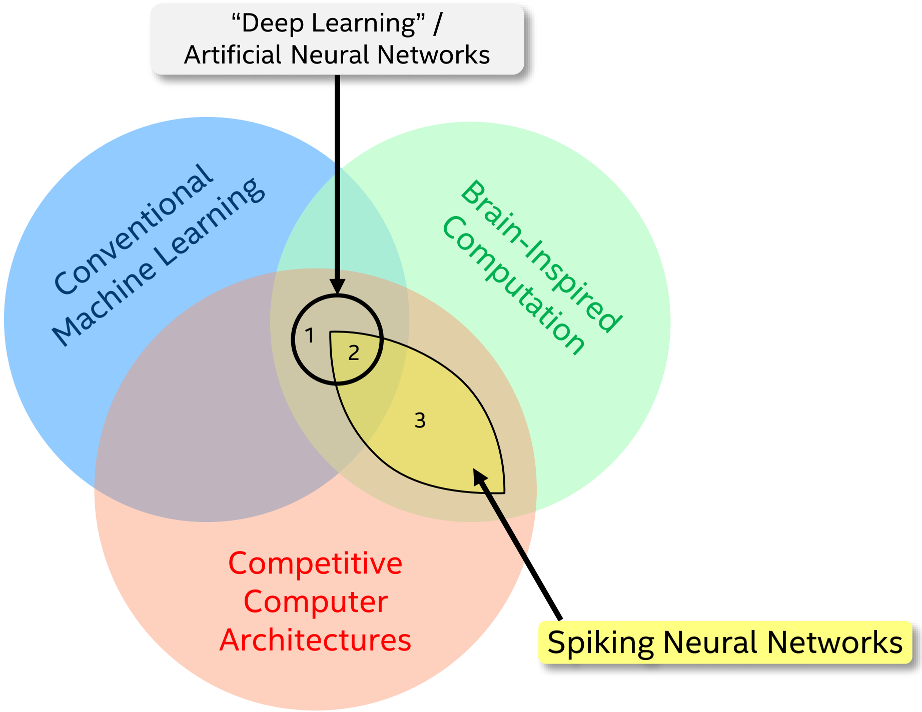
ADVANCING NEUROMORPHIC COMPUTING FROM PROMISE TO COMPETITIVE TECHNOLOGY

Mike Davies
Director, Neuromorphic Computing Lab | Intel Labs

March 27, 2019
Neuro-Inspired Computational Elements Workshop

Neuromorphic Computing Exploration Space



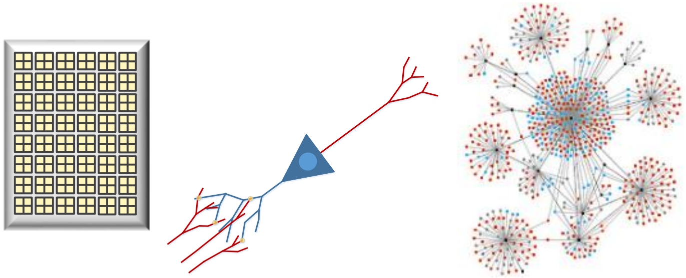
Research Goals:

- **Broad class** of brain-inspired computation
- **Efficient** hardware implementations
- **Scalable** from small to large problems and systems

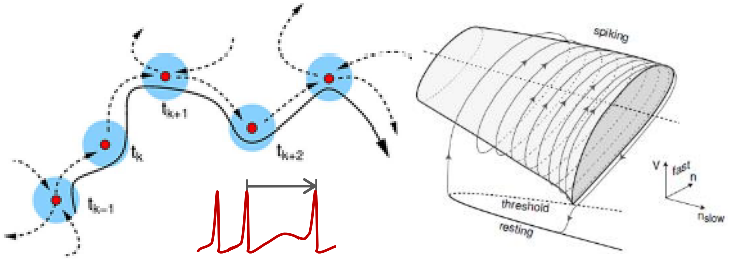
Examples:

- Online and lifelong learning
- Learning without cloud assistance
- Learning with sparse supervision
- Understanding spatiotemporal data
- Probabilistic inference and learning
- Sparse coding/optimization
- Nonlinear adaptive control (robotics)
- Pattern matching with high occlusion
- SLAM and path planning
- Dynamical systems modeling

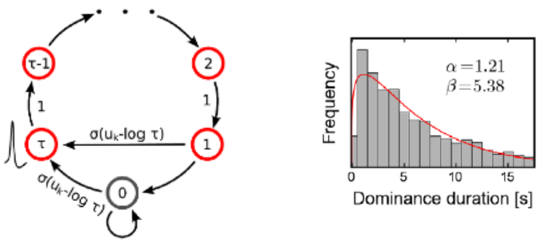
Some Principles of Neural Computation



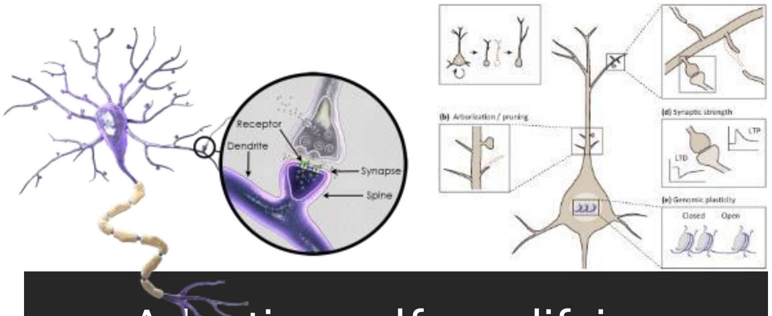
Fine-grained parallelism
with massive fanout



Event-driven computation
with time



Low precision and stochastic



Adaptive, self-modifying

Why Spikes?
Findings from our research

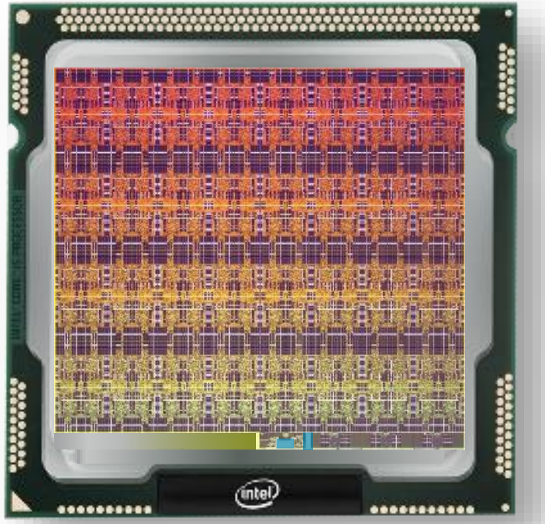
1) Sparse communication in time optimizes energy efficiency (**bits/J vs bits/s**)
2) Spikes efficiently compute many **rate-based models**
3) Spikes provide efficient and natural **processing of temporal data**
4) Spikes support **event-based algorithms** that have nothing to do with rates
5) Spikes (surprisingly) efficiently implement **phasor networks**

In all examples studied so far, benefits vs conventional architectures
increase with increasing problem scale

OUR LOIHI RESEARCH CHIP

KEY PROPERTIES

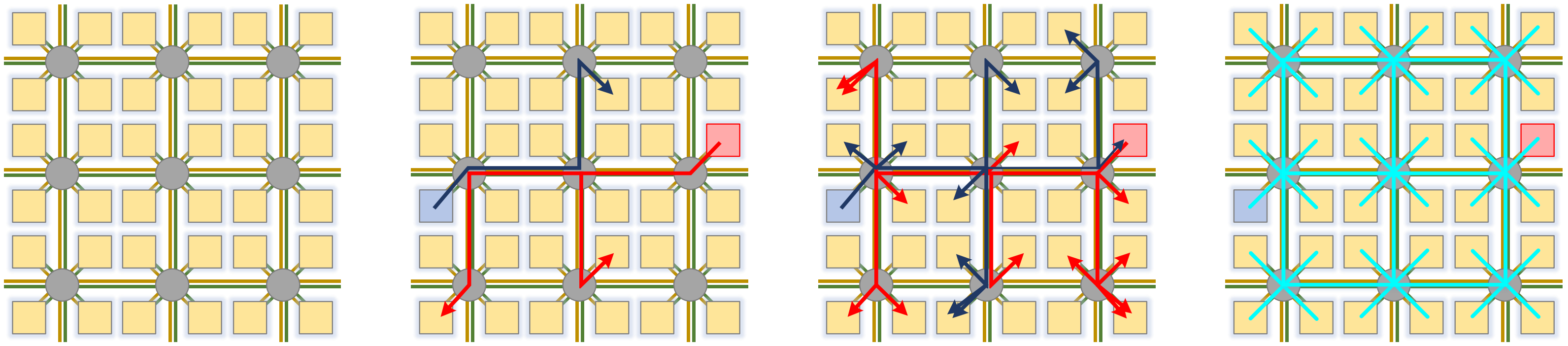
- 128 neuromorphic cores supporting up to 128k neurons and 128M synapses with an **advanced spiking neural network feature set**.
- Supports **highly complex neural network topologies**
- **Scalable on-chip learning** capabilities to support an unprecedented range of learning algorithms
- Fully digital **asynchronous** implementation
- Fabricated in Intel's **14nm FinFET process** technology



Integrated Memory + Compute Neuromorphic Architecture

Davies et al, "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning." IEEE Micro, Jan/Feb 2018.

Mesh Operation: Fine-Grained Synchronization



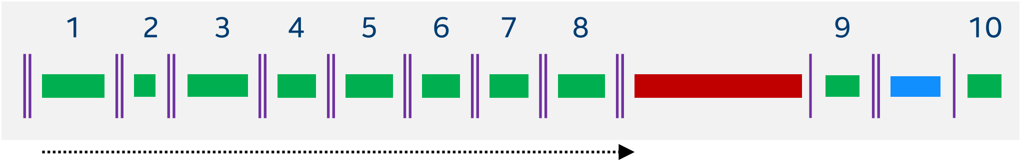
Time step T begins.

