SAND2022-3801C

Sandia National Laboratories

Exceptional service in the national interest

Spiking Neural Approaches to SAR ATR

Presented by: Craig M. Vineyard, PhD

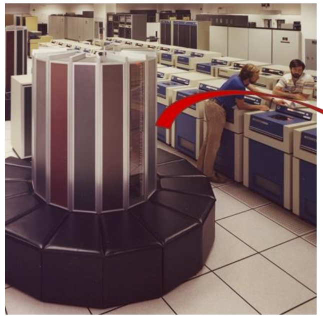
Team: James B. Aimone, Ryan Dellana, EJ Guillen, Aaron J. Hill, William M. Severa, Craig M. Vineyard, & Javier Zazueta

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. ...ation under contract DE-NA0003525.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & 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.

Introduction

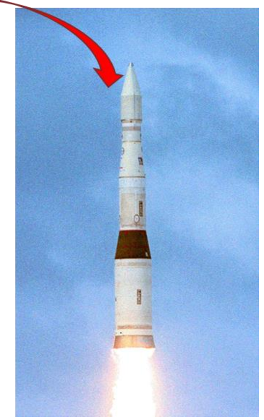
CRAY-1



https://shop.minimuseum.com/products/first-super-computer

Sandia National Labs Airborne Computer (SANDAC)

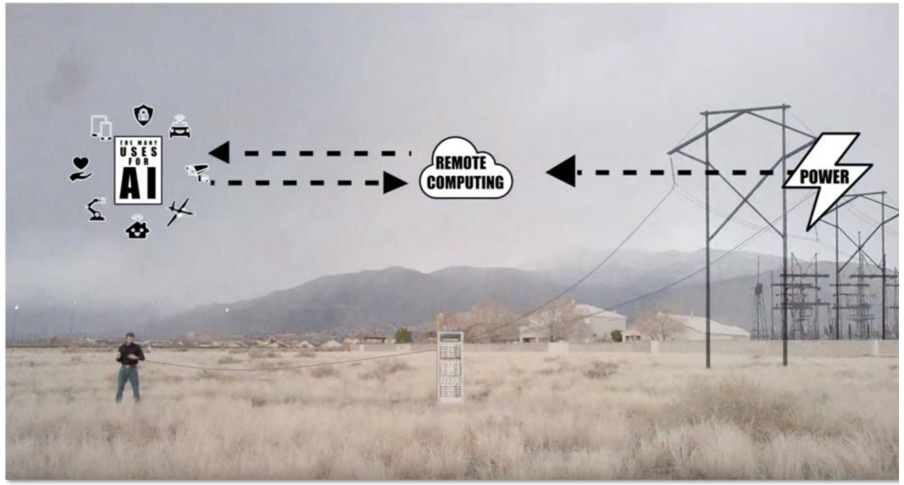
https://www.tomshardware.com/picturestory/866-supercomputer-department-of-energy-amd-intel-nvidia.html#s5



The pursuit of advanced computing technologies has seen two paths of great advances which intrinsically are entwined

- Architectures
  - Microelectronics advances have enabled immense computational power

- Algorithms
  - Increasingly sophisticated computations continue to challenge available computational platforms



