

INRC Spring 2022 Workshop

New Tools for a New Era of Neuromorphic Computing

Mike Davies
Director, Neuromorphic Computing Lab

intel
labs

April 19, 2022

Intel Neuromorphic Research Community Spring 2022 Workshop

Legal Information

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Results have been estimated or simulated.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

Welcome!

To Our Second Fully Virtual INRC Workshop

Fall Workshop in Graz, Oct 2019	Winter Workshop Feb 2021	Spring Workshop April 2022
Face-to-face	Webex	Teams
100 Attendees	350+	700+
Barely any recorded content	Nearly all content recorded	Nearly all content recorded
Most sessions closed to formally engaged members	Most sessions open to the broader community	Nearly all sessions open to the broader community
PDF agenda	Event website	Event website
Hallway chats, lunches, dinners	Slack	Slack
Acoustics problems	Connection problems	We will see...

(Next one hybrid – fingers crossed)

Workshop Goals

Long-time INRC Members	New Neuro Researchers	New Industry Participants
~40%	~40%	~20%
Learn about our new tools (Lava, algorithmic libraries, and Loihi 2) and how to start developing		
Learn and share new results, ideas, and developments with Loihi and the broader community		
Make new connections for collaboration		
Get up to speed on background (“first era” of Loihi research)		Share needs and identify compelling business opportunities

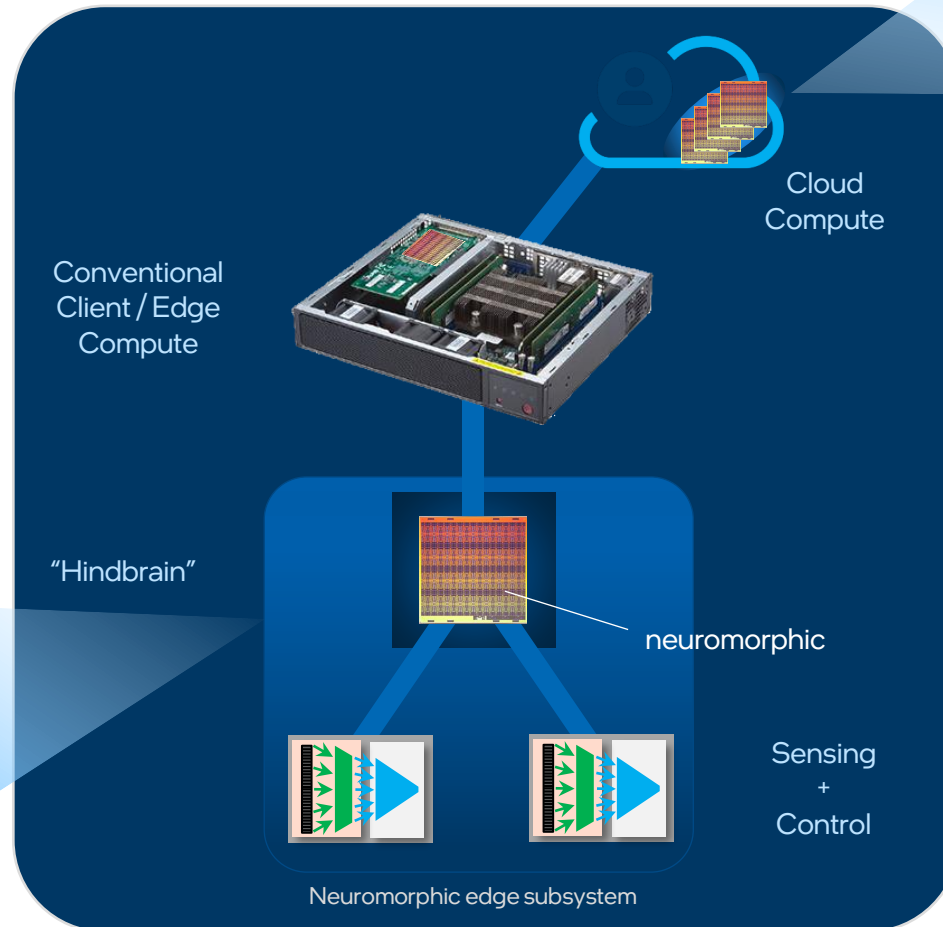
Time (PDT / CET)	Session	Speakers
8:00 / 17:00	Welcome: Workshop kick-off	
8:05 / 17:05	New Tools for a New Era of Neuromorphic Computing	Mike Davies
8:40 / 17:40	Loihi 2	
8:55 / 17:55	Lava	Mike Davies, Andreas Wild
9:40 / 18:40	Workshop and Community Orientation	Tim Shea + WG leads
10:00 / 19:00	Q&A / Break	
10:30 / 19:30	Featured Community Results	
10:35 / 19:35	<ul style="list-style-type: none"> ▪ Neuromorphic Tunneling? Comparing Loihi with Quantum Annealing 	Garrett Kenyon, LANL
10:52 / 19:52	<ul style="list-style-type: none"> ▪ Monte Carlo Simulations on Loihi 	Brad Aimone, Sandia
11:09 / 20:09	<ul style="list-style-type: none"> ▪ Loihi in Orbit: the First 90 Days 	Michael Lowry, NASA
11:26 / 20:26	<ul style="list-style-type: none"> ▪ The Backpropagation Algorithm Implemented on Loihi 	Alpha Renner, LANL
11:43 / 20:43	<ul style="list-style-type: none"> ▪ VSA With Phasor Neurons on Loihi - Towards Neuromorphic Visual Odometry 	Alpha Renner, ETHz/INI
17:00 / 02:00	Loihi 2 Deep Dive	Garrick Orchard