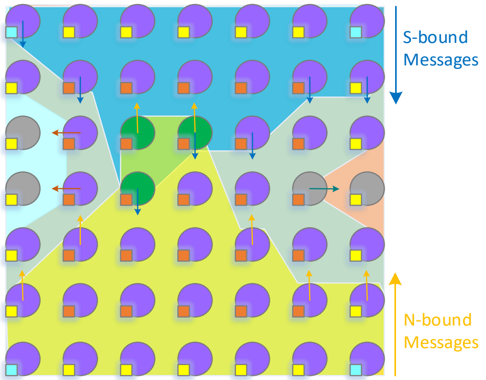
Cores update dynamic neuron state and evaluate firing thresholds

Above-threshold neurons send spike messages to fanout cores

(Two neuron firings shown.)

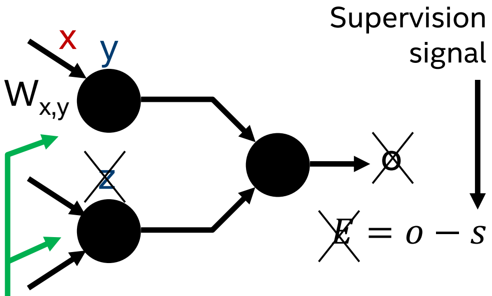
All neurons that fire in time T route their spike messages to all destination cores.





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

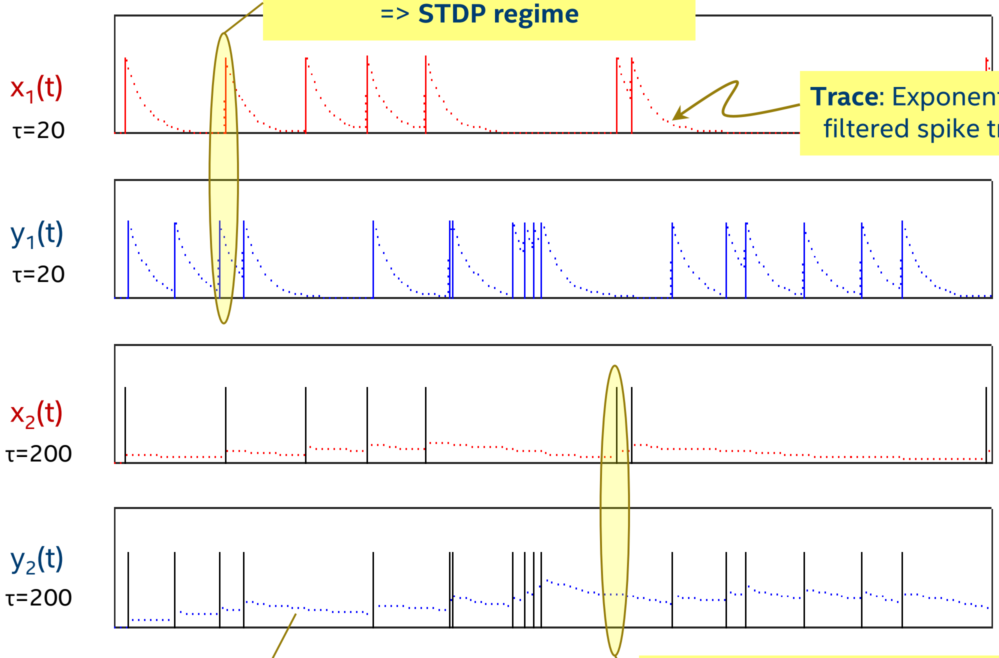


Learning rules for weight $W_{x,y}$ may *only* access presynaptic state x and postsynaptic state y

Reward spikes may be used to distribute graded reward/punishment values to a particular set of axon fanouts

Loihi's Trace-Based Programmable Learning

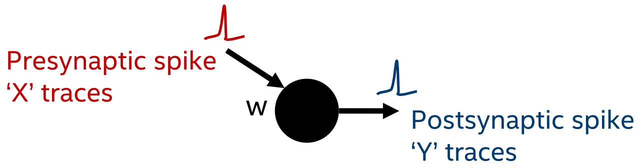
Short time scale trace correlations
=> **STDP regime**



Trace: Exponentially filtered spike train

Traces are **low precision** (7-9b) and may decay **stochastically** for implementation efficiency

Long time scale traces respond to correlations in activity rates



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

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

Synaptic Variables
Wgt, Delay, Tag
(variable precision)

Variable Dependencies
X_0, Y_0, X_1, Y_1, X_2, Y_2, R_1
Wgt, Delay, Tag, etc.

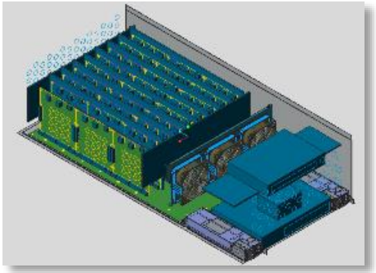
Loihi Systems

Q4 2017
Wolf Mountain
Remote Access
4 Loihi/Board

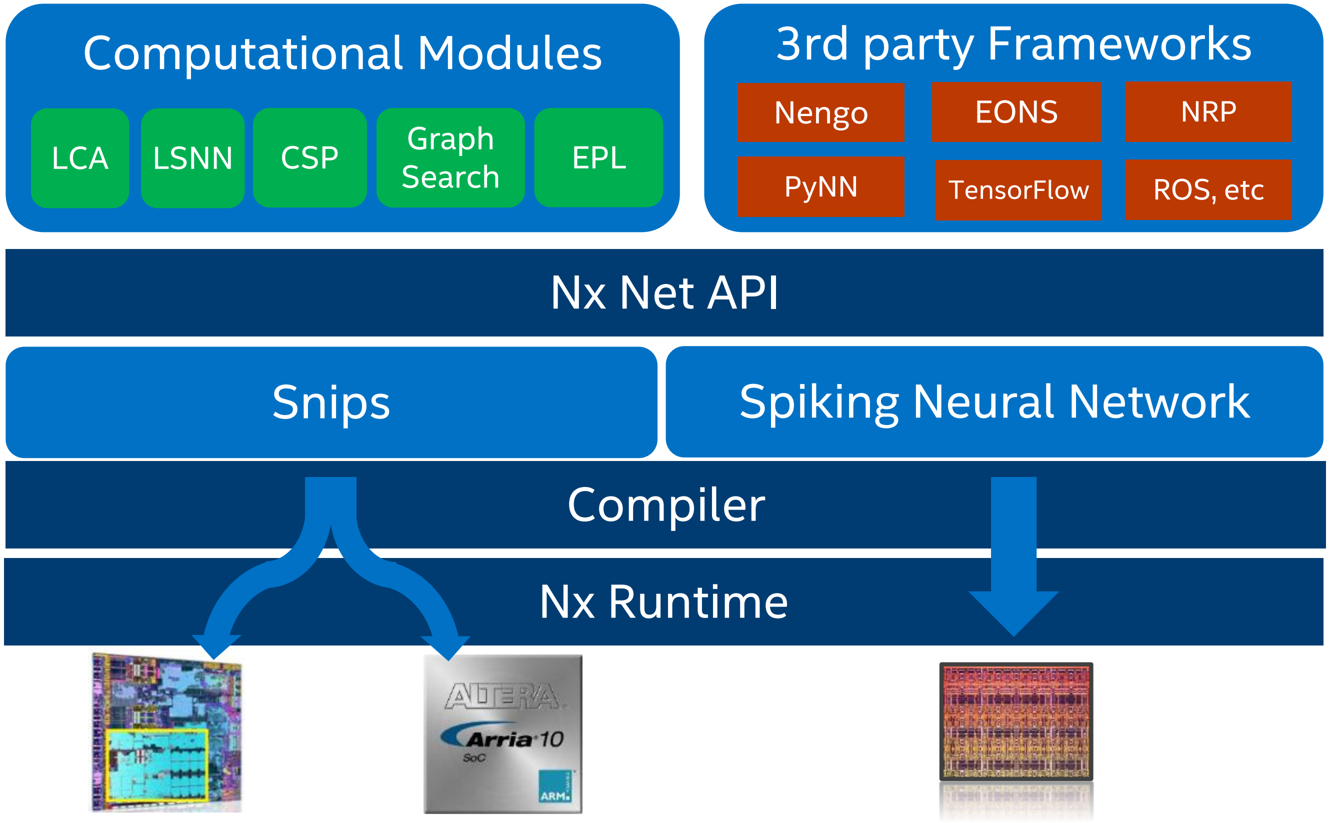
Q2 2018
Nahuku
Arria10 Expansion Board
For cloud & local use
8-32 Loihi/Board

Q3 2018
Kapoho Bay
1-2 Loihi
DVS interface
USB host interface

Q2 2019
Pohoiki Springs
Remote Access
Up to 768 chips
(100M neurons)