Neural-Inspired Computing

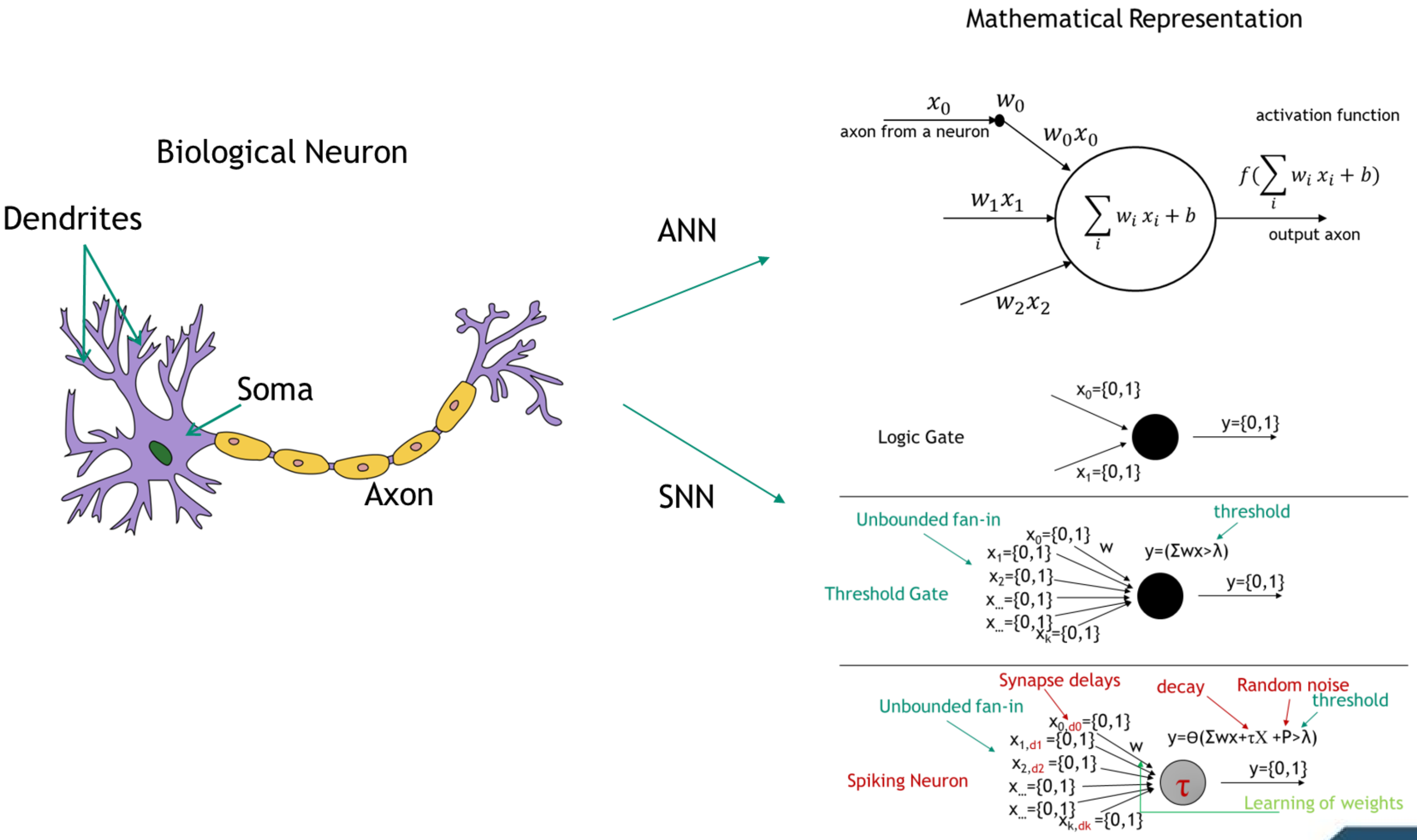
Neural-Inspired Computing

What is neural-inspired, neuromorphic, brain-inspired computing?

- Many terms
- Fundamental notion of taking inspiration from how the brain performs computation

General principles -

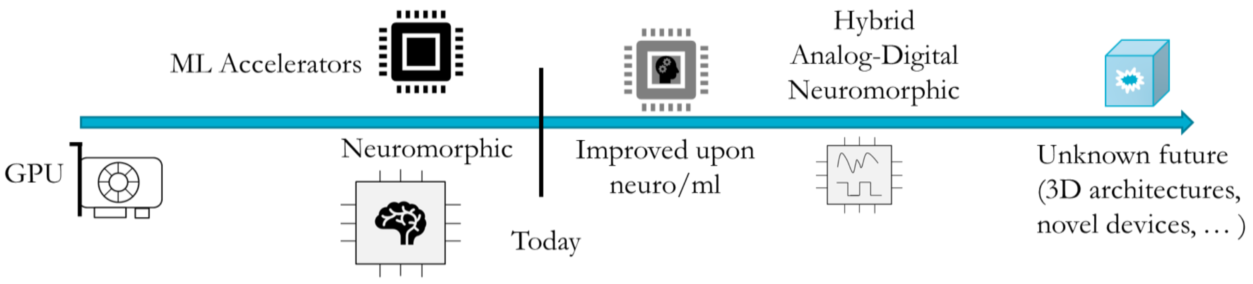
- Computing using a substrate of neuron-like components
- Each neuron connected to many other neurons
- Each connection with a 'synaptic weight' that governs how much one neuron affects another
- Highly parallel
- An algorithm is implemented by constructing a neural circuit
- ANN (Artificial Neural Networks) & SNN (Spiking Neural Networks)



