiking activations
  - Majority of accuracy degradation occurs during the sharpening of the first layer
  - Batch normalization helps mitigate this loss
  - Useful for even smaller networks
- Activation sharpening is optimizer agnostic → However, certain optimizers are better suited. Moving average modulation improves repeatability.

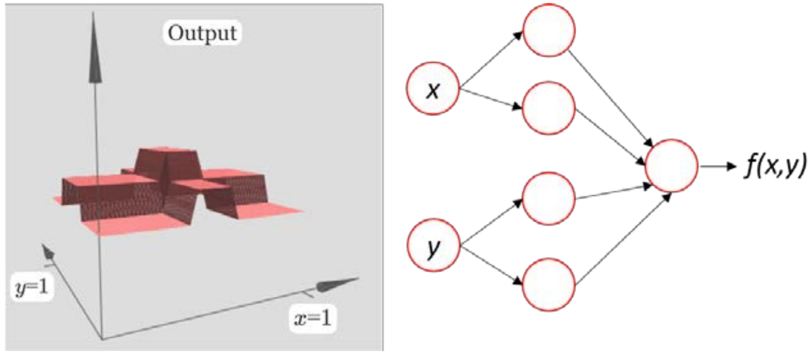




Spiking Neural Algorithms

Universal Function Approximation

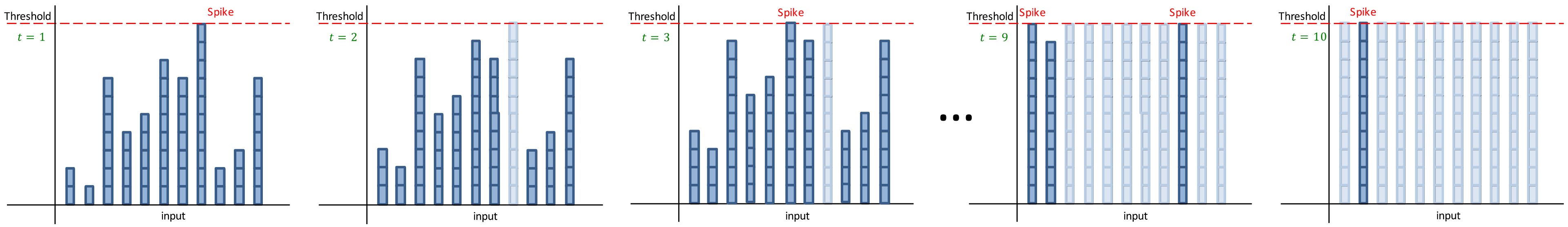




Spiking Neural Circuits

- Optimization
  - Max/Min
  - Sort
  - Median Filter
- Machine Learning
  - spiking-Nearest Neighbor
  - Spiking-ART
- PDE
  - Monte Carlo Random Walker for Diffusion
- Cross-Correlation

Spiking Optimization Algorithms



time-step	N: 1	2	3	4	5	6	7	8	9	10	11
1								1			
2											
3						1					
4			1				1				1
5											
6					1						
7				1							
8										1	
9	1								1		
10		1									

