SAND2019-3134C

Building a Comprehensive Neuromorphic Platform for Remote Computation

William Severa, Aaron J. Hill, Craig M. Vineyard, Ryan Dellana, Leah Reeder, Felix Wang, James B. Aimone, Angel Yanguas-Gil*

PRESENTED BY

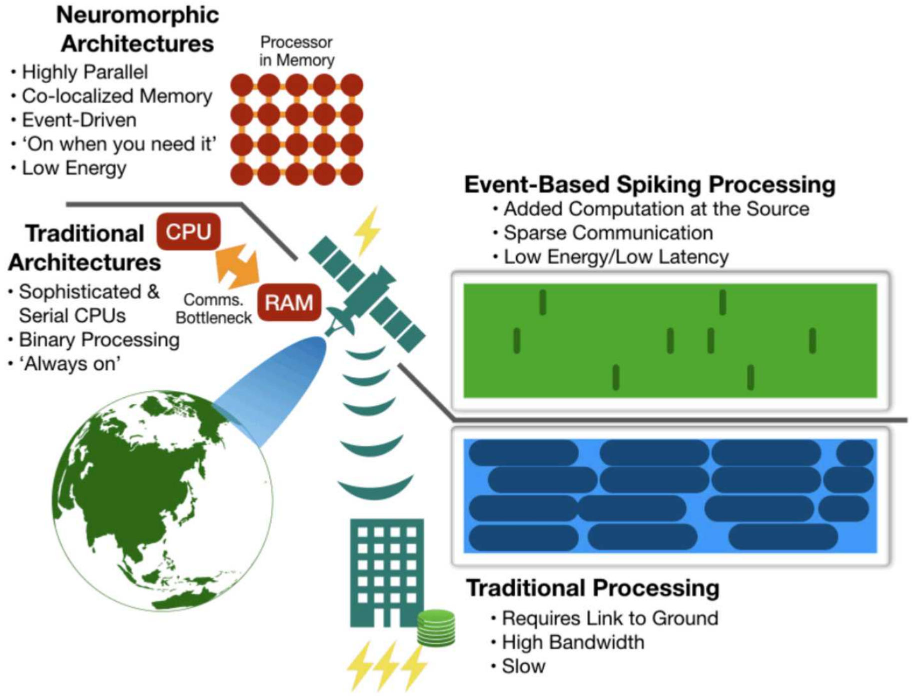
Aaron J. Hill

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.

* Argonne National Laboratory

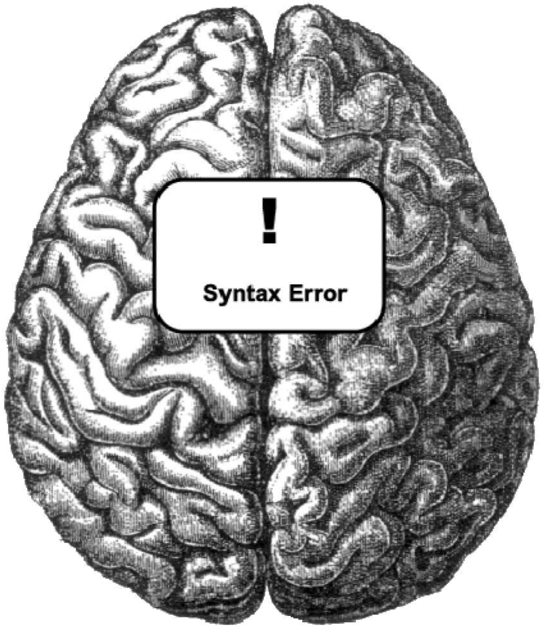
Introduction

- Remote sensor systems care about SWaP, deep learning algorithms do not
- Deep learning results are powerful, but computation platforms have kept them out of reach for edge computing
- Neuromorphic platforms may solve the 'P' problem
- Challenges
  - Algorithm compatibility
  - Programming interfaces
  - At-scale production
- This presentation
  - Porting traditional and learning based algorithms to neuromorphic platforms
  - Flexible and efficient deep learning networks
  - Programming and Performance of various neuromorphic platforms
  - Neuromorphic sensors