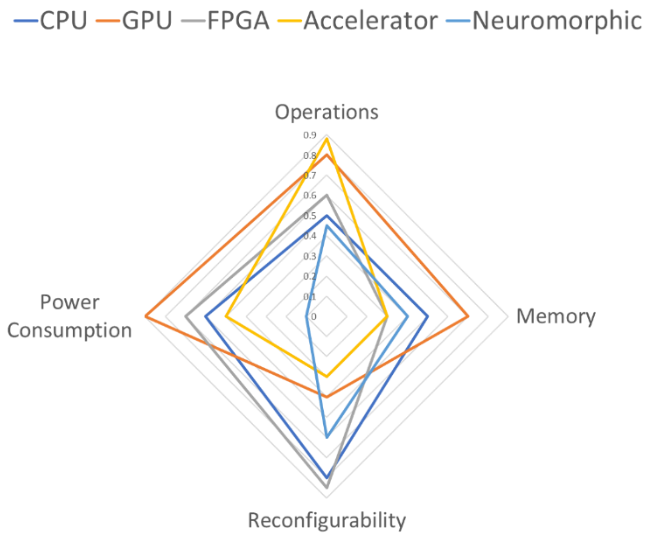
Which Neural Approach is Best?

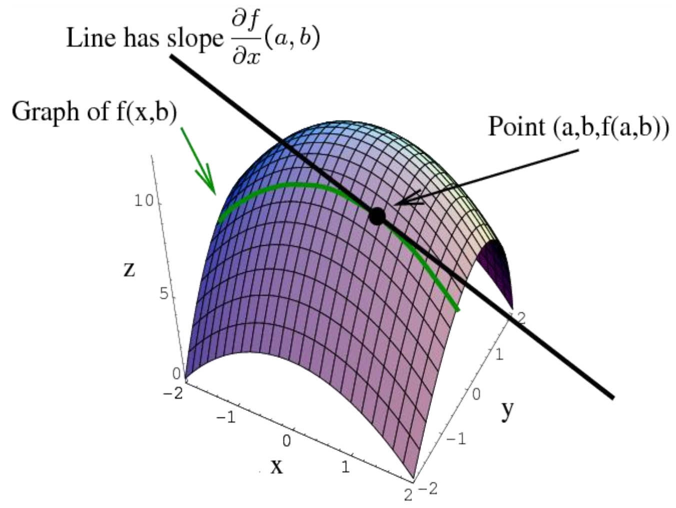
Architectural explosion

- Worldwide
- Industry as well as academia
- Leading chip vendors as well as startups
- Approaches include
  - Optimizing existing architectures for neural networks
  - Novel materials
  - Analog, digital, optical, asynchronous, event driven
- Scale & Technical maturity
- Emerging software stack



Timeline of neuromorphic impact is both now and emerging





https://mathinsight.org/image/partial_derivative_as_slope

For a fixed architecture (hardware), the algorithms (software) which are optimal does not mean they are the best overall approach → importance of co-design

SAR ATR Datasets

MSTAR & SAMPLE

Standard SAR ATR Datasets

- 10 classes
- 128x128 chips

Moving and Stationary Target Acquisition and Recognition (MSTAR)

- 17 degree for training & 15 degree collection angles for training & validation

Synthetic and Measured Paired Labeled Experiment (SAMPLE)

- EM computational tools simulate radar return of CAD models for training data set
- Validation set uses real measured data

	MSTAR		SAMPLE	
Class	17 degrees	15 degrees	Synthetic	Measured
BMP2	698	587	108	108
BTR70	233	196	96	96
T72	691	582	110	110
BTR60	256	195	---	---
2S1	299	274	177	177
BRDM2	298	274	---	---
D7	299	274	---	---
T62	299	273	---	---
ZIL131	299	274	---	---
ZSU234	299	274	177	177
M1	---	---	131	131
M2	---	---	129	129
M35	---	---	131	131
M548	---	---	129	129
M60	---	---	178	178
Totals	3671	3203	1366	1366