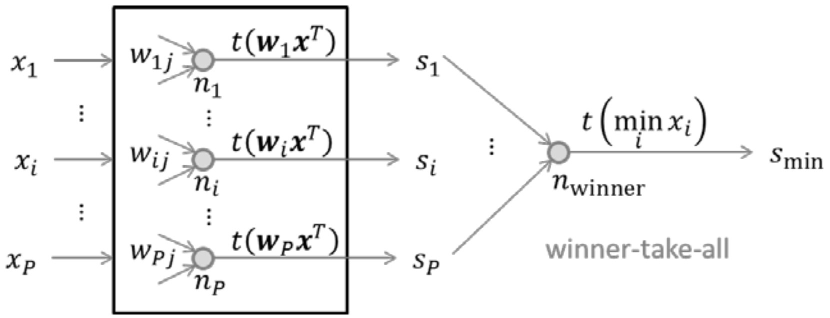
maximum

median

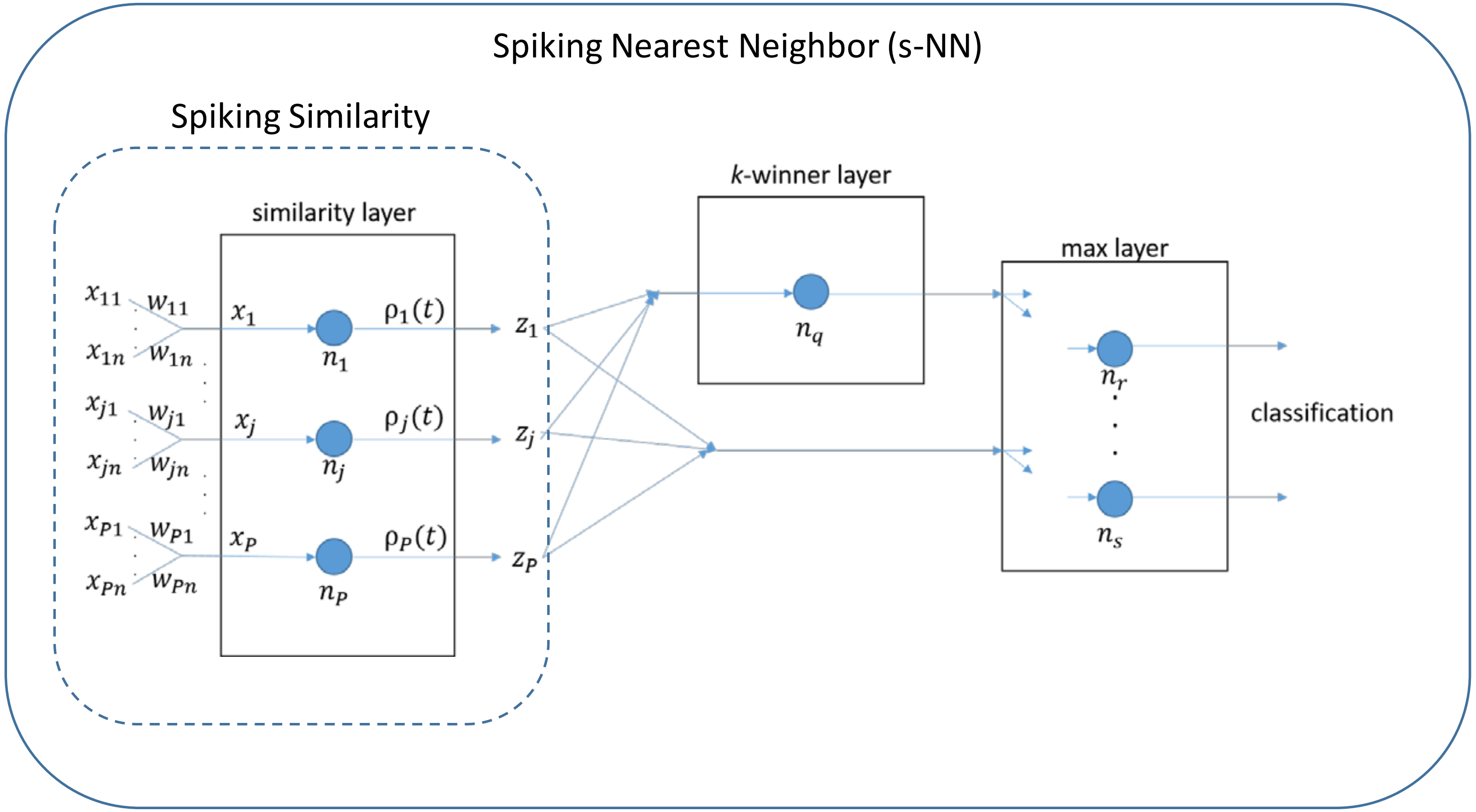
minimum

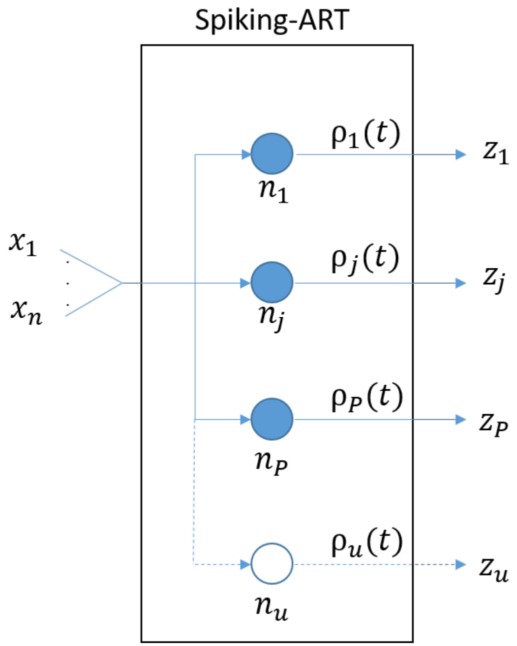
Sorted integers: {9, 7, 6, 6, 6, 5, 3, 2, 1, 1, 0}