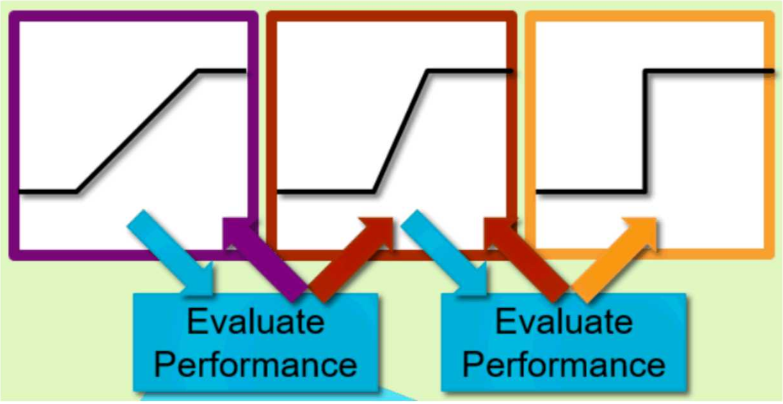


Porting Algorithms to Neuromorphic Platforms

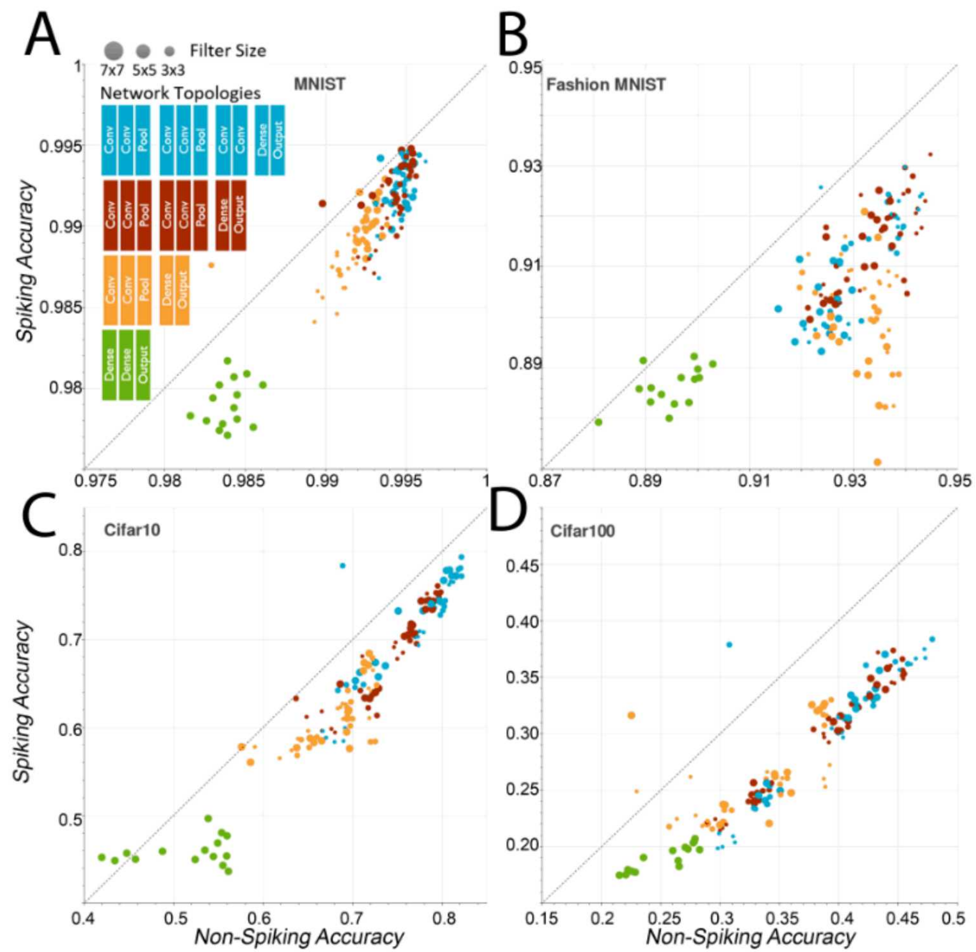
- Classical algorithms are tried-and-tested
- Neuromorphic platforms must meet and exceed classical results
- Neuromorphic has been cornered into learning based algorithms only
- View neurons as highly parallel and simple processors
  - Min, Max, Sorting, Optimization, and Filtering
  - Matrix multiplication
  - 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.



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.

Examples continued…

Aggregation of spikes weighted by their temporally code value

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

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.