SNN Algorithms

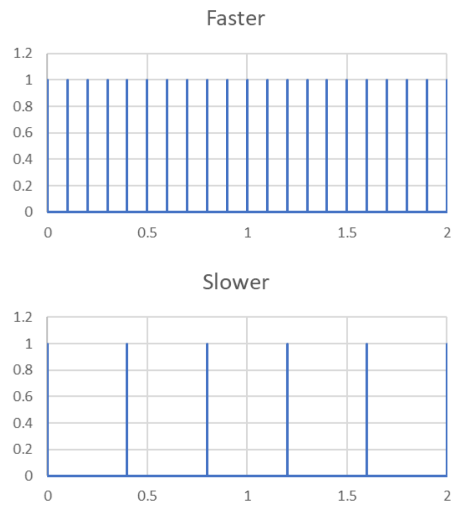
Neural Encodings

Spiking networks can encode information being transferred between neurons in a variety of ways. A few include:

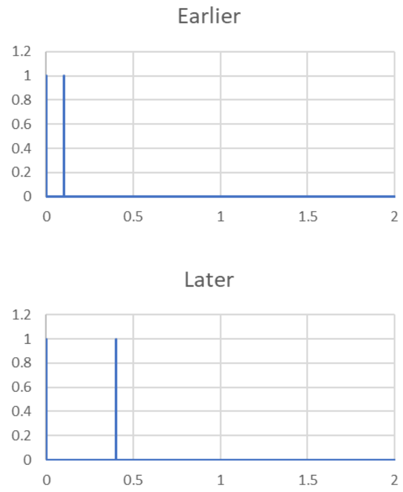
- Individual connections
  - Spike rate
  - Spike delay (time of arrival relative to some other event)
  - Height of spike
  - Energy in spike
  - Spike sequences (e.g. PWM)
- Multiple connections
  - Signal relationships (e.g. synchronizing signal on one input, data on another)
  - Binary encoding
  - Combinatorial encoding

Learning Algorithms

- Biological learning rules & algorithmic learning techniques being pursued



Rate Encoding



Phase Encoding

Whetstone

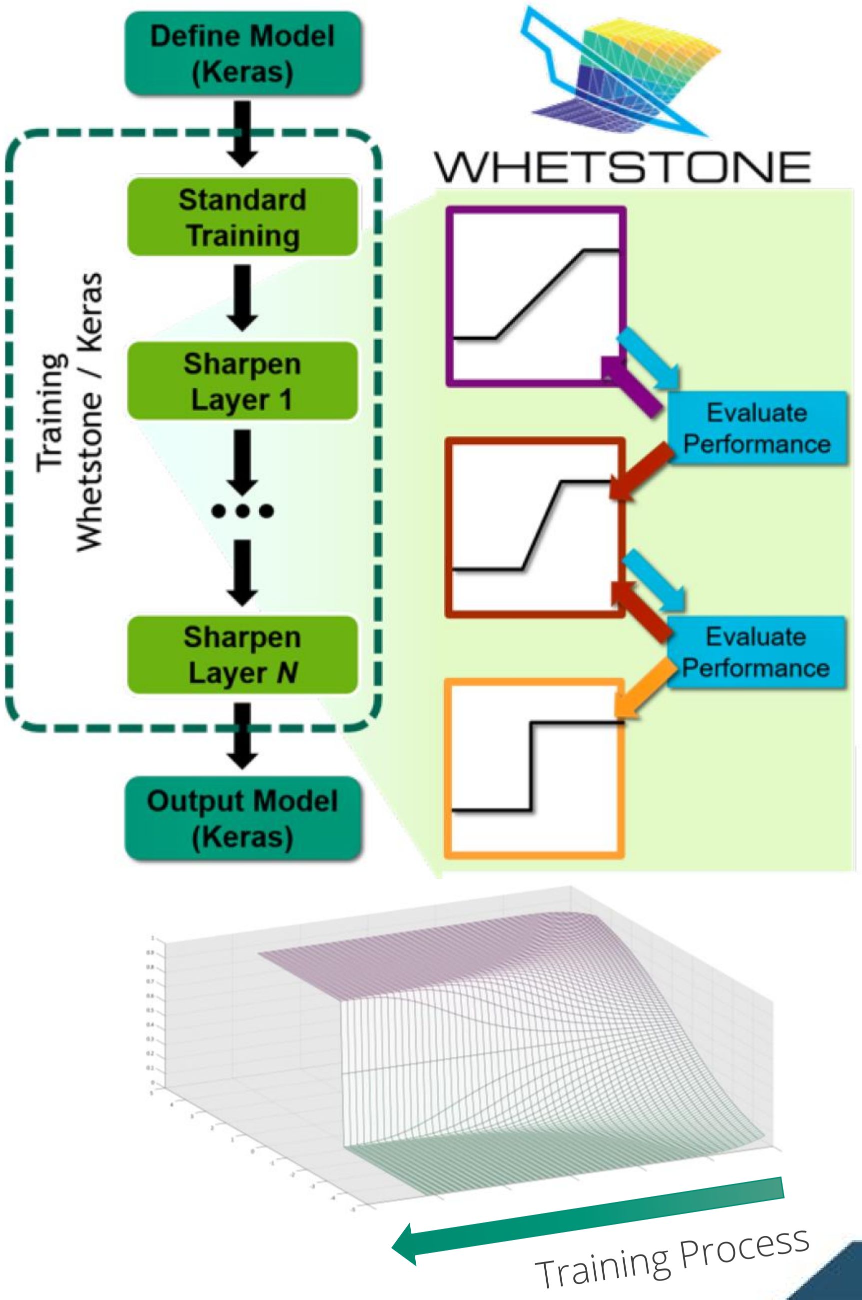
Whetstone provides a drop-in mechanism for tailoring a DNN to a spiking hardware platform (or other binary threshold activation platforms)