- Finding the min where $P \geq N$

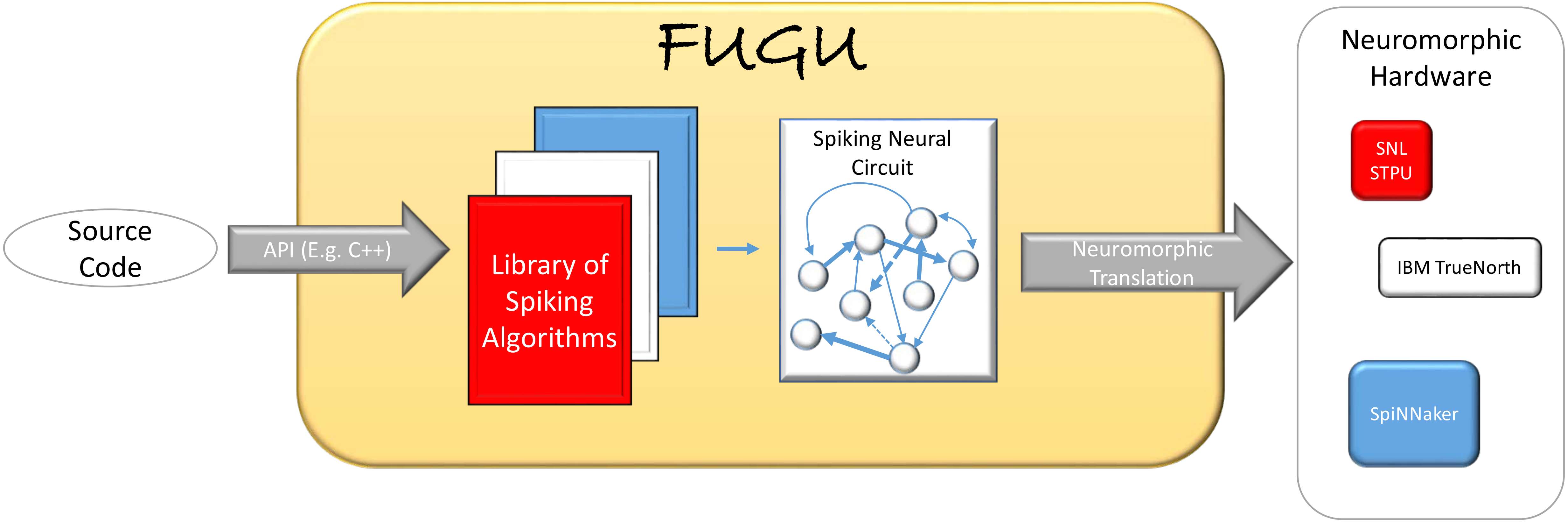


Spiking Machine Learning Algorithms





FUGU

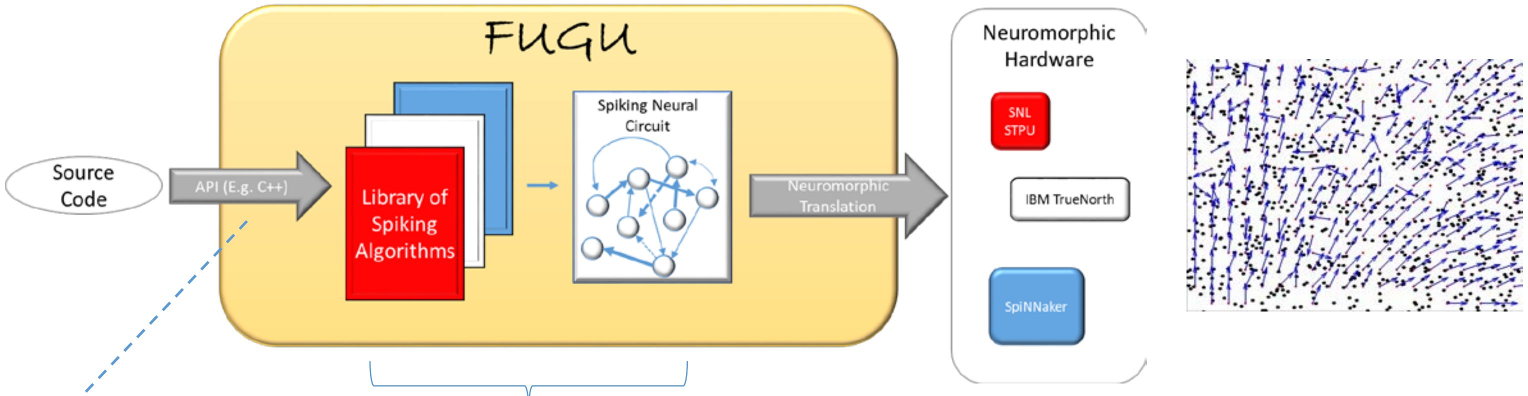

Source Code
API (E.g. C++)
FUGU
Library of Spiking Algorithms
Spiking Neural Circuit
Neuromorphic Translation
Neuromorphic Hardware
