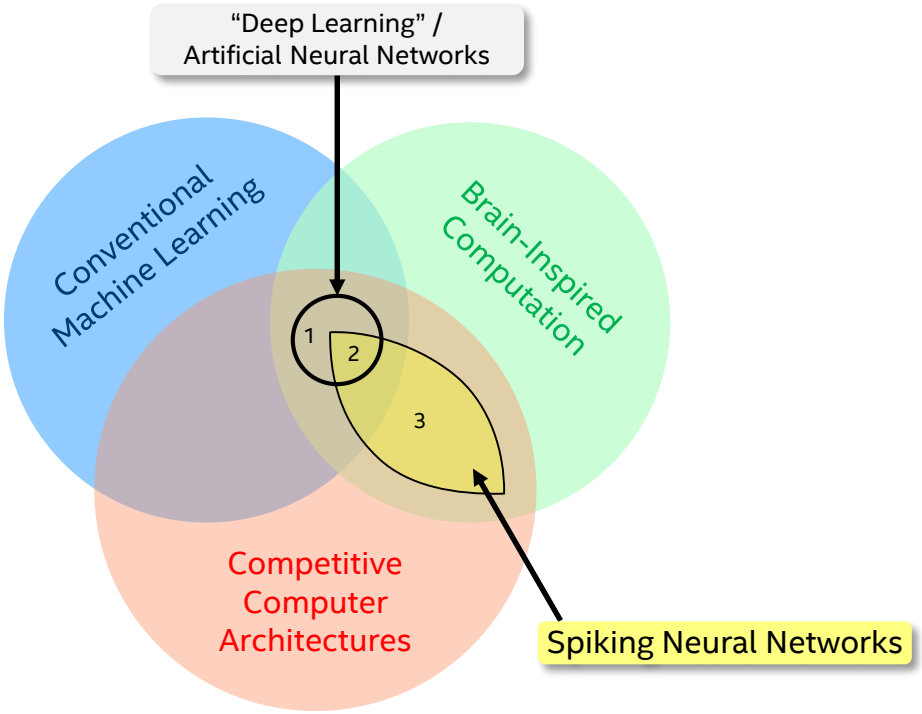
REALIZING THE PROMISE OF SPIKING NEUROMORPHIC HARDWARE

Mike Davies
Director, Neuromorphic Computing Lab | Intel Labs

Second International Workshop on Event-based Vision and Smart Cameras
June 16, 2019

LEGAL INFORMATION

Neuromorphic Computing Exploration Space



Research Goals:

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

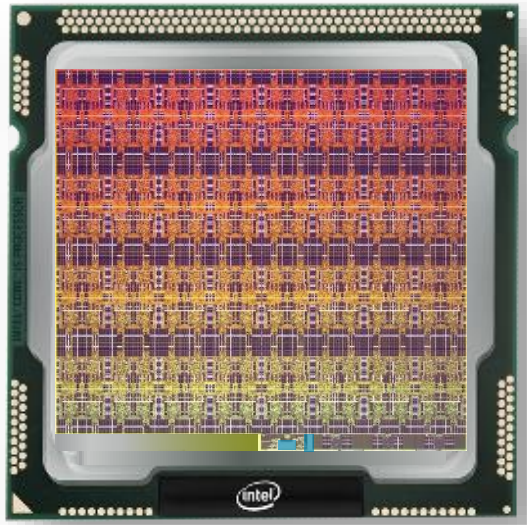
Examples:

- Learning without cloud assistance
- Learning with sparse supervision
- Online and lifelong learning
- Probabilistic inference and learning
- Sparse coding
- Associative memory, similarity matching
- Nonlinear adaptive control (robotics)
- SLAM and path planning
- Constraint satisfaction
- Dynamical systems modeling