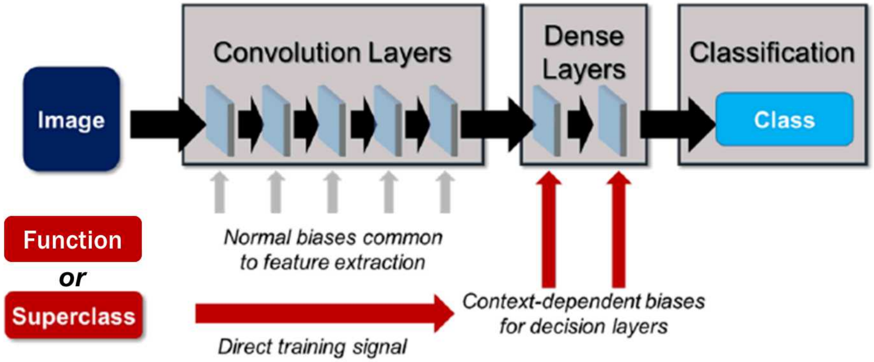
Mode	Chips	Inst.	Theoretical*	Actual* / Overhead◊	Overclocked / Overhead◊
Serial	1	1	30.8 hrs.	31.8 hrs. / 88.7 hrs.	2.8 hrs.† / 59.5 hrs.
Parallel	1	5	6.2 hrs.	6.4 hrs. / 33.9 hrs.	0.6 hrs.† / 27.6 hrs.
Parallel	16	89	20.8 min.	21.0 min. / 4.7 hrs.	4.4 min.‡ / 4.6 hrs.
Parallel	16	110	16.8 min.	– /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

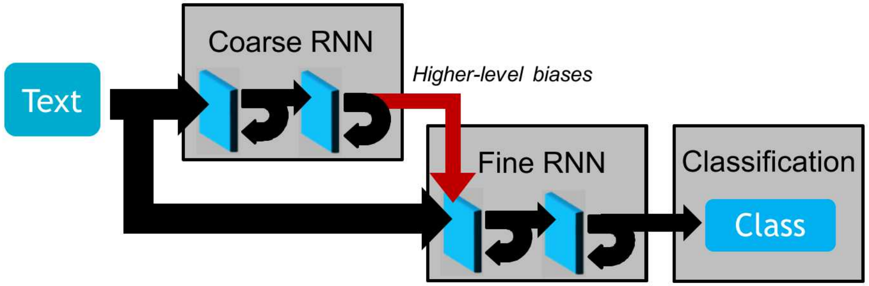
*1 tick = 1ms | †1 tick = 5μs | ‡1 tick = 200μs | ◊Includes I/O

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

Context-Sensitive Deep Learning

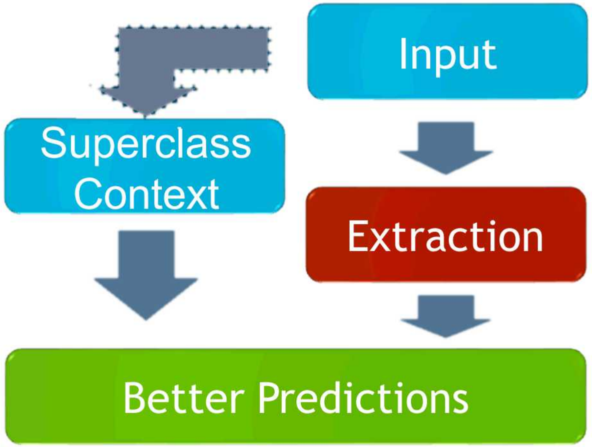
- Provide a network with the flexibility to perform different tasks without reprogramming
- Neuromodulation: The idea that diffuse, network-wide inputs can adjust behavior
  - Contextual information is fed into network through a parallel pathway
  - Context neuromodulation provides a biasing effect on 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