Our Goal

Develop a new programmable computing technology inspired by the modern understanding of brain computation



Integrate neuromorphic intelligence into computing products at all scales



Achieve brain-like efficiency, speed, adaptability, and intelligence

What is Neuromorphic Computing? A Huge Bottom-up Exploration Space

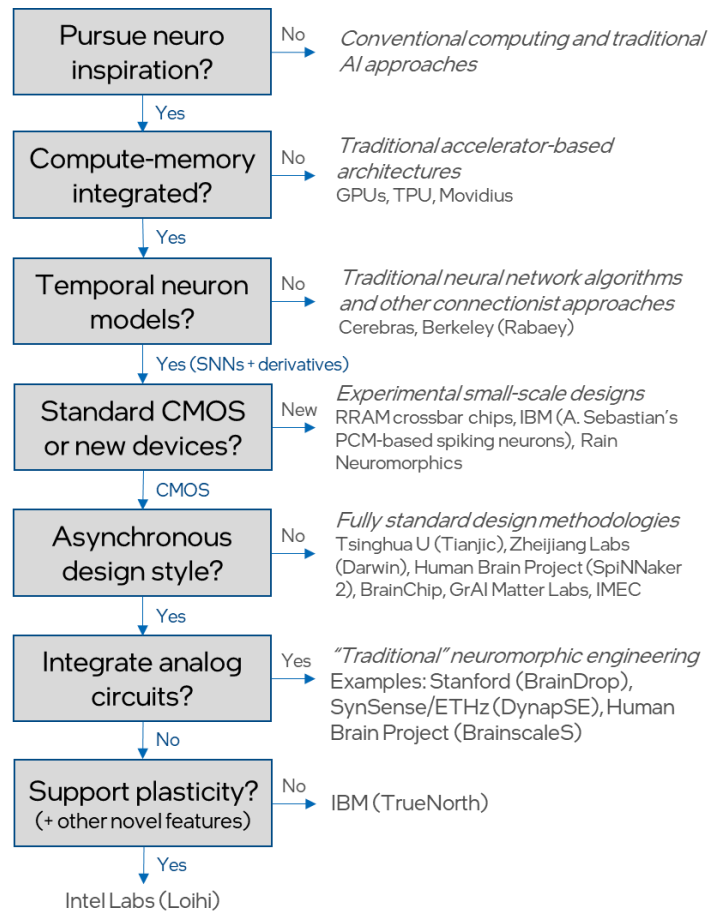
- Self-organized growth
- Autonomous healing
- Exploiting material time constants
- Oscillatory dynamics
- Stochasticity
- Local learning rules
- Very high fanout
- Distributed data representations
- Fine-grain parallelism
- Temporal data coding
- Sparse temporal activity ("Spikes")
- Sparse connectivity
- 3D wiring
- Recurrence and feedback loops
- Compute-memory integration
- Analog-valued persistent state
- Online causal adaptation
- Low precision
- Dynamics on diverse time scales
- Hybrid analog/digital computation
- Continuous time operation
- Parametric Heterogeneity