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

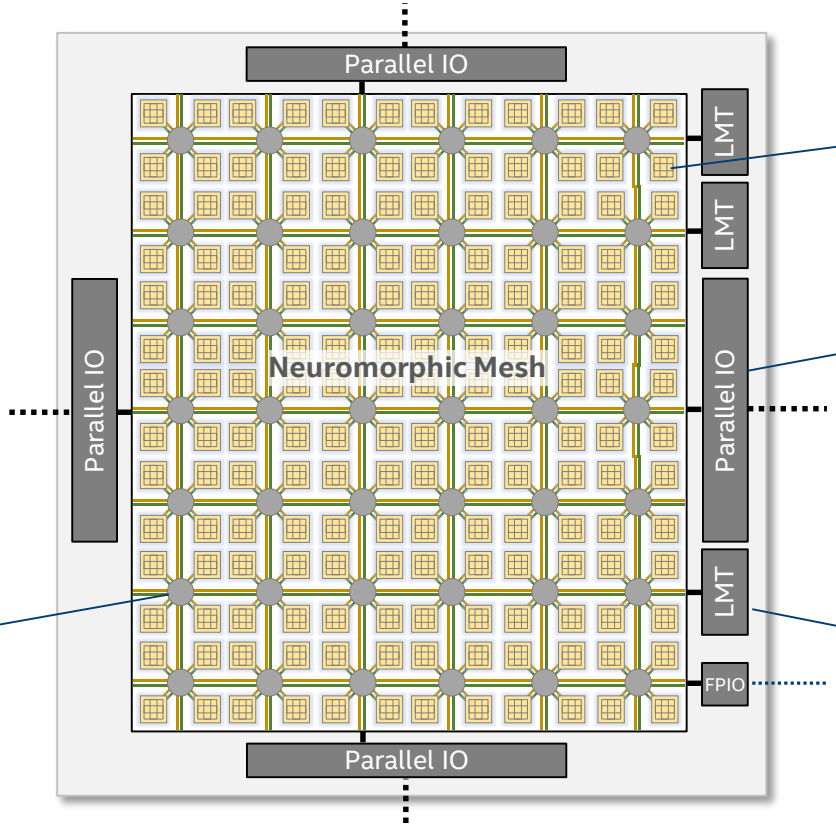


Integrated Memory + Compute Neuromorphic Architecture

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

Chip Architecture

Technology:	14nm
Die Area:	60 mm^2
Core area:	0.41 mm^2
NmC cores:	128 cores
x86 cores:	3 LMT cores
Max # neurons:	128K neurons
Max # synapses:	128M synapses
Transistors:	2.07 billion



Neuromorphic core
- LIF neuron model
- Programmable learning
- 128 KB synaptic memory
- Up to 1,024 neurons
- Asynchronous design

Parallel off-chip interfaces
- Two-phase asynchronous
- Single-ended signaling
- 100-200 MB/s BW

Embedded x86 processors
- Efficient spike-based communication with neuromorphic cores
- Data encoding/decoding
- Network configuration
- Synchronous design

Low-overhead NoC fabric
- 8x16-core 2D mesh
- Scalable to 1000's cores
- Dimension order routed
- Two physical fabrics
- 8 GB/s per hop

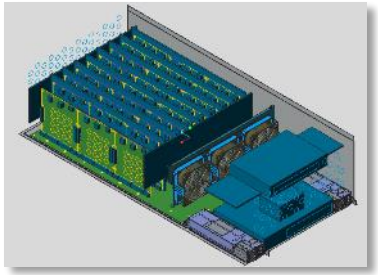
Loihi Systems

Q4 2017
Wolf Mountain
Remote Access
4 Loihi/Board

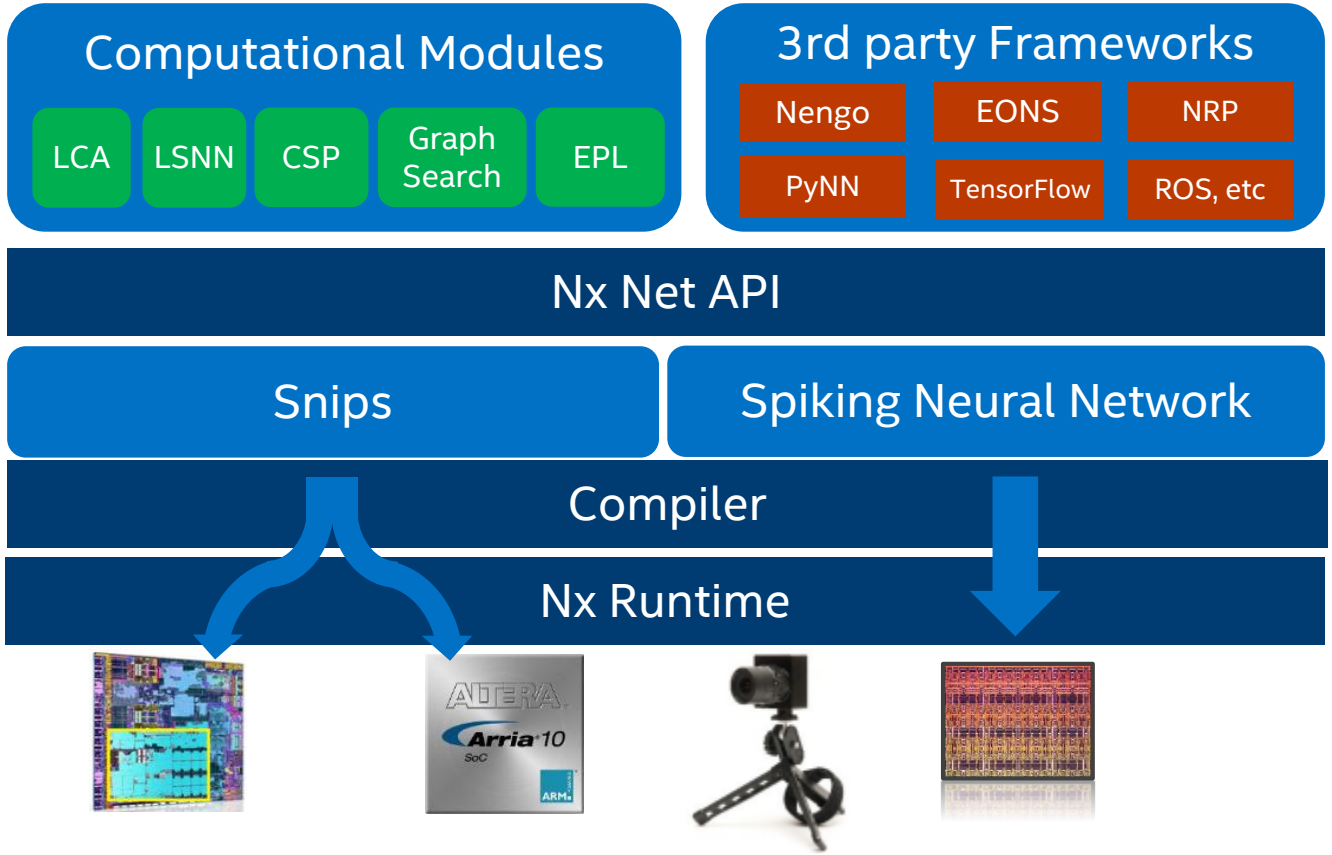
Q2 2018
Nahuku
Arria10 Expansion Board
For cloud & local use
8-32 Loihi/Board