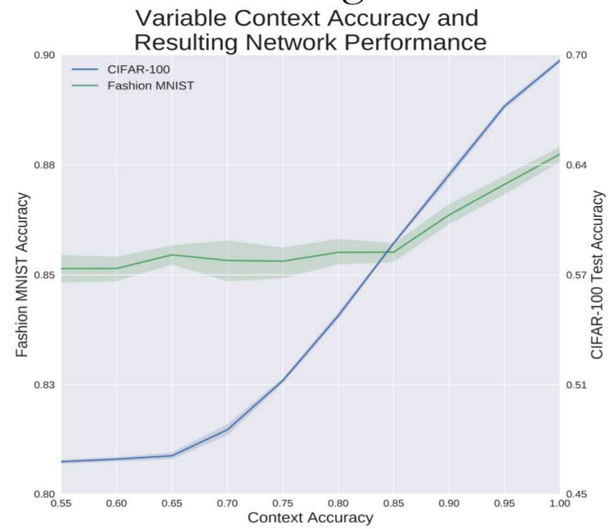


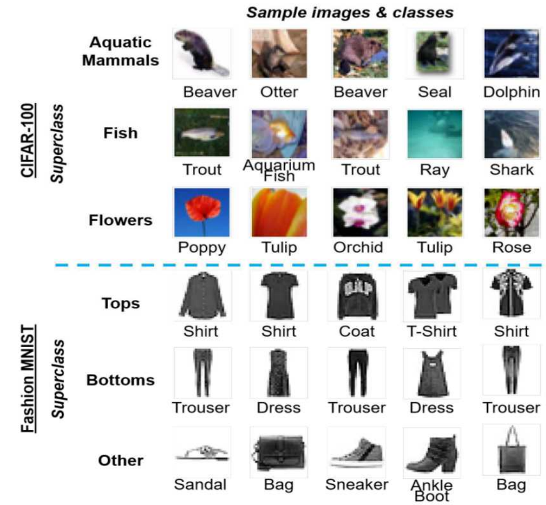
Superclass Exclusion

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



Images: CIFAR100 and FashionMNIST

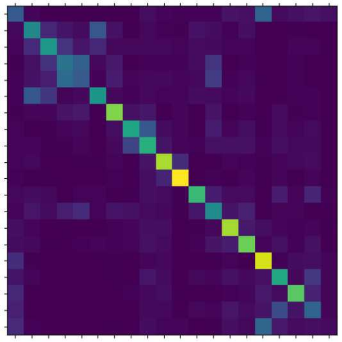




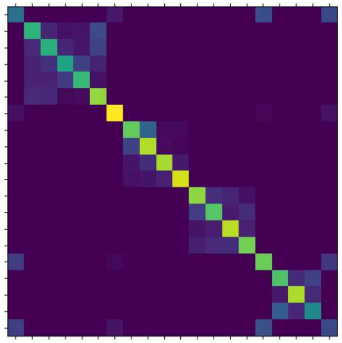
Text: 20 Newsgroups

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

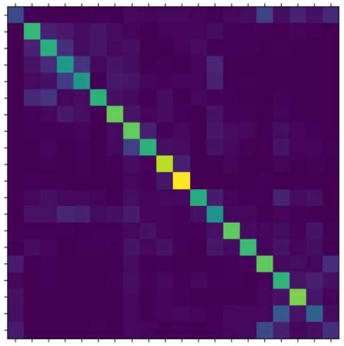
Control (No Context)



Direct (Provided Context)

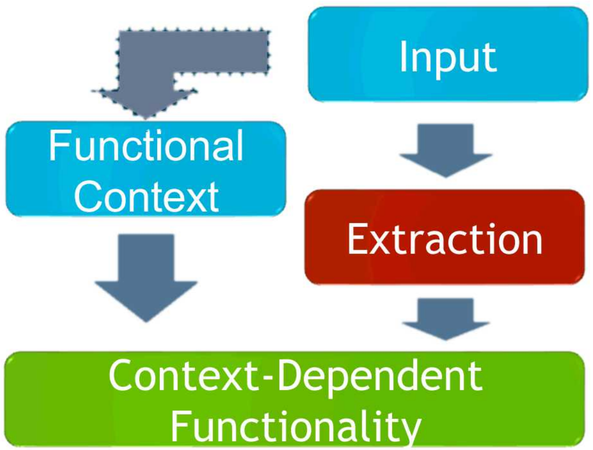


Learned (Inferred Context)