- Hardware platform agnostic
- Compatible with a wide variety of DNN topologies
- No added time or complexity cost at inference
- Simple neuron requirements: Integrate and fire

Open Source

https://github.com/SNL-NERL/Whetstone

Severa, William, Craig M. Vineyard, Ryan Dellana, Stephen J. Verzi, and James B. Aimone. "Training deep neural networks for binary communication with the whetstone method." Nature Machine Intelligence 1, no. 2 (2019): 86-94.



Whetstone

Study exploring range of precisions and training configurations

- VGG type CNN with 10 convolution layers, 5 max-pooling layers, 2 dense layers, and a 4-hot output layer
- Not optimized results, but shows ability to explore architectural tradeoffs such as the impact of bit precision
- ‘Acc. X-clip Y-bit’ represents X bits of integer dynamic range, and Y bits of decimal precision
  - Note Intel Loihi architecture has 9 bits (including sign) available so some of these results are incompatible but offer comparison reference
- Comparable to accuracy results from the IBM TrueNorth architecture
  - 95.67% accuracy
  - Renz, M., & Wu, Q. (2017, November). An energy-efficient embedded implementation for target recognition in SAR imageries. In 2017 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 1-5). IEEE.

Property	Network Design											
Batchnorm	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Batchnorm Constraint	None	None	None	0.25	0.5	0.25	0.5	1	None	None	0.5	None
Weight Constraint	1	None	1	0.5	0.5	1	1	1	1	None	0.5	1
Noise	0	0	0.1	0	0	0	0	0	0	0	0.1	0.1
Results												
Pre-Sharp Loss	2.541	2.304	2.54	3.731	6.778	0.201	0.065	3.799	4.291	1.502	1.045	1.221
Pre-Sharp Acc.	0.183	0.086	0.086	0.381	0.28	0.933	0.98	0.271	0.37	0.681	0.77	0.748
Post-Sharp Loss	2.303	2.303	2.54	1.234	0.66	1.184	0.423	0.63	0.931	0.462	0.271	0.586
Post-Sharp Acc	0.086	0.086	0.086	0.696	0.901	0.713	0.94	0.904	0.874	0.934	0.954	0.916
Acc 1-clip 6-bit	0.086	0.086	0.086	0.306	0.256	0.221	0.483	0.16	0.28	0.096	0.614	0.22
Acc 1-clip 7-bit	0.086	0.086	0.086	0.648	0.905	0.761	0.891	0.185	0.234	0.106	0.946	0.24
Acc 1-clip 8-bit	0.086	0.086	0.086	0.662	0.89	0.742	0.905	0.19	0.263	0.121	0.948	0.237
Acc 1-clip 9-bit	0.086	0.086	0.086	0.673	0.916	0.741	0.906	0.18	0.268	0.109	0.943	0.231
Acc 1-clip 10-bit	0.086	0.086	0.086	0.681	0.919	0.734	0.915	0.178	0.265	0.11	0.947	0.233
Acc 2-clip 6-bit	0.086	0.086	0.086	0.306	0.248	0.221	0.535	0.336	0.458	0.254	0.619	0.352
Acc 2-clip 7-bit	0.086	0.086	0.086	0.648	0.897	0.761	0.91	0.357	0.387	0.234	0.948	0.34
Acc 2-clip 8-bit	0.086	0.086	0.086	0.662	0.859	0.742	0.93	0.354	0.414	0.254	0.956	0.373
Acc 4-clip 6-bit	0.086	0.086	0.086	0.306	0.248	0.221	0.535	0.766	0.891	0.864	0.619	0.858
Acc 4-clip 7-bit	0.086	0.086	0.086	0.648	0.897	0.761	0.91	0.914	0.763	0.927	0.948	0.931
Acc 4-clip 8-bit	0.086	0.086	0.086	0.662	0.859	0.742	0.93	0.928	0.881	0.937	0.956	0.915