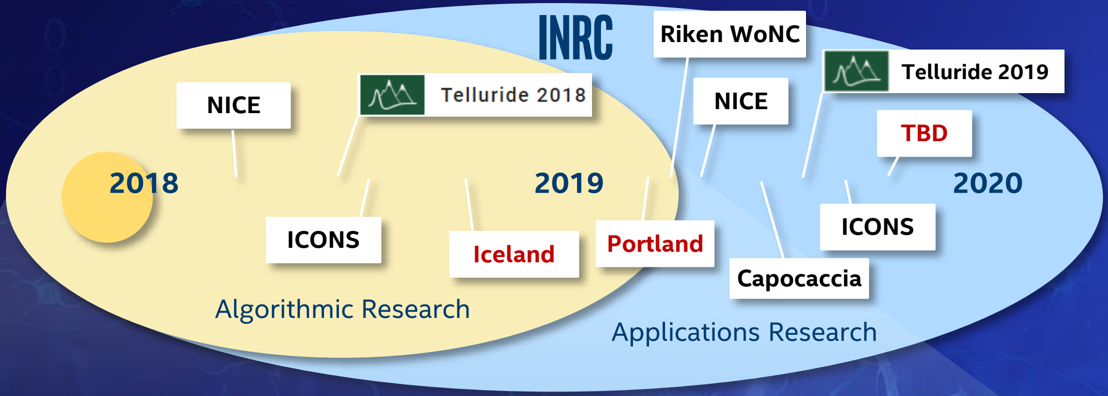


Nx SDK Software Architecture



INTEL NEUROMORPHIC RESEARCH COMMUNITY

Collaborating to Accelerate Progress

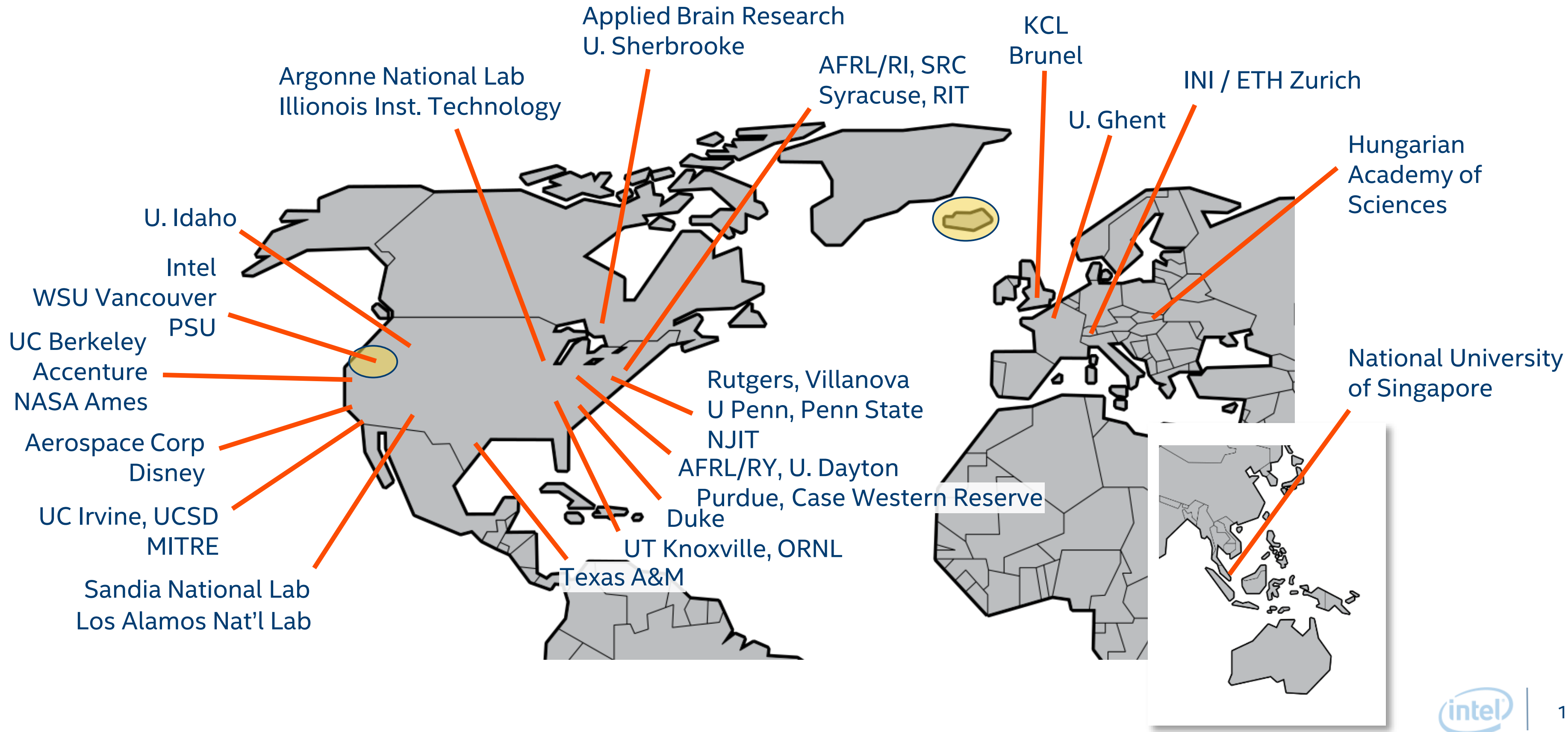


44+ active projects, 50+ organizations

Iceland Workshop (Sep 28 – Oct 2) attended by 62 researchers

Winter Workshop (Feb 11-15) attended by 90+ researchers

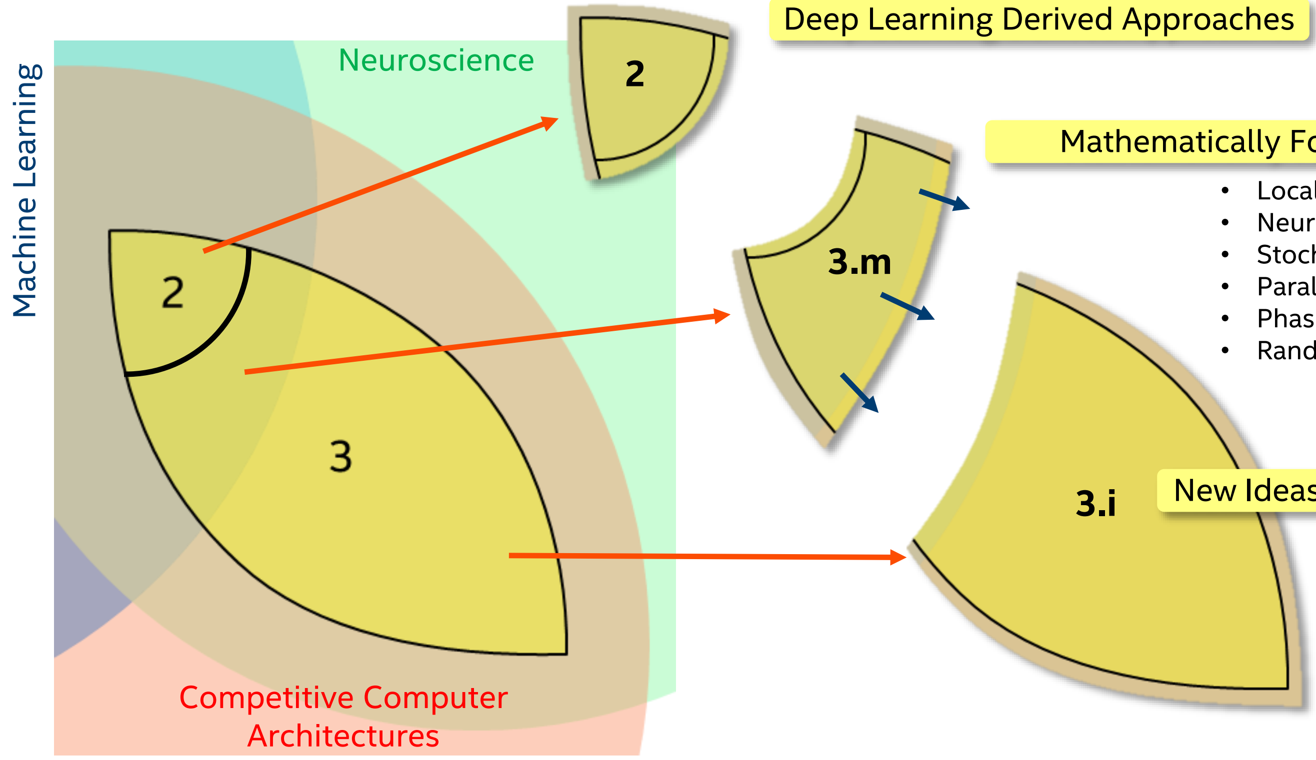
INRC Winter Workshop Attendance



JOIN THE COMMUNITY

E-mail: inrc_interest@intel.com

SNN Algorithms Discovery and Development



Deep Learning Derived Approaches

- DNN -> SNN conversion
- SNN backpropagation
- Online SNN pseudo-backprop

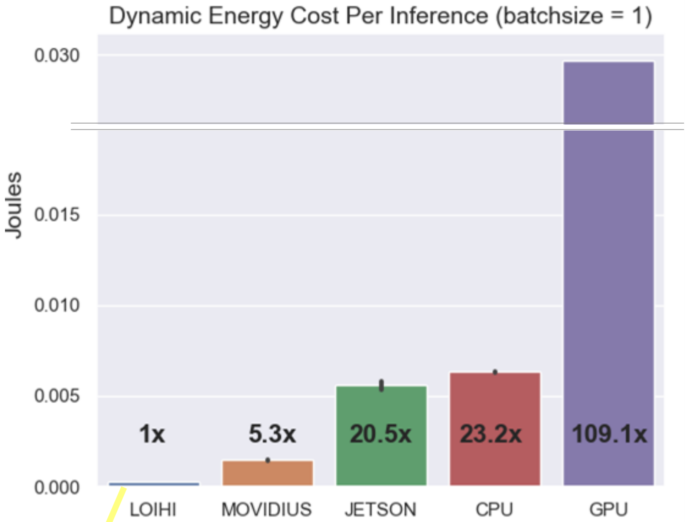
Mathematically Formalized

- Locally Competitive Algorithm for LASSO
- Neural Engineering Framework (NEF)
- Stochastic SNNs for solving CSPs
- Parallel graph search
- Phasor associative memories
- Random diffusion walkers