Increasingly exotic or uncommon properties in conventional computing systems

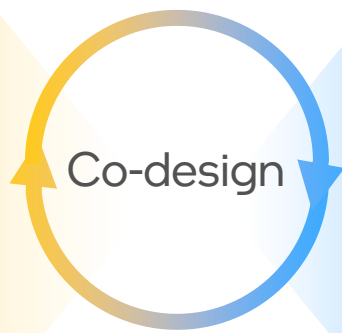
Our Approach: Iterative architecture-algorithms co-design

Neuro-Inspired Silicon



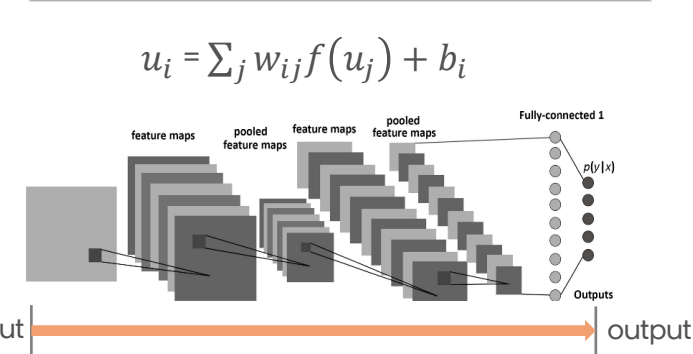
Novel Neuro-Inspired Algorithms

Category	Example applications
Deep learning: backprop-trained event-based DNNs	Object and gesture recognition for event-based vision sensors, slip detection for event-based tactile sensors, ANNs with sparsely changing input data
Deep learning: DNNs with online adaptation	Few-shot new gesture learning, Adaptive control,
Vector Symbolic Architectures (VSA), aka Hyperdimensional Computing (HDC)	Semantic factorization, relational reasoning, symbolic and analogical reasoning
Neural Engineering Framework (NEF)	Adaptive control systems, state machines
Dynamic Neural Fields (DNF)	SLAM, object tracking, dynamic control, attention
Neural sampling e.g. spiking Boltzmann machines	Constraint satisfaction, probabilistic inference
Oscillatory computation	Optimization, event-based spectral transforms, optic flow, audio spectral normalization
Recurrent Excitation/Inhibition-balanced networks	LASSO regression, sparse feature coding
Event-based networks with temporally coded information	Graph search, similarity search