Intel Loihi neuromorphic architecture has 9 bits (including sign) available so some of these results are incompatible but offer comparison reference

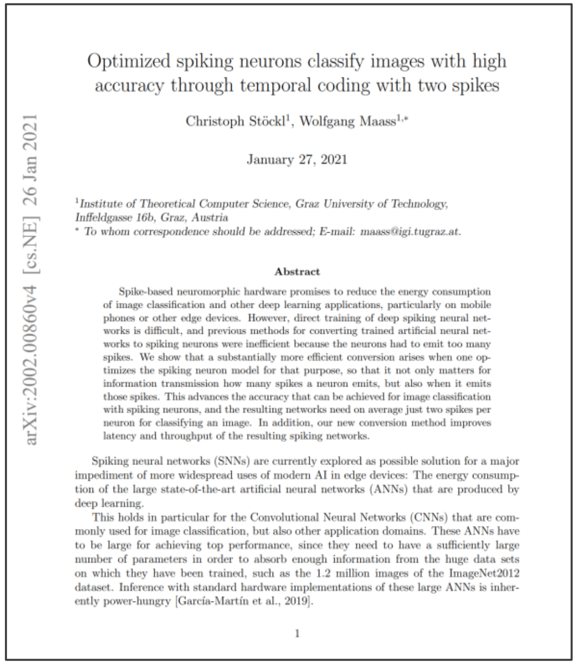
FS-Neuron

Goal: to produce SNNs that achieve similar performance as ANNs with more efficient communication

- Many conversion techniques use a firing rate that requires inefficient communication
- Incompatible with state-of-the-art activation functions

Approach:

- FS = Few Spikes
- Emulates any ANN activation function using a spiking neuron over K time steps
  - Introduces additional approximation properties
- Energy consumption on digital neuromorphic hardware is proportional to the number of spikes which are needed for a computation
  - FS-Neuron method minimizes spikes while maintaining accurate approximation power



Optimized spiking neurons classify images with high accuracy through temporal coding with two spikes

Christoph Stöckl[1], Wolfgang Maass[1,*]

January 27, 2021

[1]*Institute of Theoretical Computer Science, Graz University of Technology, Inffeldgasse 16b, Graz, Austria*
[*] *To whom correspondence should be addressed; E-mail: maass@igi.tugraz.at.*

arXiv:2002.00860v4 [cs.NE] 26 Jan 2021

Abstract

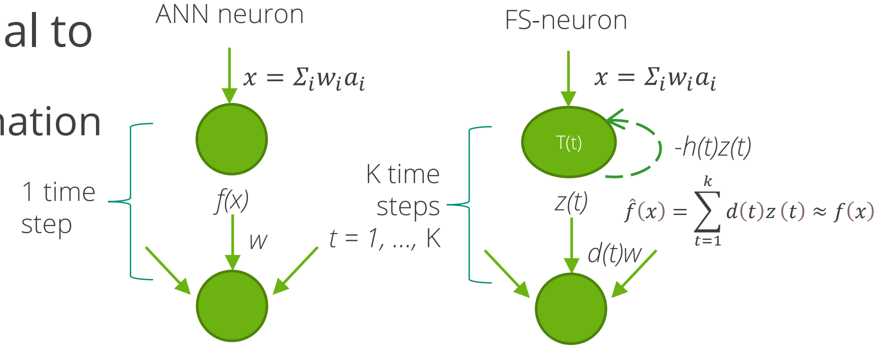
Spike-based neuromorphic hardware promises to reduce the energy consumption of image classification and other deep learning applications, particularly on mobile phones or other edge devices. However, direct training of deep spiking neural networks is difficult, and previous methods for converting trained artificial neural networks to spiking neurons were inefficient because the neurons had to emit too many spikes. We show that a substantially more efficient conversion arises when one optimizes the spiking neuron model for that purpose, so that it not only matters for information transmission how many spikes a neuron emits, but also when it emits those spikes. This advances the accuracy that can be achieved for image classification with spiking neurons, and the resulting networks need on average just two spikes per neuron for classifying an image. In addition, our new conversion method improves latency and throughput of the resulting spiking networks.