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

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

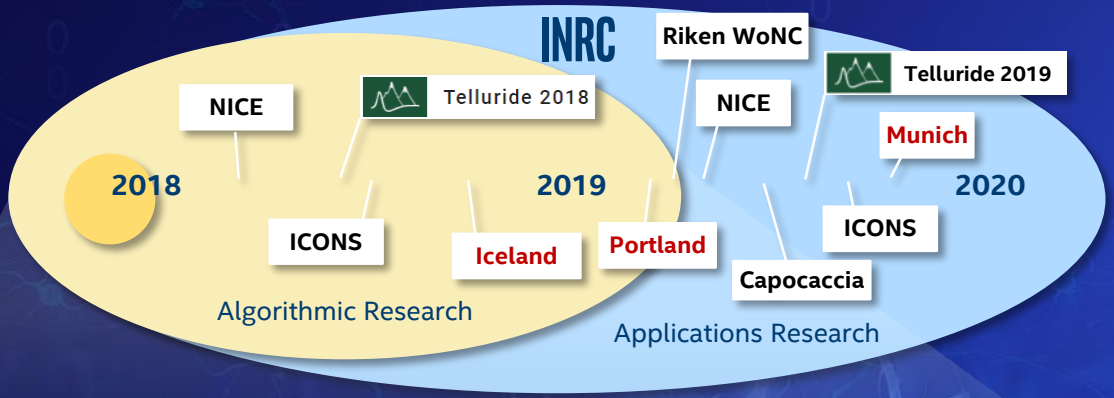


Nx SDK Software Architecture



INTEL NEUROMORPHIC RESEARCH COMMUNITY

Collaborating to Accelerate Progress

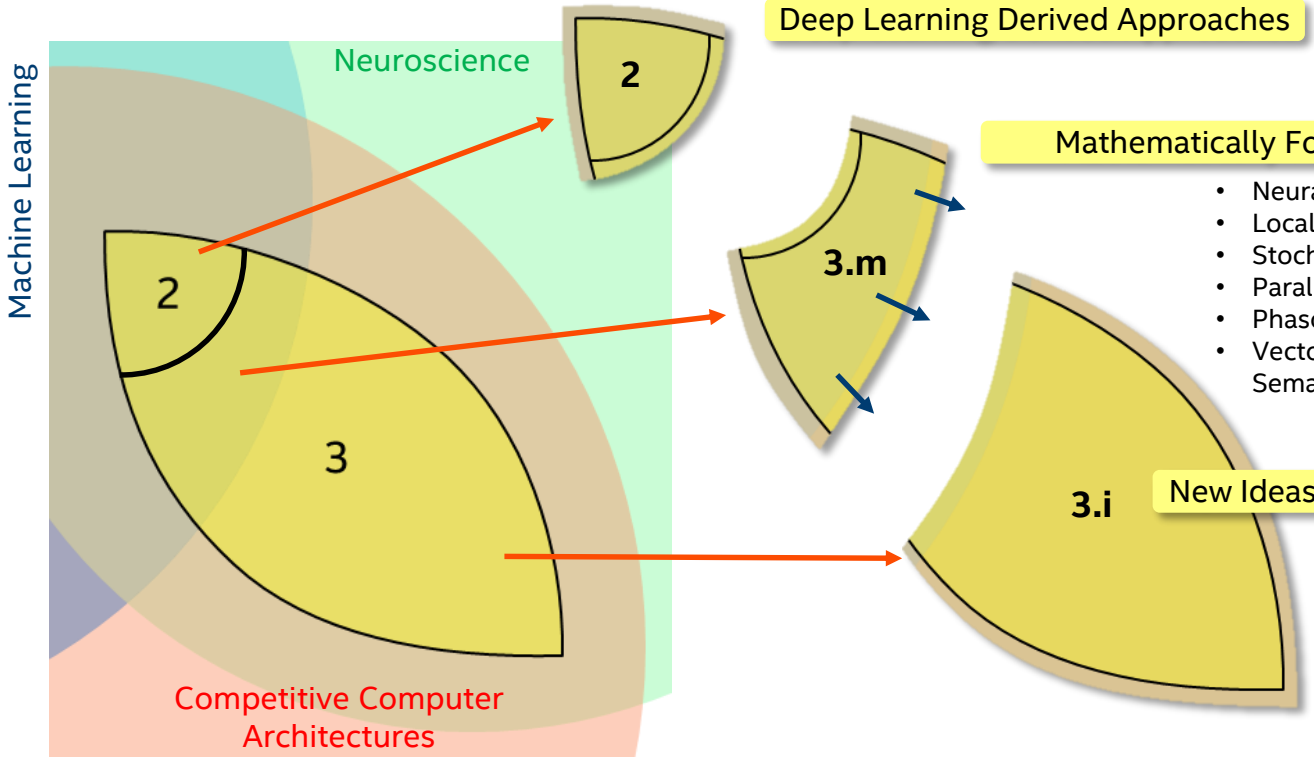


Over 50 active projects

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

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

The Challenge: SNN Algorithm Discovery



Deep Learning Derived Approaches

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