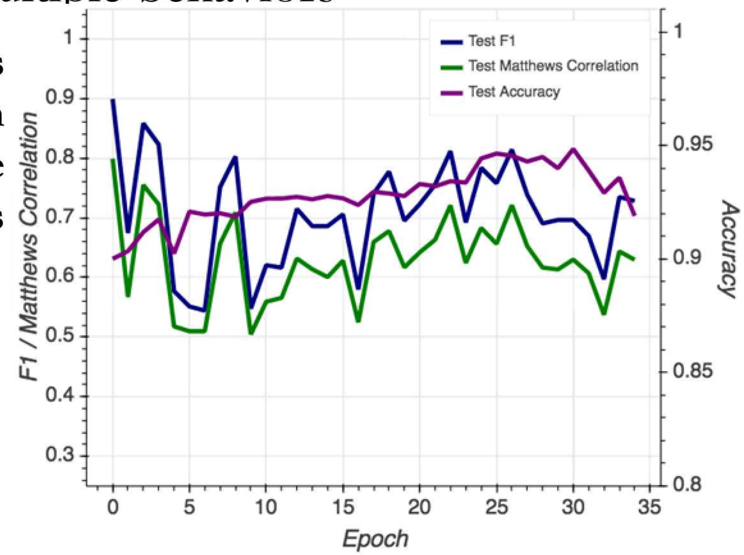


Context-dependent function

Ability of a singular network to incorporate multiple behaviors



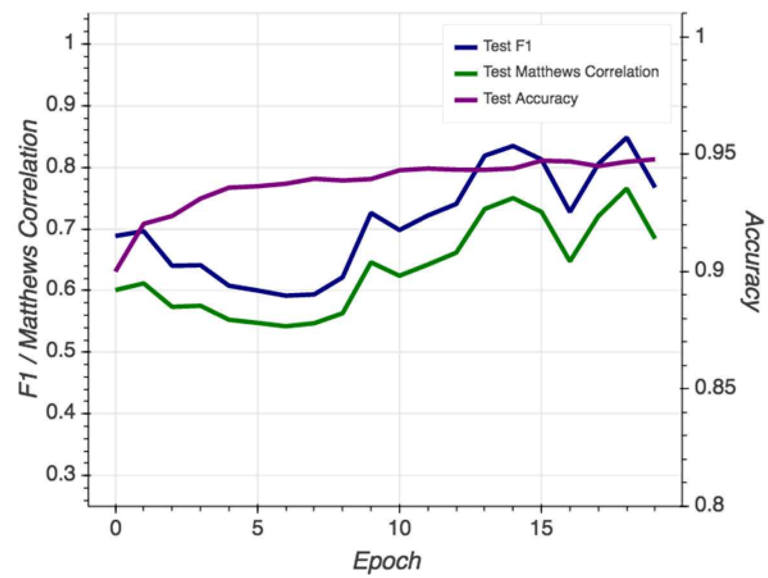
A single network is trained to perform multiple alternative functions



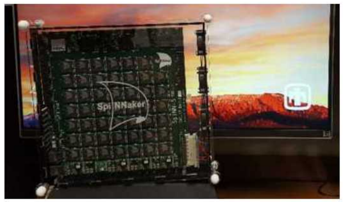
Detecting class 3 vs class 5

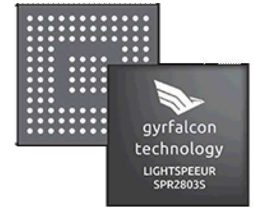
Context	Accuracy
3 (Cat)	.9448
5 (Dog)	.9603
7 (Horse)	.9752
9 (Truck)	.9108

Detecting four separate classes dependent on context

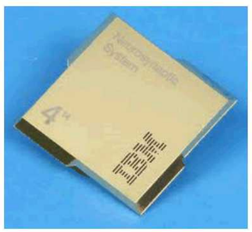


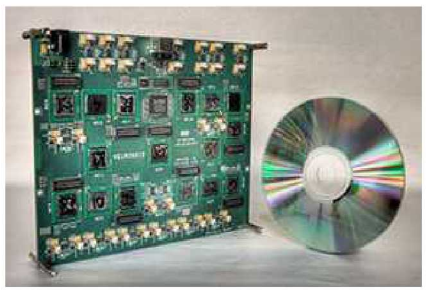
Neuromorphic Hardware

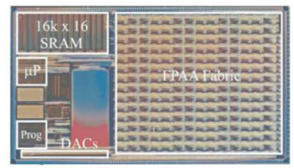




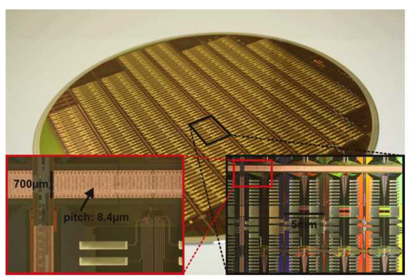
https://www.gyrfalcontech.ai/solutions/2803s/