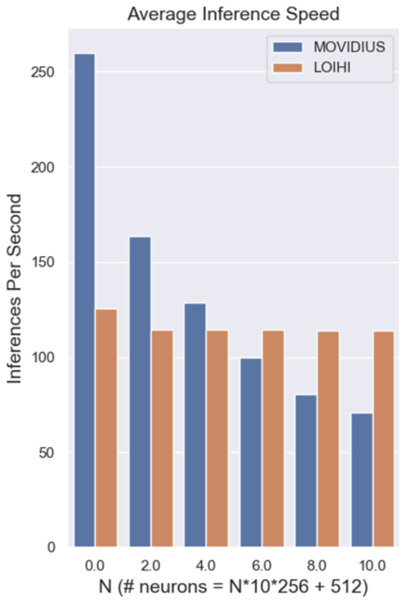
New Ideas Guided by Neuroscience

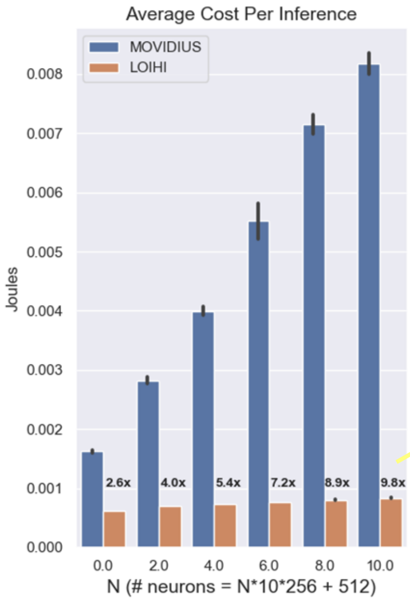
- Olfaction-inspired rapid learning
- Dynamic Neural Fields
- SLAM
- Evolutionary search
- Cortical models

DNN-to-SNN conversion for keyword spotting



Loihi is the most energy-efficient architecture for real-time inference (batchsize=1 case)





Loihi provides extremely good scaling vs conventional architectures as network size grows by 50x

- Loihi provides 5-10x lower energy than closest conventional DNN architecture
- Caveats: batchsize=1 and reduced accuracy (90.6% SNN vs 92.7% DNN)

Results from: Blouw et al, “Benchmarking Keyword Spotting Efficiency on Neuromorphic Hardware.” arXiv:1812.01739

3.m

Case Study: LASSO Sparse Coding

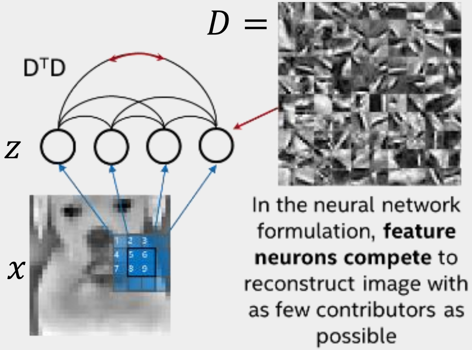
The *Spiking Locally Competitive Algorithm (S-LCA)*

Problem

$$\min_z \frac{1}{2}\|x - Dz\|_2^2 + \lambda\|z\|_1$$

Input — Reconstruction — Sparse regularization

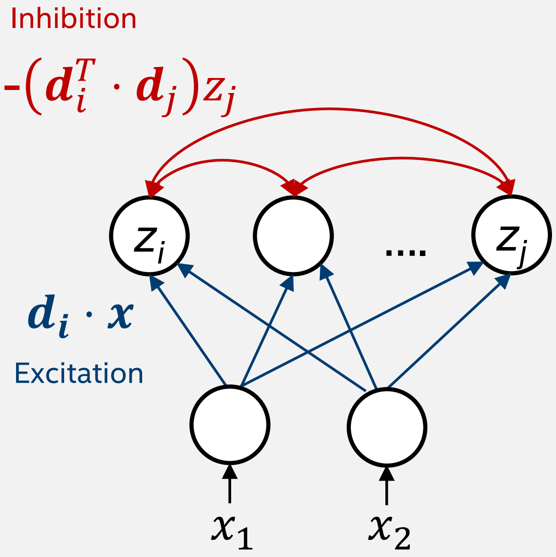
Implementation



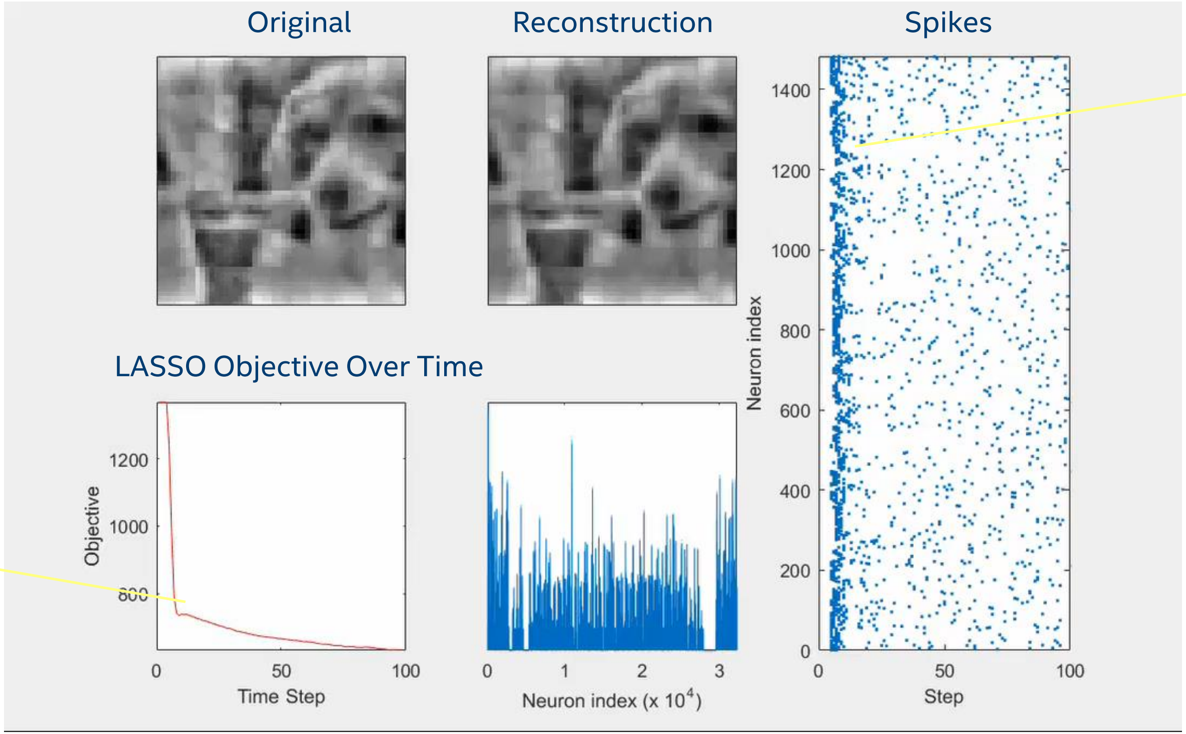
In the neural network formulation, **feature neurons compete** to reconstruct image with as few contributors as possible

Tang et al, arxiv: 1705:05475

Neural Network Structure



Spiking LCA dynamics on Loihi



Intense but very brief period of competition

Much faster convergence on a neuromorphic architecture

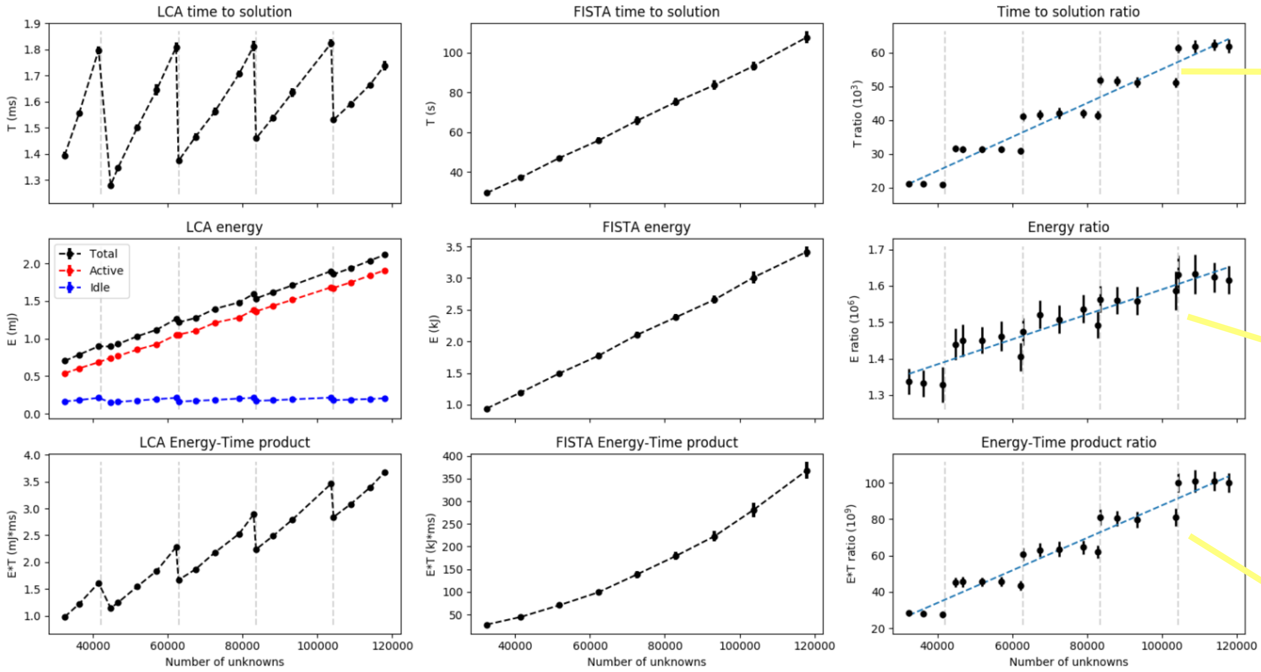
Loihi compared to Core i7 CPU

Loihi | **CPU*** | **CPU/Loihi Ratios**

Time:

Energy:

Energy x Time:



>10,000x faster

~1,000,000x lower energy

$10^{10} – 10^{11}$ lower EDP

* **Intel Core i7-4790 3.6GHz w/ 32GB RAM.** FISTA solver: **SPAMS** http://spams-devel.gforge.inria.fr/
Performance results are based on testing as of December 2018 and may not reflect all publicly available security updates. No product can be absolutely secure.