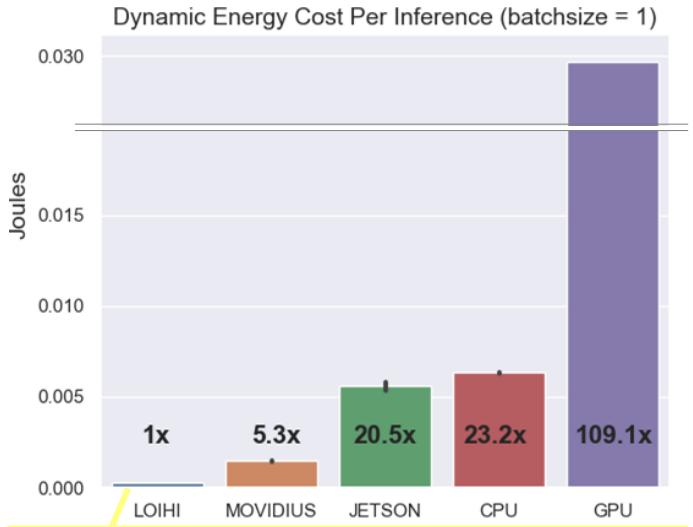
Mathematically Formalized

- Neural Engineering Framework (NEF)
- Locally Competitive Algorithm for LASSO
- Stochastic SNNs for solving CSPs
- Parallel graph search
- Phasor associative memories
- Vector symbolic architecture (VSA) *aka* Semantic pointer architecture (SPA)

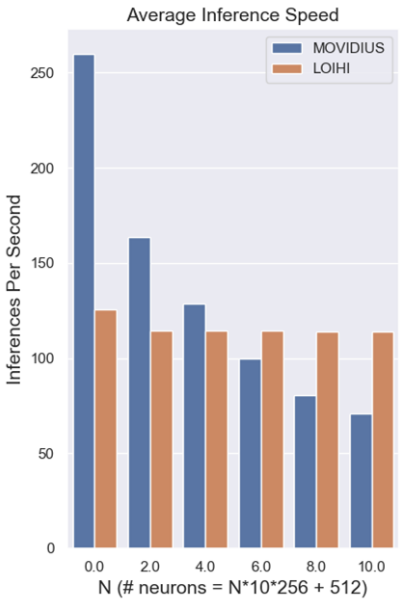
New Ideas Guided by Neuroscience

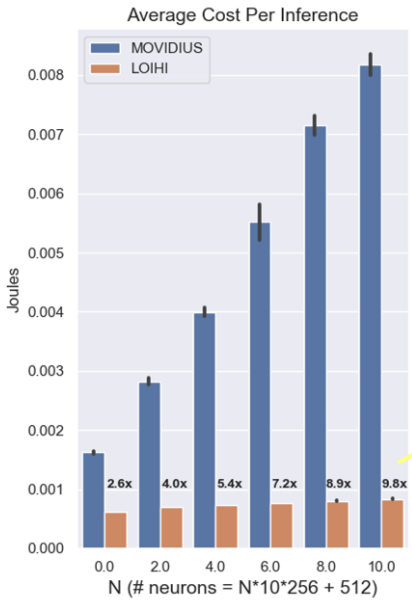
- Olfaction-inspired rapid learning
- Dynamic Neural Fields
- SLAM
- Evolutionary search
- Cortical models