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

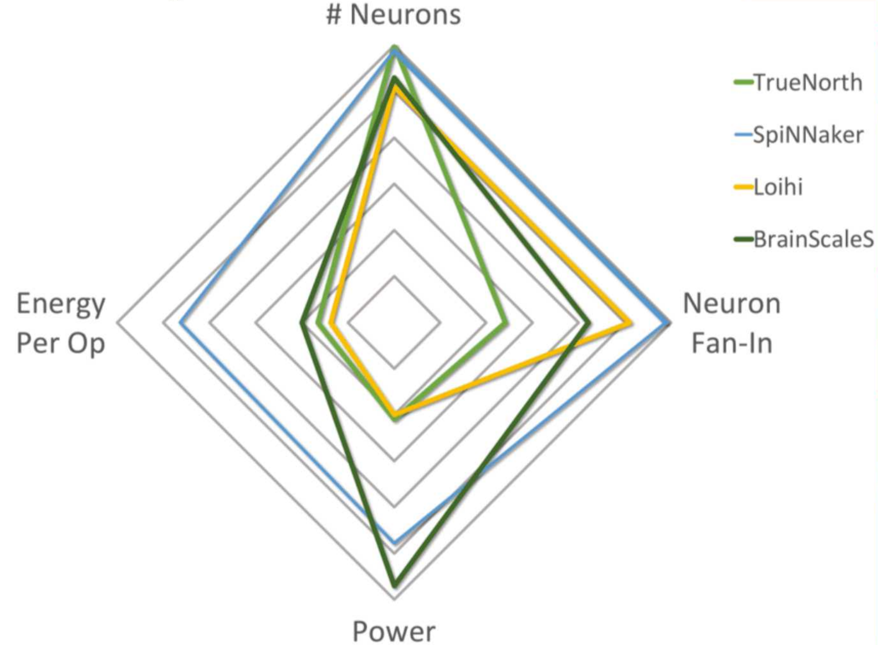


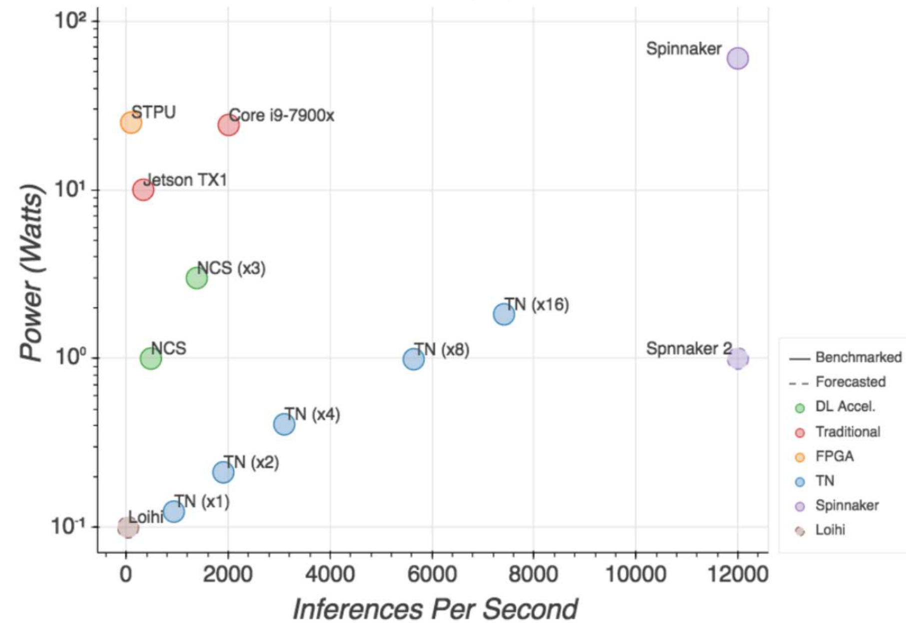
http://www.artificialbrains.com/brainscales

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

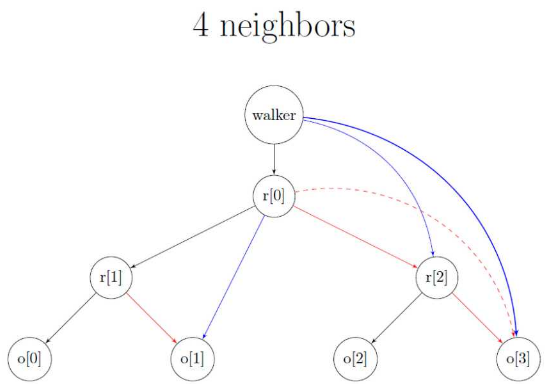
Programming and Performance of Neuromorphic Hardware