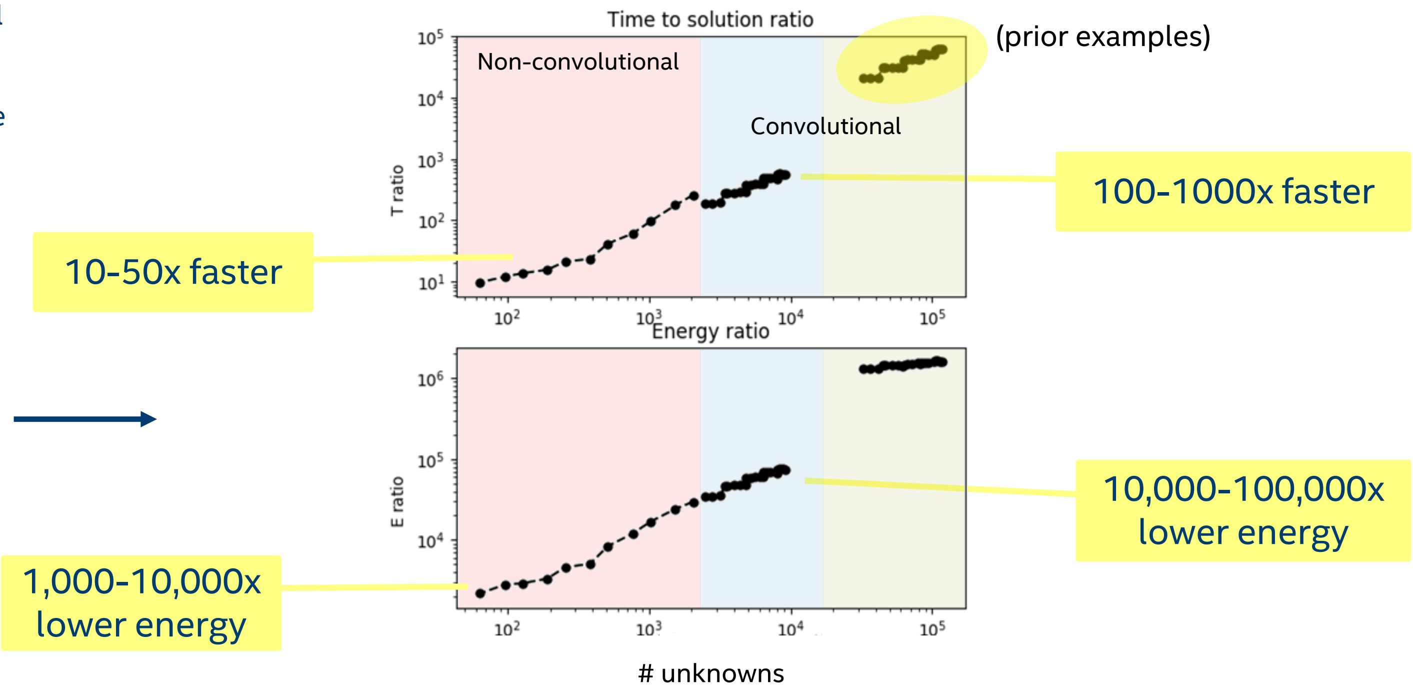
intel

Loihi compared to Core i7 CPU (smaller problems)

Note: Previous examples are all large convolutional LASSO problems that may be unfair to the SPAMS FISTA solver since it includes no optimizations for convolutional problems.

But general scaling trend is clear across small-to-large problems spanning non-convolutional and convolutional examples.

CPU/Loihi Ratios



* **Intel Core i7-4790 3.6GHz w/ 32GB RAM.** FISTA solver: **SPAMS** http://spams-devel.gforge.inria.fr/
Performance results are based on testing as of December 2018 and may not reflect all publicly available security updates. No product can be absolutely secure.

Next Steps: Generalizations & Learning

Unsupervised dictionary learning:

Lin, Tsung-Han, and Ping Tak Peter Tang. 2018. "Dictionary Learning by Dynamical Neural Networks." arXiv preprint. https://arxiv.org/abs/1805.08952.

Yijing Watkins and Garret Kenyon – upcoming NICE talk & poster

Generalization to data manifold learning:

Pehlevan, Cengiz. 2019. "A Spiking Neural Network with Local Learning Rules Derived From Nonnegative Similarity Matching." arXiv preprint. https://arxiv.org/abs/1902.01429.

Hierarchical LCA for adversarial-robust inference:

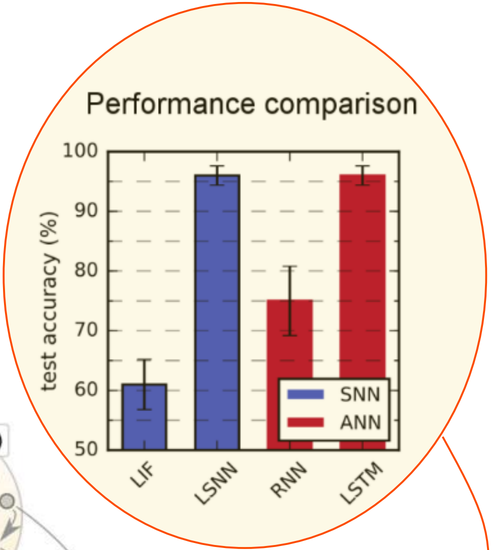
Jacob M Springer, et al. "Classifiers Based on Deep Sparse Coding Architectures are Robust to Deep Learning Transferable Examples." arXiv preprint. https://arxiv.org/abs/1811.07211

Spike-based LSTMs – "LSNNs"

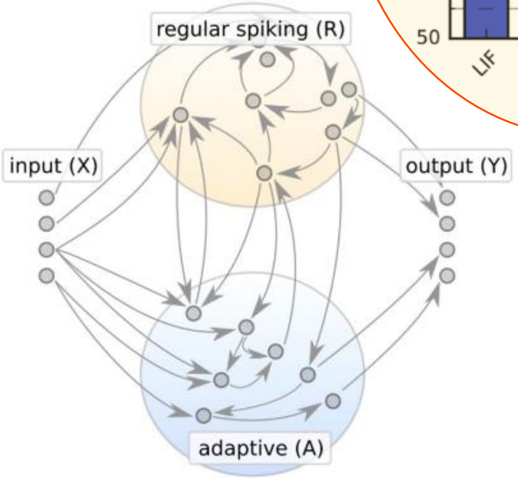
Simple adaptive spiking model achieves LSTM-level accuracy

- SNN reservoir augmented with adaptive neurons
- Thresholds rise on each spike, decay exponentially
  ☞ *Highly energy-efficient adaptation*
- Trained offline with BPTT (TensorFlow)
- Achieves 96% accuracy on sequential MNIST, same as equivalent LSTMs
- **Runs on Loihi today with 94% accuracy**

[Bellec et al, arXiv preprint arXiv:1803.09574]



First case of an **SNN matching LSTM accuracy**



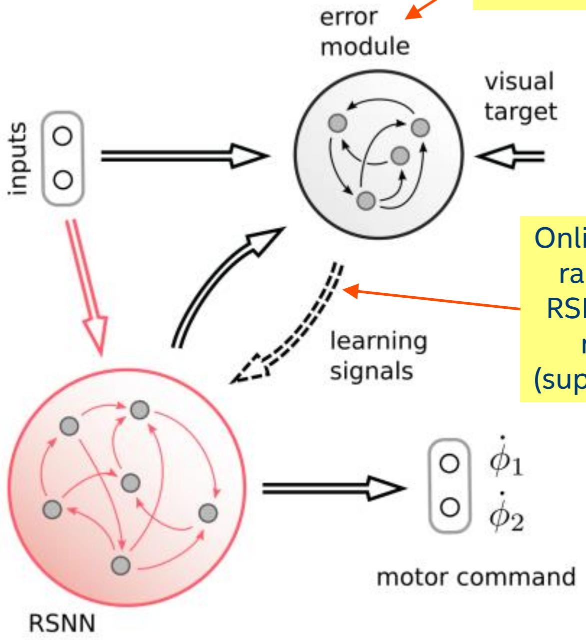
“Neuromorphic Backpropagation”

Numerous promising approaches:

- **Eligibility Propagation**
  Bellec, et al (TU Graz), on arxiv Jan 25, 2019.
- **Surrogate Gradient Learning**
  Mostafa, Neftci, Zenke (Tue/Wed),
  on arxiv Jan 28, 2019.
- **Dendritic cortical microcircuits approximate the backpropagation algorithm**
  J Sacramento, et al. NeurIPS 2018.

Soon we will be able to train **multi-layer** and **recurrent LSNNs** with local three-factor learning rules on Loihi.

Error module is trained offline with BPTT in learning-to-learn framework



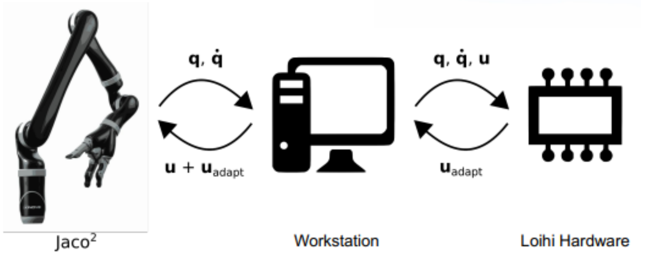
Online error signals rapidly train the RSNN reservoir to match visual (supervised) targets

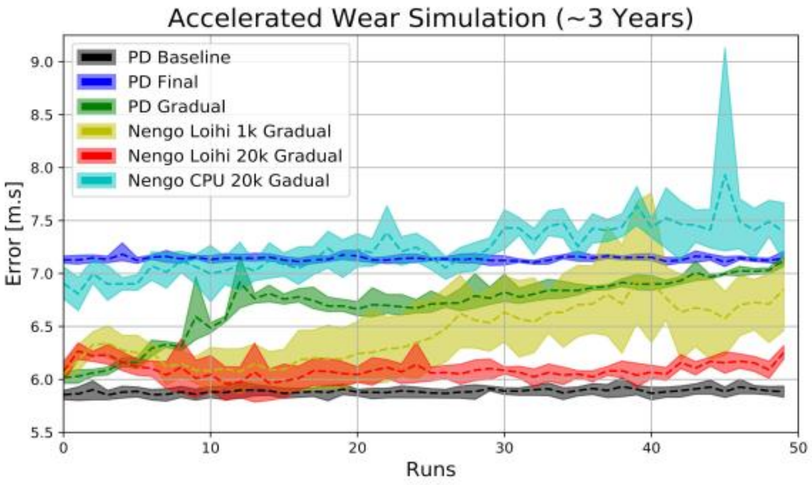
[Bellec et al, arXiv preprint arXiv:1901.09049]

Adaptive Control of a Robot Arm Using Loihi

SNN adaptive dynamic controller implemented on Loihi allows a robot arm to adjust in real time to nonlinear, unpredictable changes in system mechanics[1][2].