SNL STPU
IBM TrueNorth
SpiNNaker

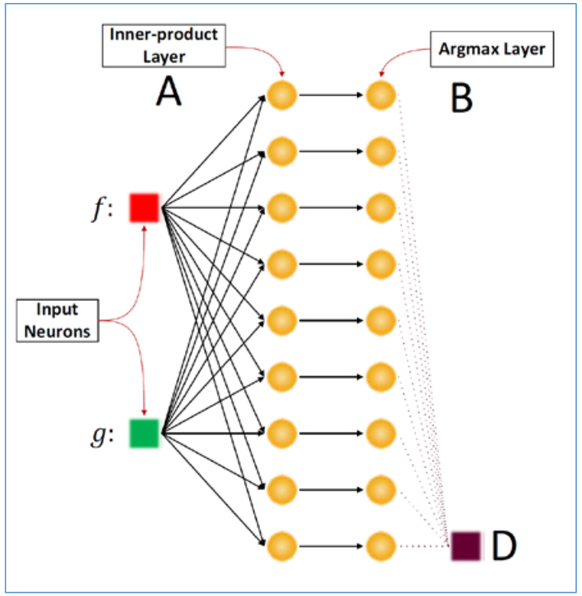
FUGU: PIV Cross-Correlation Example

- Particle Image Velocimetry (PIV) is a well studied method for using particles to determine the local velocity flow in many applications throughout science and engineering