Spiking neural networks (SNNs) are currently explored as possible solution for a major impediment of more widespread uses of modern AI in edge devices: The energy consumption of the large state-of-the-art artificial neural networks (ANNs) that are produced by deep learning.

This holds in particular for the Convolutional Neural Networks (CNNs) that are commonly used for image classification, but also other application domains. These ANNs have to be large for achieving top performance, since they need to have a sufficiently large number of parameters in order to absorb enough information from the huge data sets on which they have been trained, such as the 1.2 million images of the ImageNet2012 dataset. Inference with standard hardware implementations of these large ANNs is inherently power-hungry [García-Martín et al., 2019].

1

Stöckl, Christoph, and Wolfgang Maass. "Optimized spiking neurons can classify images with high accuracy through temporal coding with two spikes." *Nature Machine Intelligence* 3.3 (2021): 230-238.



FS-Neuron

	Model	Baseline Train Accuracy	Baseline Validation Accuracy	FS-Neuron Train Accuracy	FS-Neuron Validation Accuracy
MSTAR	EfficientNet-B1	100%	99.187%	100%	96.035%
	ResNet50	100%	99.469%	100%	99.500%
	VGG16	100%	97.281%	100%	96.316%
SAMPLE	EfficientNet-B1	100%	81.473%	99.85%	79.777%
	ResNet50	100%	85.863%	100%	86.543%
	VGG16	100%	89.063%	100%	89.516%

- These models were explored: for model variability, tie to Stöckl & Maass' initial work, & familiarity with performance of these model families
- Higher accuracies possible with optimizations – here we first showing viability of technique
- Care required for neuromorphic hardware compatibility