Result outperforms standard PD & PID control algorithms.





Different control methods adapting to a gradual, linear increase in friction, over the course of 50 runs. This simulates ~3 years of wear over the course of 16.67 minutes of run time, a 90K times speed up. Only 20K neurons on Loihi is able to successfully cope with this perturbation.

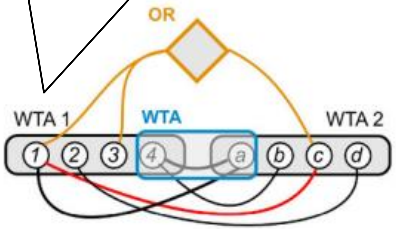
[1] DeWolf, T., Stewart, T. C., Slotine, J. J., & Eliasmith, C. (2016, November). A spiking neural model of adaptive arm control. In *Proc. R. Soc. B* (Vol. 283, No. 1843, p. 20162134). The Royal Society.
[2] Eliasmith, "Building applications with next generation neuromorphic hardware." *NICE Workshop 2018*

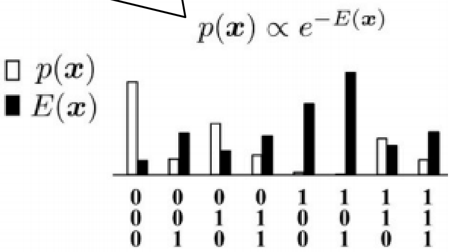
Solving Constraint Satisfaction Problems

SNN with noise stochastically searches to find the minimum energy solution:

Variables represented by Winner-Take-All (WTA) circuits

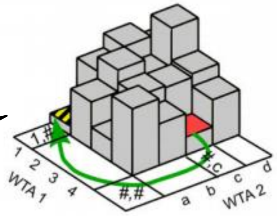
Minimization ⇔ Sampling from probability distribution $p(x)$



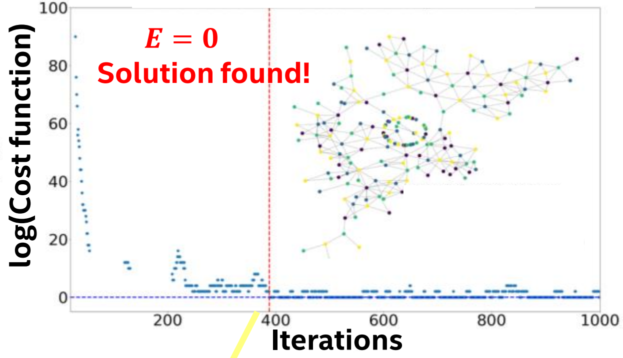


Encode constraints into interconnectivity between WTAs

Stochastic search via SNN enables faster convergence than pure gradient dynamics



Example: 4-coloring of world map



$\approx 10\mu s$/step results in $\approx 4ms$ time to solution.

WIP: Self-checking validation network to stop execution when solutions are found.

Jonke et at., "Solving Constraint Satisfaction Problems with Networks of Spiking Neurons." Front. Neurosci. 2016

Graph Search – Path Planning

Runtime comparison to best Djikstra optimizations:

- Neuromorphic: $O(L \cdot \sqrt{V})$
- Standard: $O(E)$

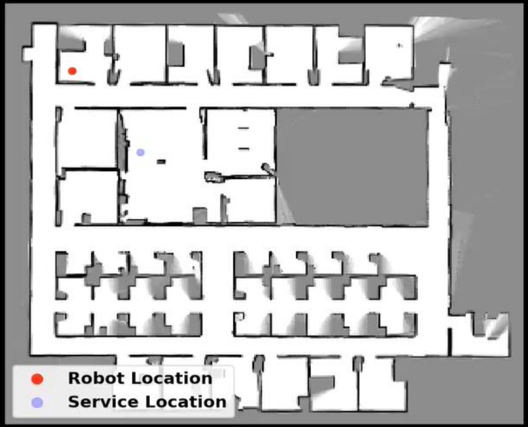
For most nontrivial problems:

- L<<E
- V<<E

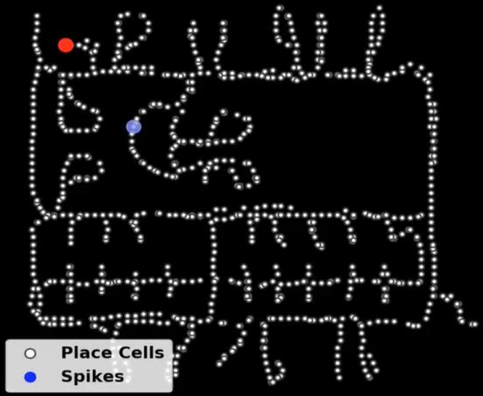
Neuromorphic solution uses *fine-grain parallelism* an *temporal wavefront-driven computation* to potentially provide great performance gains for large problems.

Based on *Ponulak F., Hopfield J.J. Rapid, parallel path planning by propagating wavefronts of spiking neural activity. Front. Comput. Neurosci. 2013. V. 7. Article № e98.*

Robot Motion



Loihi Representation



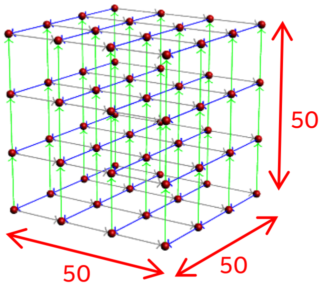
DARPA SDR Site B
(Data from Radish Robotics Dataset)

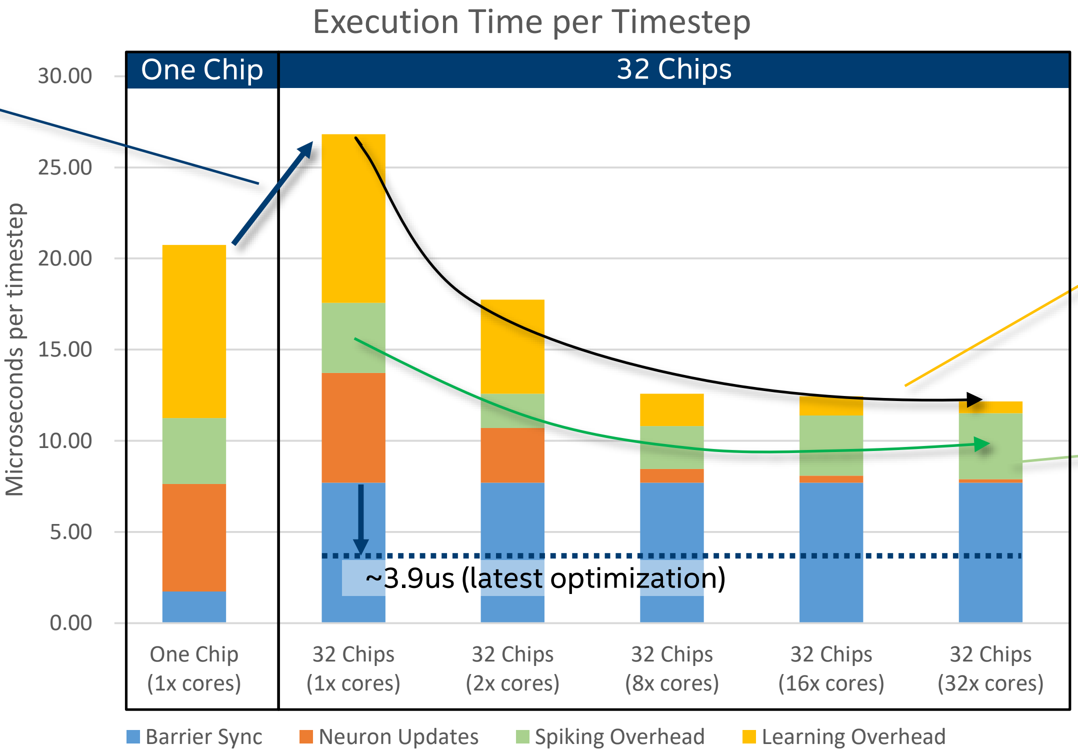
Graph Search on Nahuku (32-chip Loihi System)