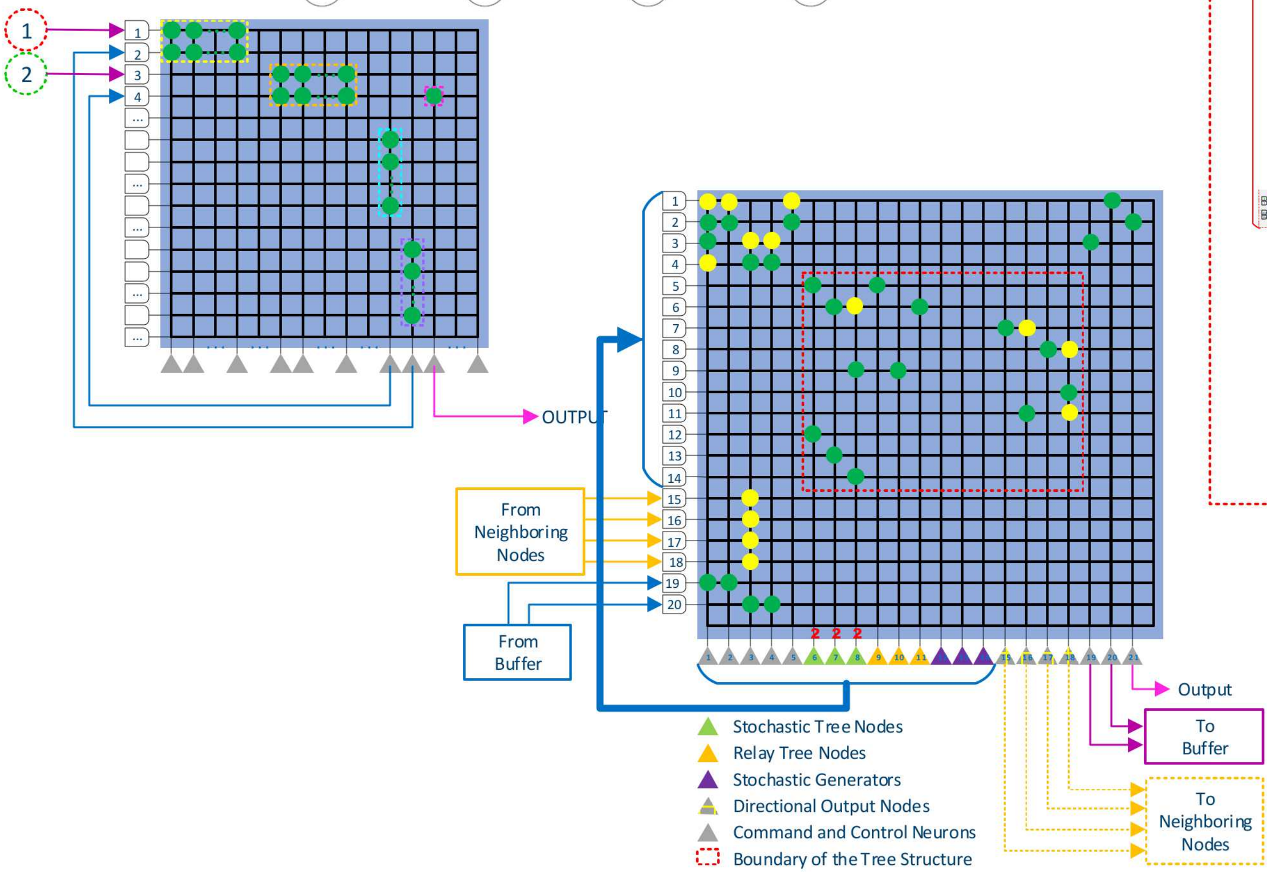
- There are many different emerging neuromorphic architectures
  - Design tradeoffs focus upon different features making them better suited for different applications
  - Architectural differences result in performance differences for different tasks
- Bottom figure shows benchmark results across a suite 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