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

LASSO Sparse Coding

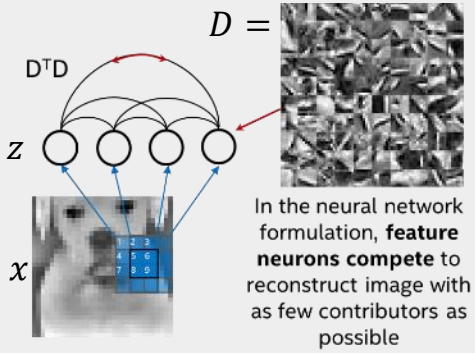
The *Spiking Locally Competitive Algorithm (S-LCA)*

Problem

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

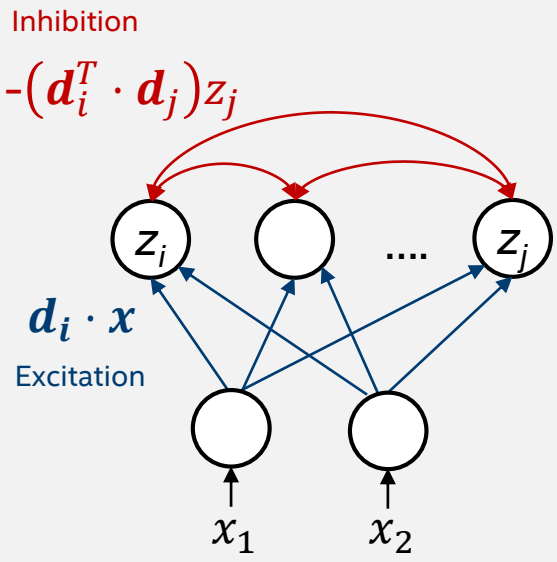
Input — Reconstruction — Sparse regularization

Implementation



Tang et al, arxiv: 1705:05475

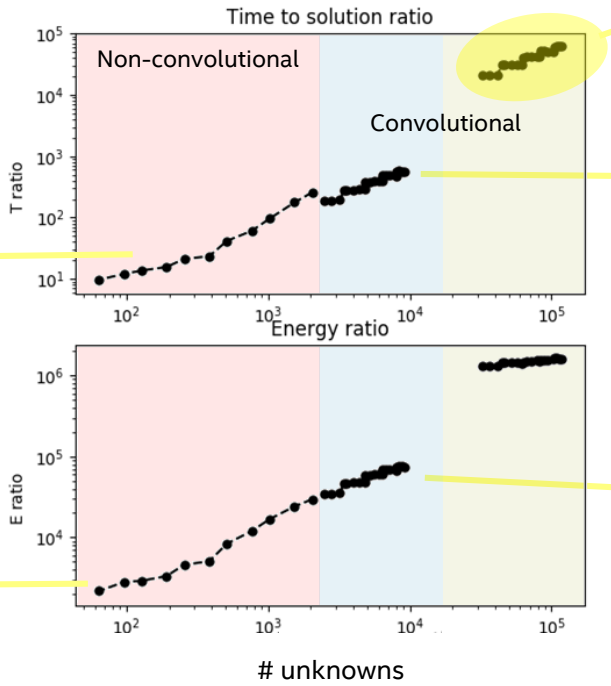
Neural Network Structure



LCA on Loihi compared to FISTA on Core i7 CPU

Clear, compelling scaling trend across both non-convolutional and convolutional examples.

CPU/Loihi Ratios



>10,000x faster

(Possibly unfair to CPU since SPAMS is not optimized for convolutional LASSO.)

100-1000x faster

10-50x faster

10,000-100,000x lower energy

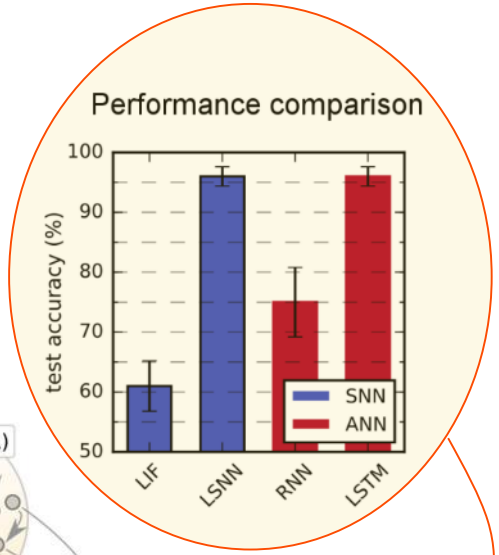
1,000-10,000x lower energy

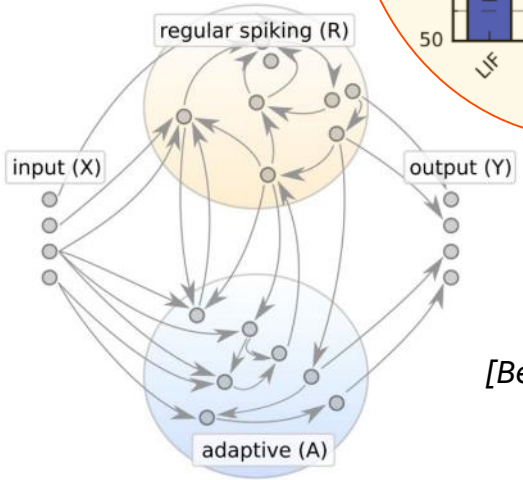
* **Intel Core i7-4790 3.6GHz w/ 32GB RAM.** FISTA solver: **SPAMS** http://spams-devel.gforge.inria.fr/
Performance results are based on testing as of December 2018 and may not reflect all publicly available security updates. No product can be absolutely secure.