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

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

50x50x50 3D lattice





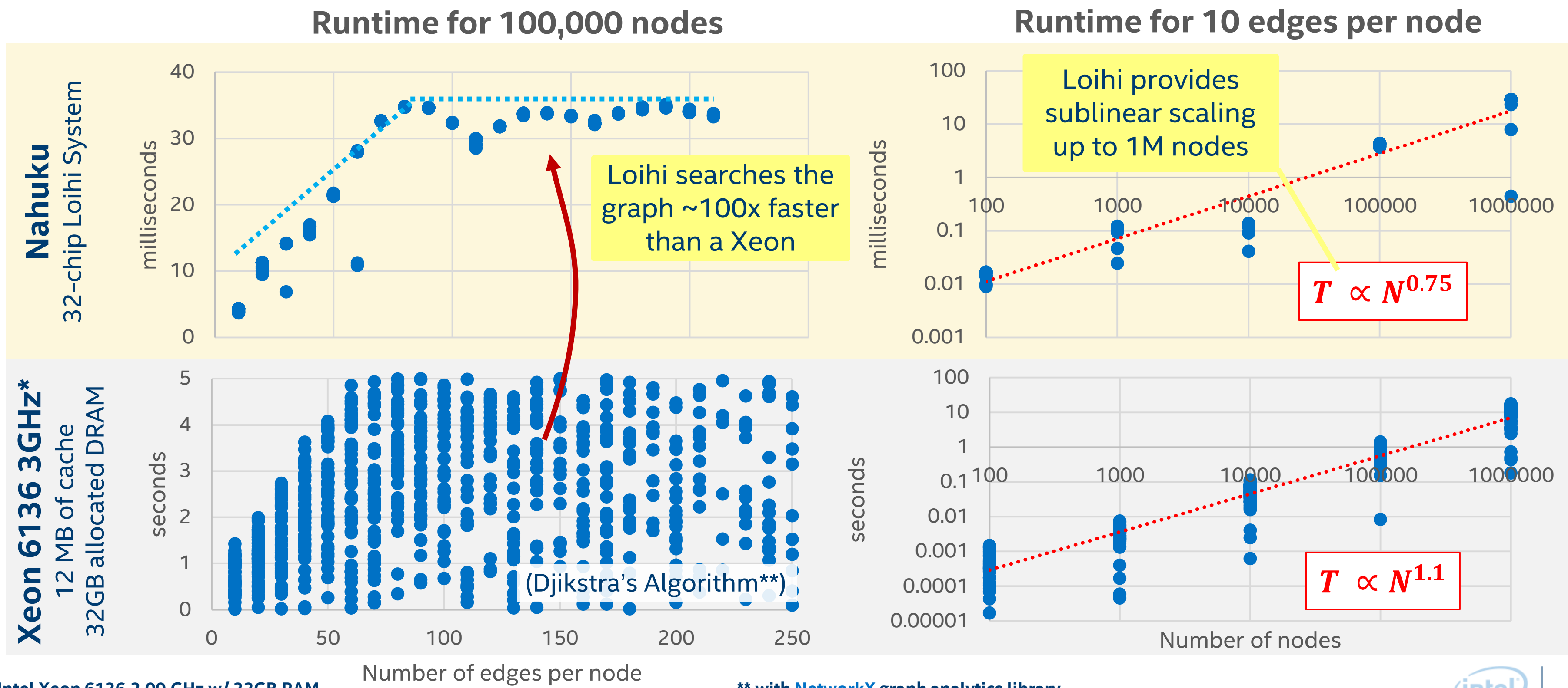
Learning overhead **decreases** with increasing core parallelism

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.

Searching Small World Networks with Loihi

Watts-Strogatz network model with rewiring probability 20%.



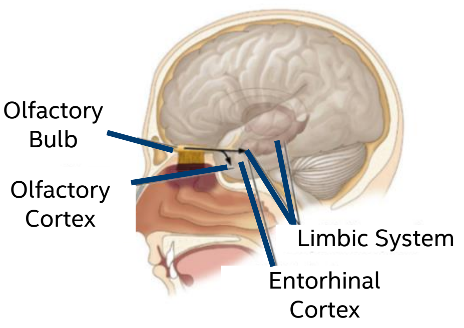
* Intel Xeon 6136 3.00 GHz w/ 32GB RAM.

** with NetworkX graph analytics library

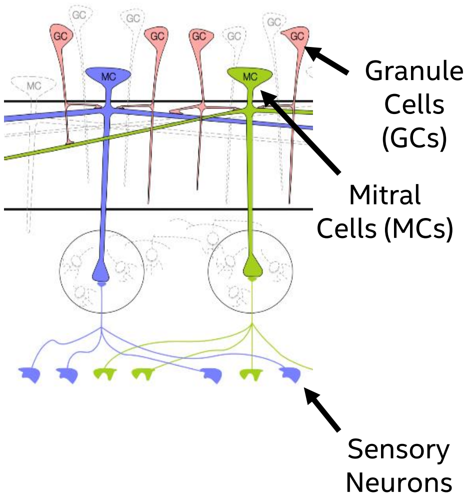
Performance results are based on testing as of December 2018 and may not reflect all publicly available security updates. No product can be absolutely secure.

Olfaction-Inspired One Shot Learning

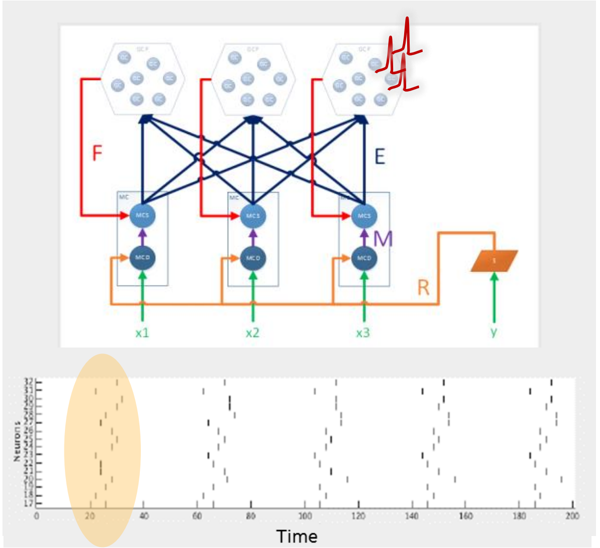
Olfactory System



Olfactory Bulb Neural Circuit



Spatiotemporal Attractor Model



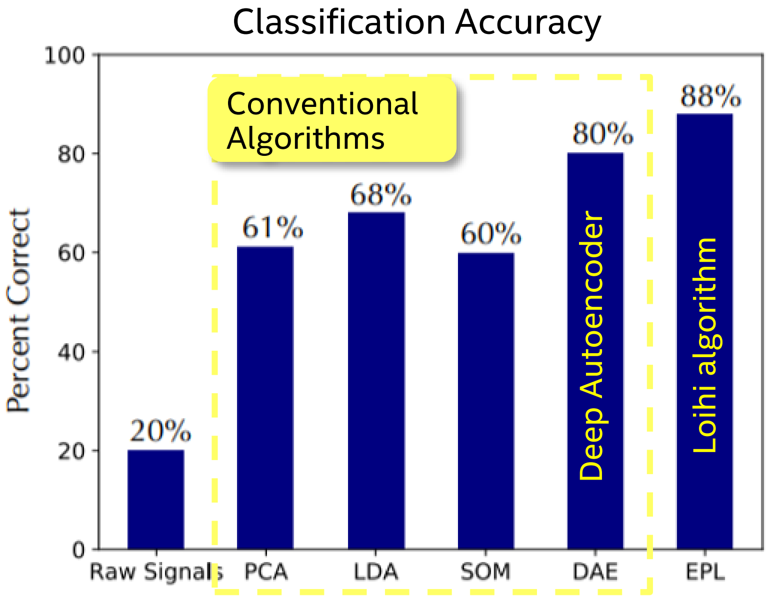
Nabil Imam (Intel) with Thomas Cleland (Cornell) – submitted

Outperforms Conventional Algorithms

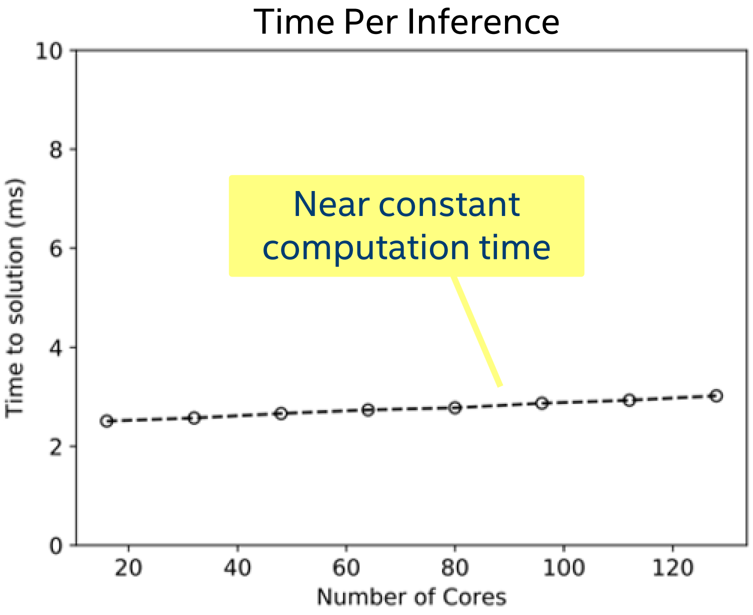
Provides average of **8% accuracy improvement** vs deep autoencoder

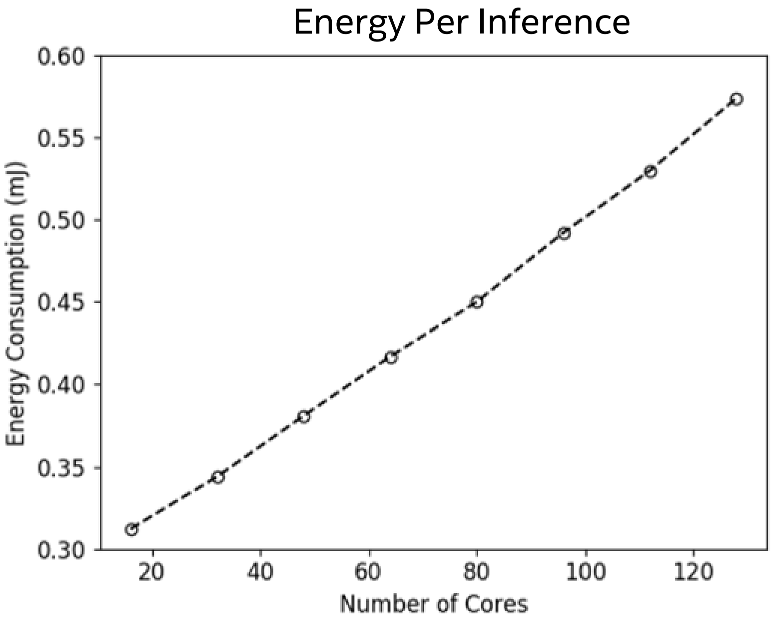
40x more data efficient learning vs backpropagation

Supports **online learning** (robust to catastrophic forgetting)



Excellent Scaling to Larger Network Sizes





Performance results are based on testing as of December 2018 and may not reflect all publicly available security updates. No product can be absolutely secure.

Phasor Neural Networks

An emerging paradigm for SNN computation?