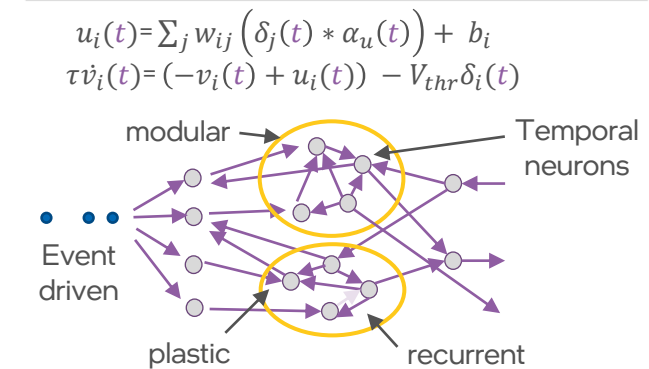


Rigorous Benchmarking

Conventional Deep Networks

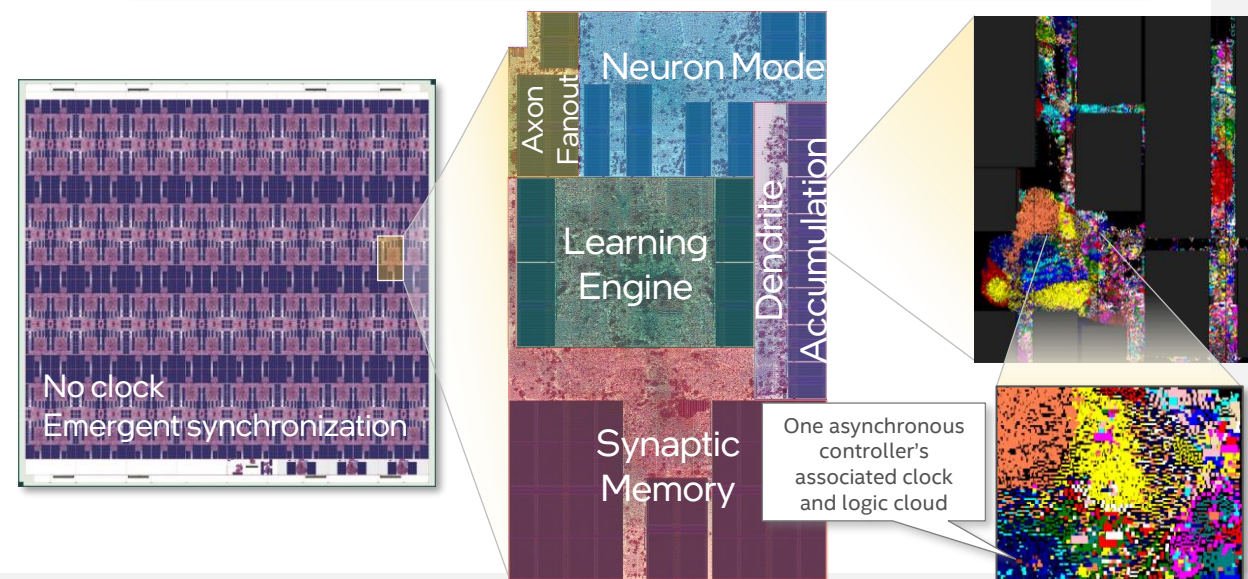
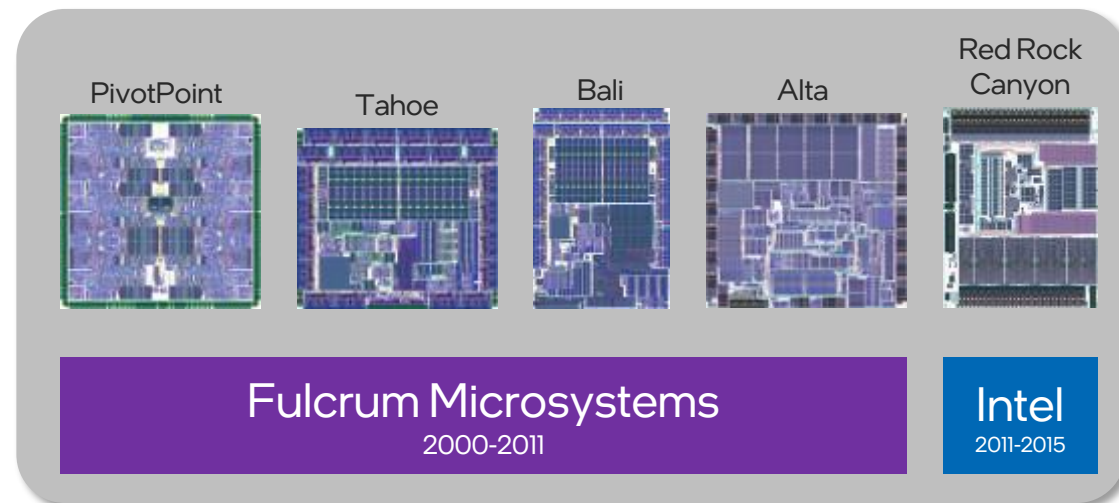