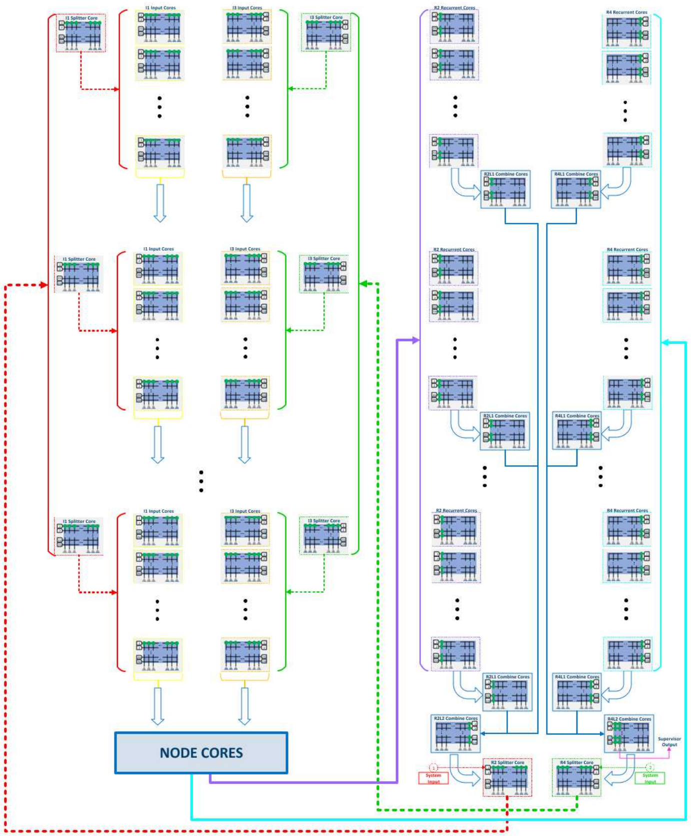




Programming and Performance of Neuromorphic Hardware



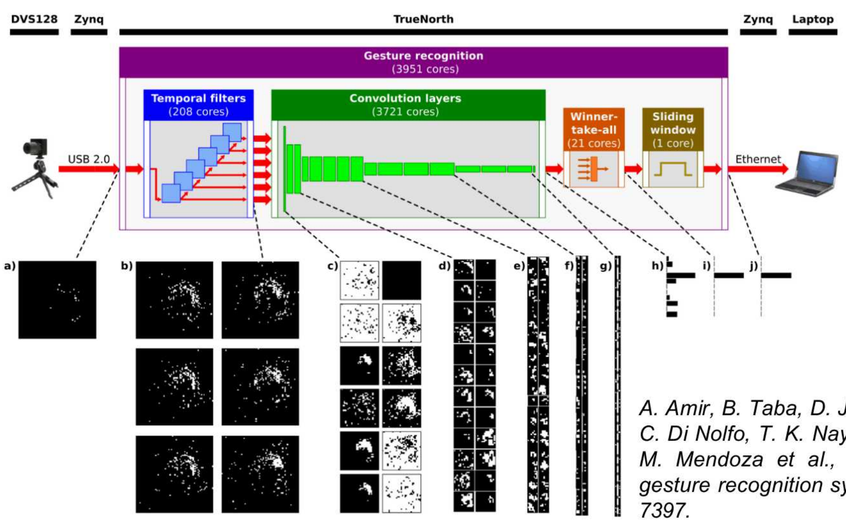


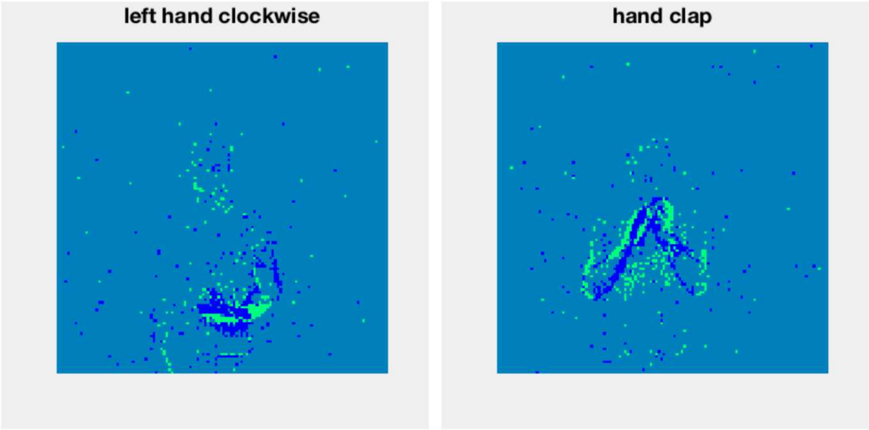


640,000 nodes in an 800x800 2D mesh topology consuming 63,378 TrueNorth Cores across 16 TrueNorth processors

Neuromorphic Sensing

- Traditional neural networks operate on real number valued data.
- Neuromorphic networks operate on spiking data.
- Transduction, the process of converting non spike data to spikes
  - Network and dataset dependent
  - Adds negatively to the overall network performance
- Sensors that produce native spike data outputs are advantages to neuromorphic hardware
  - Dynamic Vision Sensor (silicon retina)
  - Dynamic Audio Sensor (silicon cochlea)





http://research.ibm.com/dvsgesture/

A. Amir, B. Taba, D. J. Berg, T. Melano, J. L. McKinstry, C. Di Nolfo, T. K. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza et al., "A low power, fully event-based gesture recognition system." in CVPR, 2017, pp. 7388–7397.

Conclusion and Discussion

- Deep neural networks are ubiquitous in many fields
- Classical architectures are not ideally suited for these algorithms, especially for resource constrained platforms
- Co-development of algorithms and architecture can efficiently exploit neuro-dynamics
  - Parallelism
  - Sparse event-driven computation
  - Simple computation elements with complex connectivity.
- Neuromorphic platforms offer substantial advantages for sophisticated remote sensing domains while operating within size, weight, and power constraints.