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)
- For Sequential MNIST dataset:
  - **Loihi achieves 94% accuracy**
  - LSTM: **98%** (simple RNNs: **68-89%**)





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

[Bellec et al, NeurIPS 2018]

LSNN Benchmarking Results

Algorithm	Dataset	Training	#Param	Best Accuracy
LSNN	Sequential MNIST	SNN Backprop + DEEP-R w/ TensorFlow (Adam Optimizer)	68210	94.1%
LSTM	Sequential MNIST	Standard Backprop w/ TensorFlow (Adam Optimizer)	67850	98.5%

Loihi is **best on all metrics**, *including throughput* (with batch size = 1)

Architecture	Batch size	Energy per inference (mJ)		Latency per inference (ms)		Inference Throughput (1/s)	
Loihi[1]	1	2.68	1x	21.5	1x	47	1x
Intel Core i5-7440HQ[2]	1	1740	649x	83.2	3.9x	12	1/4x
Intel Core i7-7700HQ[2]	1	2510	937x	77.7	3.6x	13	1/3.6x
NVIDIA GeForce GTX 1050 Ti[3]	1	n/a	n/a	66.8	3.1x	15	1/3.1x
NVIDIA Tesla P100[4]	1	3480	1298x	94.9	4.4x	11	1/4.3x
NVIDIA Tesla P100[4]	64	171	64x	148	6.9x	435	9.3x

Best GPU is **worst** on all metrics except **throughput w/ large batch size**

[1] Loihi Wolf Mountain running NxSDK 0.85
[2] 2.8-3.8 GHz CPU with 16 GB RAM. TensorFlow v1.14.1 on Windows 10.
[3] 4GB RAM, CUDA v10.0. Driver v419.17. TensorFlow v1.13.1
[4] 16GB RAM, CUDA v10.0. Driver v410.104. TensorFlow v1.10.1
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.

Graph Search – Path Planning

Runtime comparison to best Dijkstra optimizations:

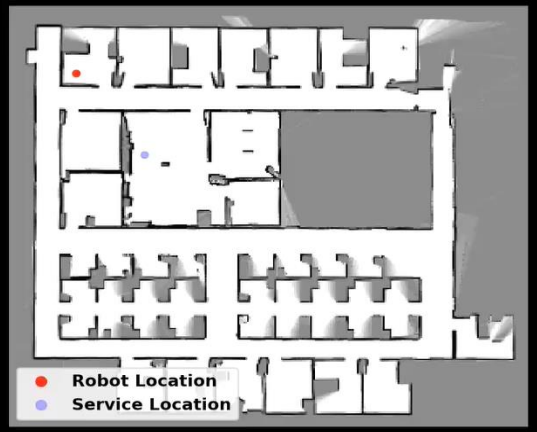
- Neuromorphic: $O(L \cdot \sqrt{V})$
- Standard: $O(E)$

For most nontrivial problems:

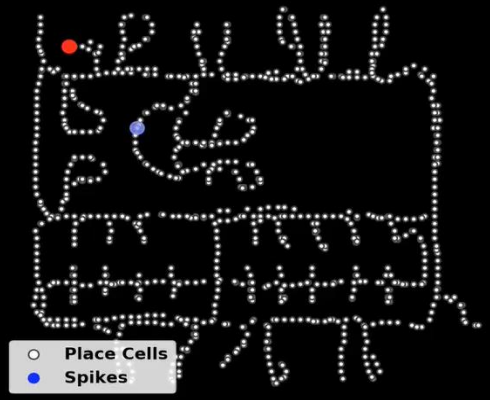
- L<<E
- V<<E

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

Robot Motion



Loihi Representation

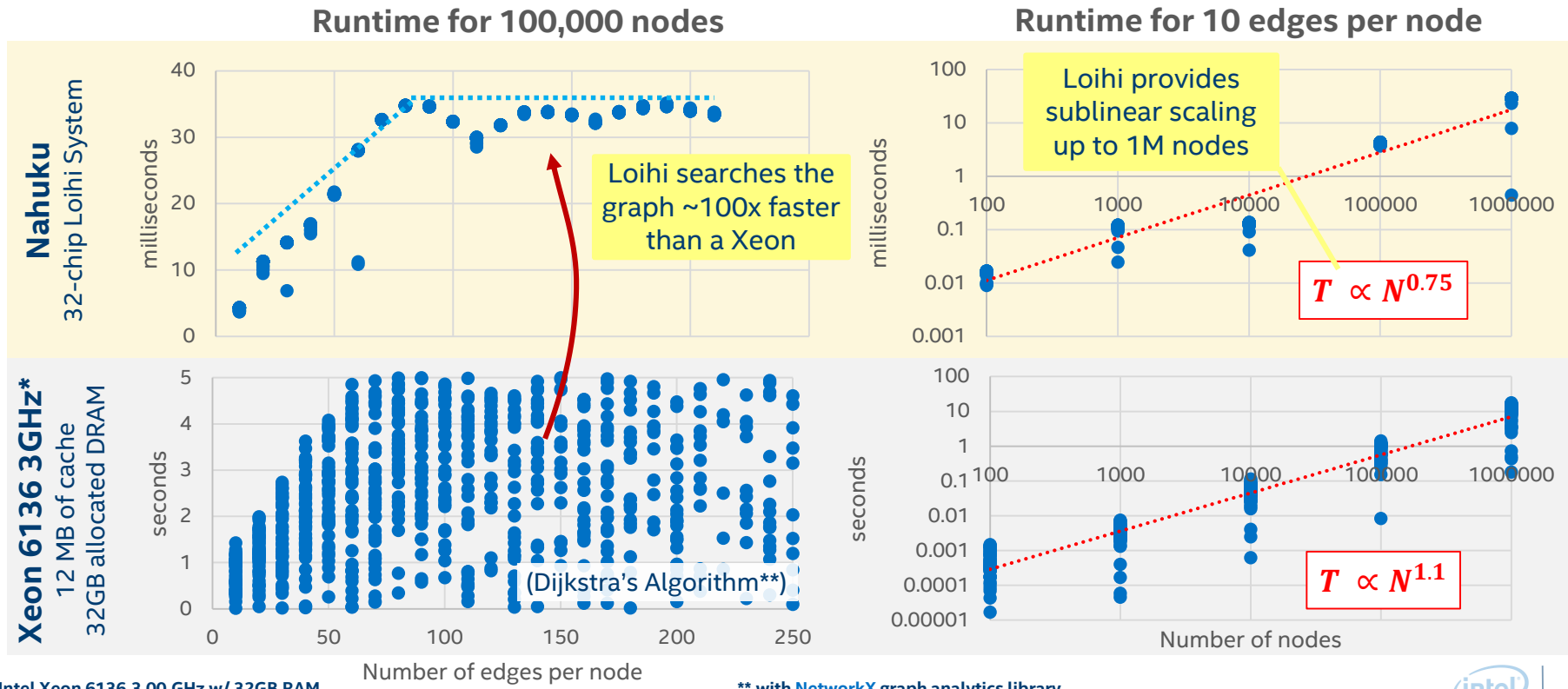


DARPA SDR Site B
(Data from Radish Robotics Dataset)

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.*

Searching Small World Networks with Loihi

Watts-Strogatz network model with rewiring probability 20%.



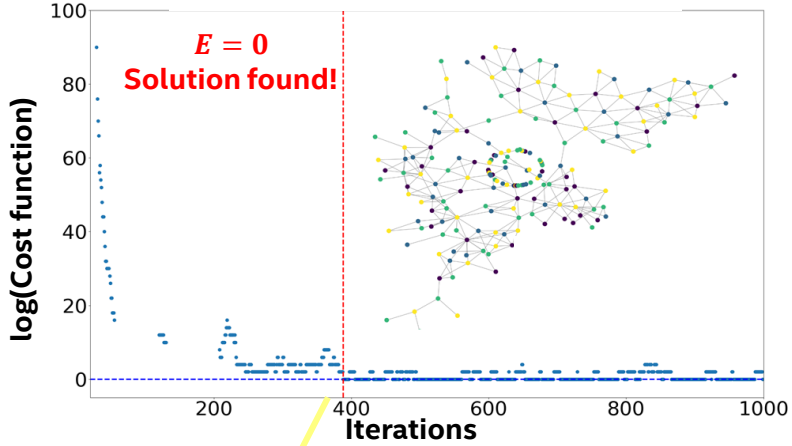
* Intel Xeon 6136 3.00 GHz w/ 32GB RAM.