Highest accuracies for SNNs applied to MSTAR & SAMPLE

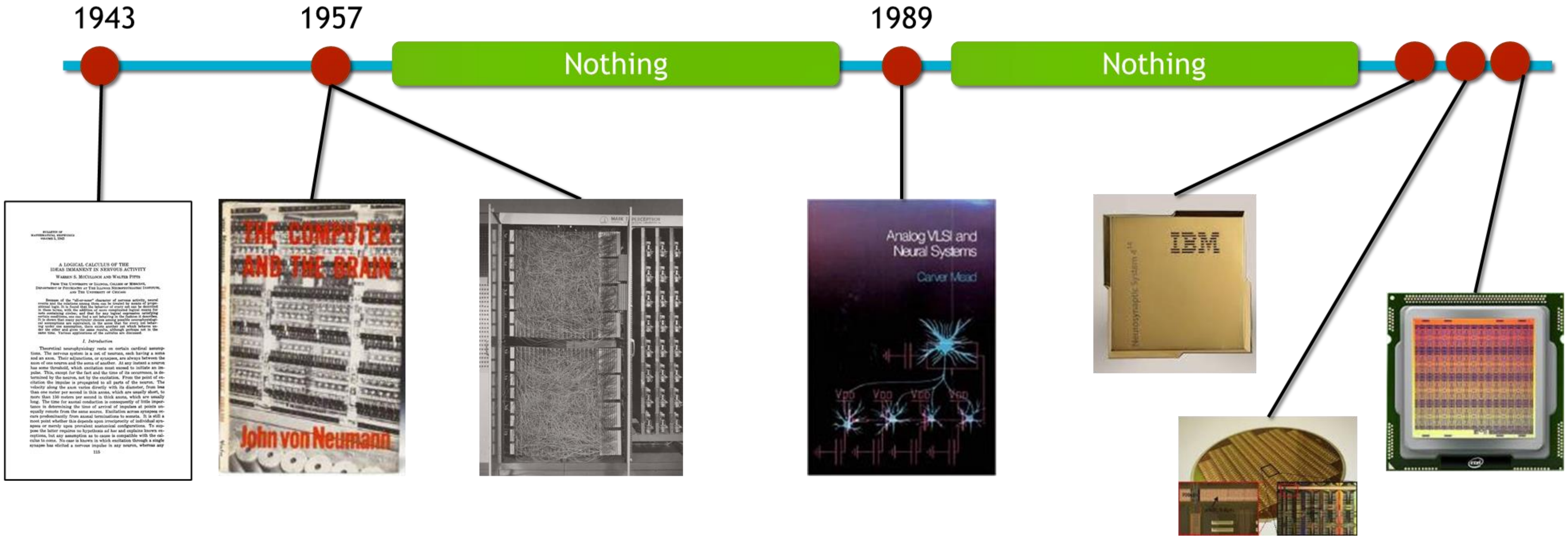
SNN Architectures

Neuromorphic Hardware

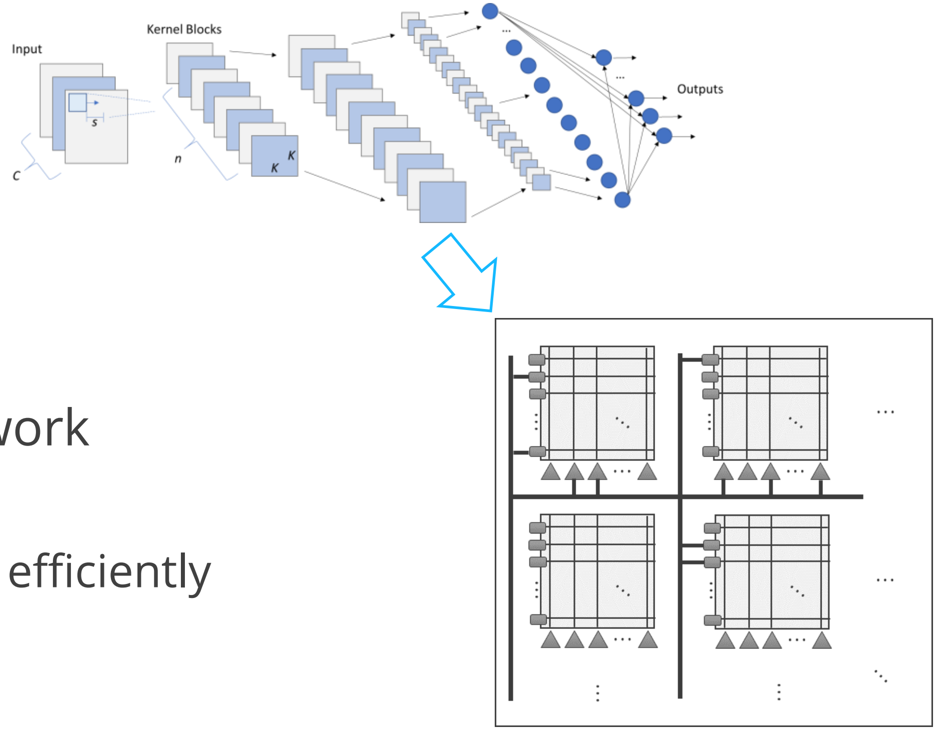
Neuromorphic computing parallels much of the progression of the field of computing at large

- Digital: employs design advances of digital computing, enabling large scale designs while emphasizing sparsity and event driven operation
- Analog: advantages in physics of computation, research pursuing precision and scaling

Forward looking – heterogenous as biological brains employ digital & analog concepts



TrueNorth SAR ATR



IBM TrueNorth

- 1M neuron, 256M synapses - digital neuromorphic architecture
- Composed of 4096 neurosynaptic cores (each with 256 neurons)

EEDN (Energy-efficient deep neuromorphic networks) framework

- Supports implementing CNNs on TrueNorth
- Structures the computation as an all CNN approach and seeks to efficiently map to cores/neurons

Network

- 15 EEDN layers – comparable to a 4 conv layer & 4 pool layer CNN

Preprocessing & Optimizations:

- Center crop 128x128 to 64x64
- Resize to 32x32 (lossy)
- Noise effectively added by mean shift SAMPLE training data by 30

Results -

	SAMPLE
Accuracy	82.857%
Size	4042 cores
Power*	148 mW

*First order energy approximation based upon core counts

Conclusions

Conclusions

While there is an abundance of future work to pursue - the potential of spiking neuromorphic computing for enabling SAR ATR is exciting

- As we've shown - high accuracies are possible

- Neuromorphic computational architectures are amenable for resource constrained deployment

- Further advantages on the horizon include resilience, potential sensor innovation, & further novel algorithm developments in an emerging SNN paradigm

Questions?