Idea: Represent neural activities with **complex numbers**

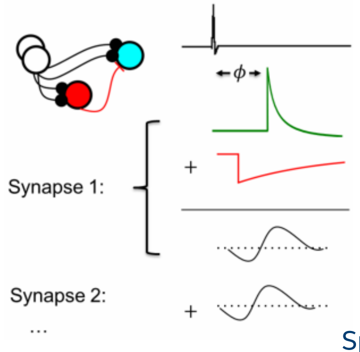
Offer benefits for associative memory capacity, backprop gradient propagation, VSA factoring, among others.

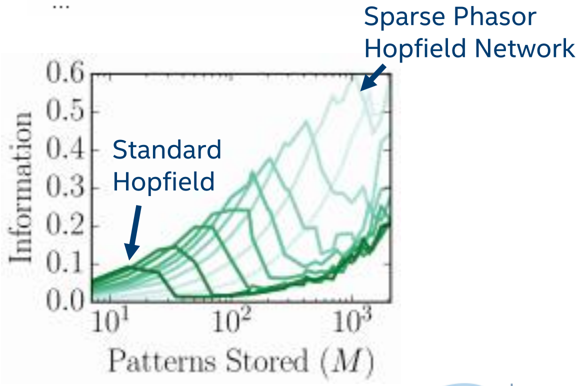
Many SNN implementation benefits:

- Simple LIF implementation w/ different E/I decays
- Constant guaranteed sparse activity
- Synaptic delays provide non-trivial phase transformations
- Fast, bounded response time vs rate coding

Sparse SNN phasor generalization of Hopfield network provides up to **6x higher information per synapse** vs real-valued Hopfield network.

EP Frady, F Sommer, "Robust computation with rhythmic spike patterns." arXiv:1901.07718





The Frontier Ahead

Advancing from Compelling Example Results to Valuable Real-World Technologies

- Inference and learning of sparse feature representations
- Video and speech recognition
- Event-based camera processing
- Chemosensing
- Adaptive dynamic control
- Anomaly detection for security and industrial monitoring
- Optimization: Constraint Satisfaction, QUBO, Convex optimization
- Autonomy: SLAM, Planning, closed-loop behavior

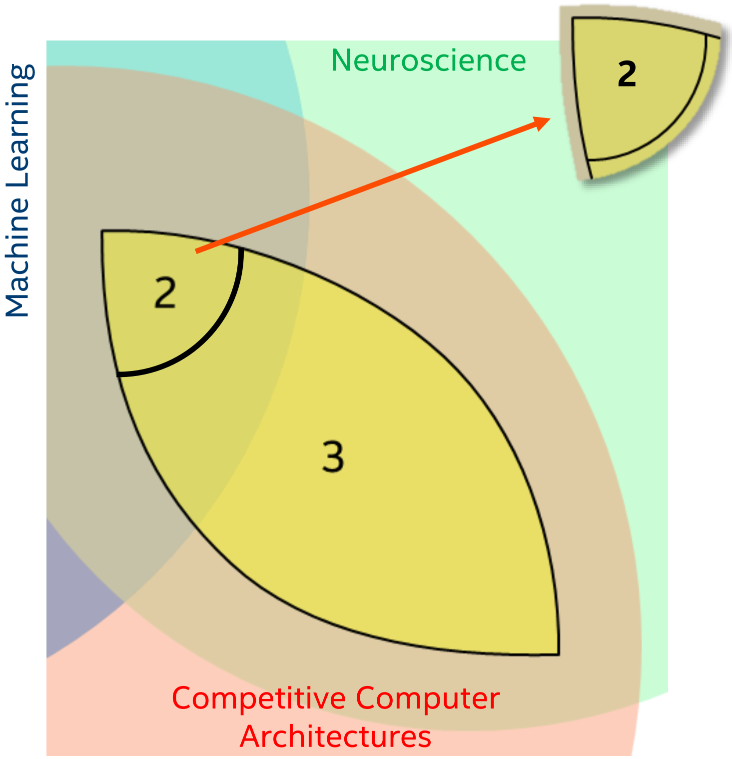
Low Energy	Low Latency	Adaptive	Batch Size = 1	High Cost

Thank You!

Email inrc_interest @ intel.com for more information

LEGAL INFORMATION

Loihi Results Summary: Deep Learning Inspired



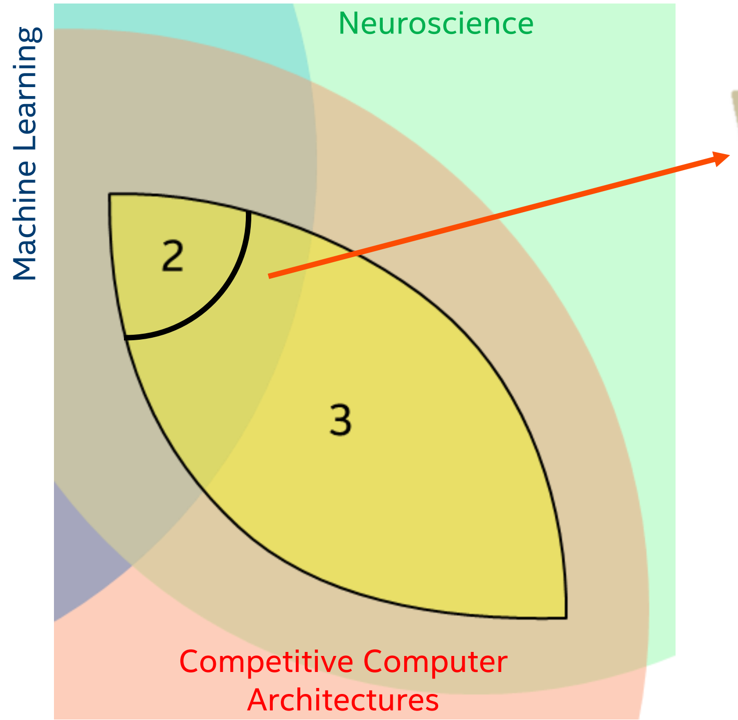
Deep Learning Derived Approaches

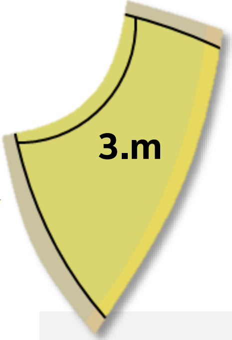
- DNN -> SNN conversion
- SNN backpropagation
- Online SNN pseudo-backprop

Key Results

- For **real-time audio inference**: 1-2 orders of magnitude lower energy than conventional architectures.
  - ☞ *5-10x lower vs Movidius Neural Compute Stick with similar or better performance*
- **SNN backpropagation**: SNNs with long-term adaptive state elements running on Loihi give comparable accuracy to LSTMs.
  - ☞ *94% accuracy on Sequential MNIST vs 96% best LSTM result. Perf/energy benchmarking is WIP – expected to be very good.*
- **Online approximate backpropagation running on Loihi**: Rapid progress on algorithmic formulation & modeling. Neuromorphic architectures will soon support this.
  - ☞ *Suggests a new deep learning approach: offline batch pre-training on CPU/GPU followed by batchsize=1 fine-tuning when deployed on neuromorphic edge processors.*

Loihi Results Summary: Mathematically Formalized





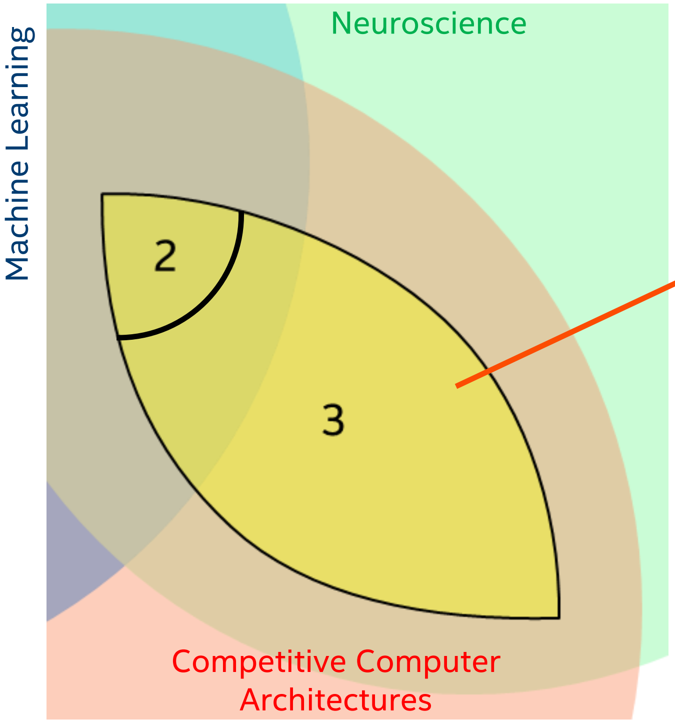
Mathematically Formalized

- Locally Competitive Algorithm for LASSO
- Neural Engineering Framework (NEF)
- Stochastic SNNs for solving CSPs
- Parallel graph search
- Phasor associative memories
- Random diffusion walkers

Key Results

- Loihi solves **LASSO** problems up to **10^4 faster** w/ **10^6 lower energy** than state-of-the-art CPU solvers.
- NEF with local learning rules support **adaptive control algorithms** (e.g. for compliant robotic control) providing **15x lower power** and **3x accuracy improvement** vs similar algorithms on CPU/GPU.
- Loihi can solve **constraint satisfaction problems** with ~100 unknowns in milliseconds.
- Loihi networks can solve million-vertex graph search problems **100x faster** than Djikstra's algorithm on a Xeon.

Loihi Results Summary: Neuro-Inspired Examples



New Ideas Guided by Neuroscience

- Olfaction-inspired rapid learning
- SLAM
- Dynamic Neural Fields
- Evolutionary search
- Cortical models

Key Results

- Loihi supports an olfaction-inspired associative memory autoencoder algorithm with **one-shot** and **online learning**
  - ☞ *Outperforms state-of-art alternatives (PCA, SVM, deep autoencoder) on chemosensor datasets*
- Neuro-inspired one-dimensional **SLAM** algorithm runs on Loihi with **100x lower power** vs GMapping on a CPU.
  - ☞ *Enhancement to 2D is in progress*