** 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.

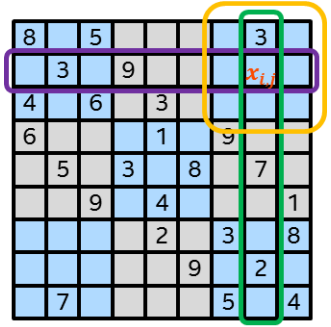
Solving Constraint Satisfaction Problems

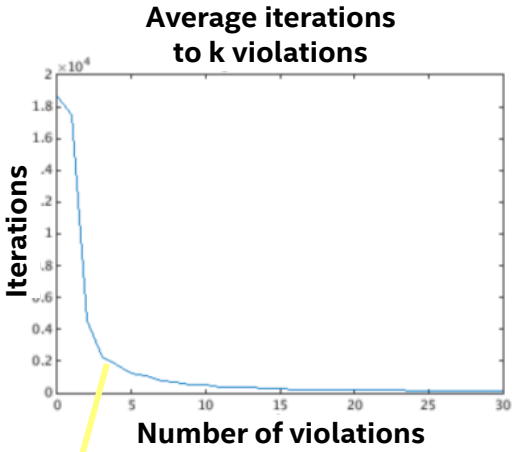
Example: 4-coloring of world map



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

Example: Sudoku

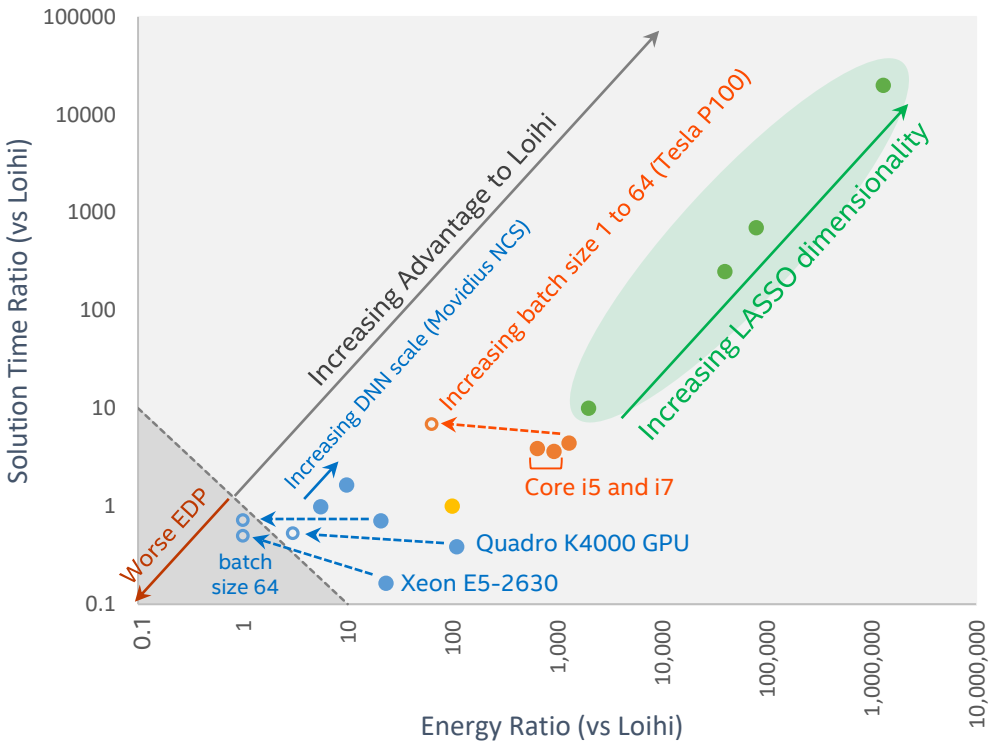




Finds solutions with few violations quickly & improves solution over time.

Uses stochastic SNN MCMC algorithm from Jonke et at., Front. Neurosci. 10.118, 2016
See also G Guerra and S Furber, Front. Neurosci. 11.714, 2017

Loihi Benchmarking Summary



- Keyword Spotter DNN*
- Keyword Spotter DNN* (batch size >1)
- 1D SLAM**
- Sequential MNIST (LSNN***)
- Sequential MNIST (batch size 64)
- LASSO
- Unit energy delay product (EDP)

** P Blouw et al, 2018.* arXiv:1812.01739
*** G Tang et al, 2019. arXiv:1903.02504*
**** Bellec et al, 2018. arXiv:1803.09574*

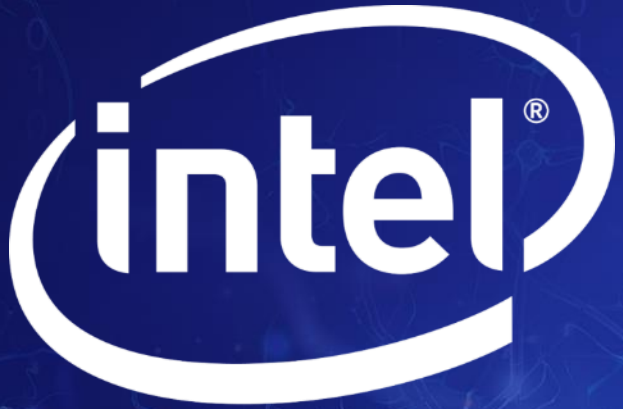
The Research Frontier

Advancing from Compelling Algorithms to Viable Applications

- Inference and learning of sparse feature representations
- Video and speech recognition
- Event-based camera processing
- Chemosensing
- Robotics
- 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