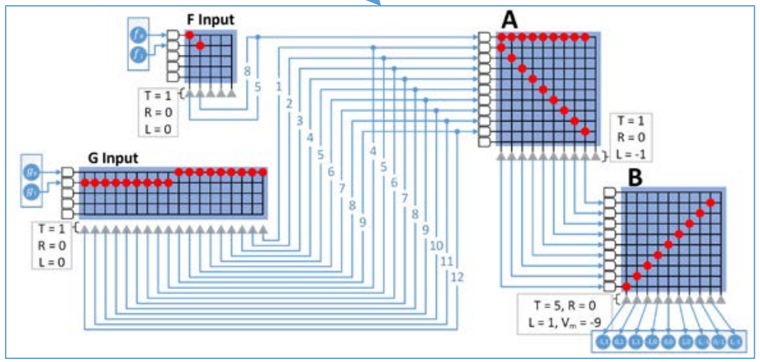
- Cross-Correlation finds agreement in signals
  - Computed as a sliding scalar product
  - $(f \star g)(n) = \Sigma_m f(n) g(m+n)$

- Mapped to the SNL STPU & IBM TrueNorth Neuromorphic architectures





Severa et al.,ICRC2016

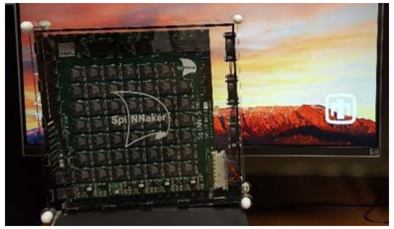


Neural Exploration & Research Lab (NERL)

Sandia National Laboratories

- Enables researchers to explore the boundaries of neural computation
- Consists of a variety of neuromorphic hardware & neural algorithms providing a testbed facility for comparative benchmarking and new architecture exploration

SpiNNaker 48 Node Board



IBM TrueNorth*

IBM TrueNorth NS16e*

Intel Neural Compute Stick

Cognimem CM1K

KnuPath Hermosa

SNL STPU on FPGA

Xilinx ZYNQ-7000 FPGA

Xilinx PYNQ FPGA

Inilabs DAVIS 240C DVS

Nvidia Jetson TX1

GPU Workstations

* Remote Access

Conclusion

- There are some bold & exciting claims surrounding neuromorphic computing

- The interplay of algorithms, architectures, and hardware is incredibly important
  - In our approach, we’ve been focusing upon the significance neuroscience & fundamental theory

- Sandia Labs is working to understand this landscape & employ neural-inspired computing for scientific computing and other domains

Sandia National Laboratories

Neuromorphic Hardware in Practice and Use

Description of the workshop

- Abstract – This workshop is designed to explore the current advances, challenges and best practices for working with and implementing algorithms on neuromorphic hardware. Despite growing availability of prominent biologically inspired architectures and corresponding interest, practical guidelines and results are scattered and disparate. This leads to wasted repeated effort and poor exposure of state-of-the-art results. We collect cutting edge results from a variety of application spaces providing both an up-to-date, in-depth discussion for domain experts as well as an accessible starting point for newcomers.

Goals & Objectives

- This workshop strives to bring together algorithm and architecture researchers and help facilitate how challenges each face can be overcome for mutual benefit. In particular, by focusing on neuromorphic hardware practice and use, an emphasis on understanding the strengths and weaknesses of these emerging approaches can help to identify and convey the significance of research developments. This overarching goal is intended to be addressed by the following workshop objectives:
  - Explore implemented or otherwise real-world usage of neuromorphic hardware platforms
  - Help develop ‘best practices’ for developing neuromorphic-ready algorithms and software
  - Bridge the gap between hardware design and theoretical algorithms
  - Begin to establish formal benchmarks to understand the significance and impact of neuromorphic architectures

http://neuroscience.sandia.gov/research/wcci2018.html

Call: https://easychair.org/cfp/nipu2018