Neuromorphic Networks

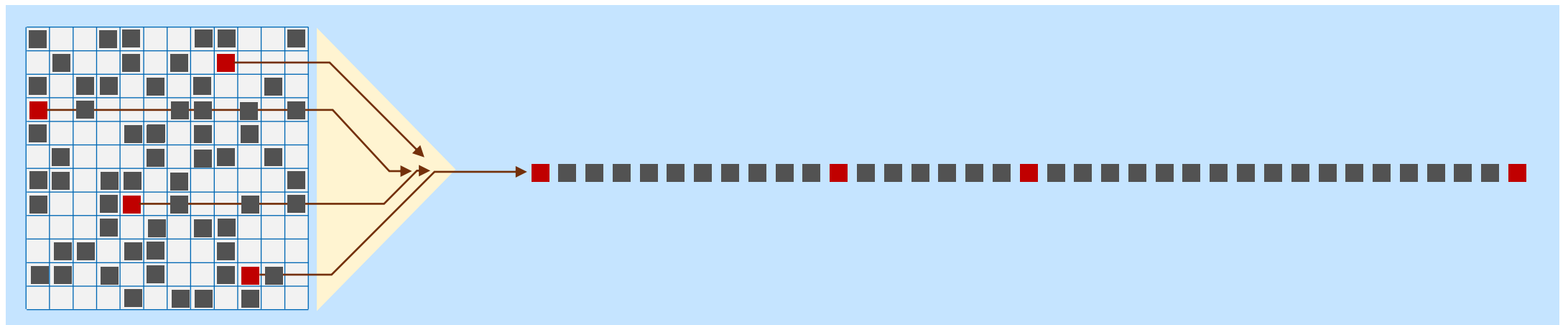
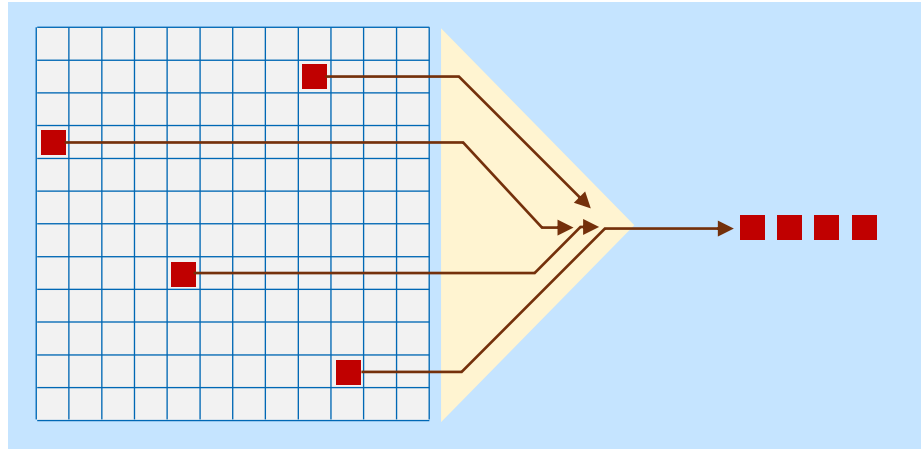


Chip implementation with asynchronous design

- Commercially developed over 2000-2015 at Fulcrum Microsystems and Intel
- Applied to five generation of commercial Ethernet switches
- Power consumption scales with activity – matched for spiking neurons and sparse interconnect
- Low latency communication – enables scaling to large systems
- Allows neuromorphic mesh to operate over 1000x faster than biological speeds using emergent synchronization
- Supports low energy, highly-ported SRAM arrays



Sparse, asynchronous communication is fast



Leads us to a new class of computer architecture

Standard Computing



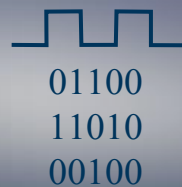
**PROGRAMMING BY
ENCODING ALGORITHMS**

**SYNCHRONOUS
CLOCKING**

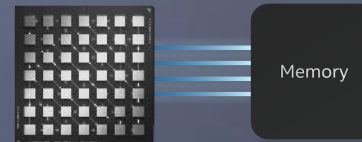
**SEQUENTIAL THREADS
OF CONTROL**

```
if X then
...
else
...

```



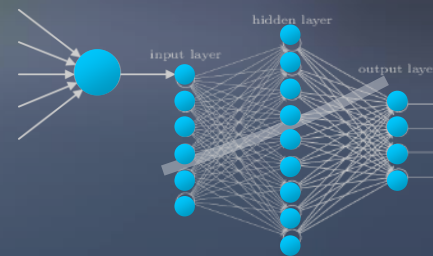
Parallel Computing



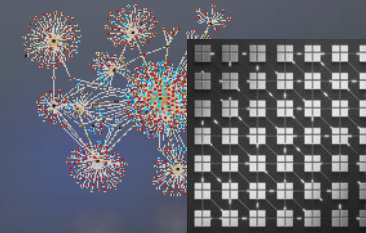
**OFFLINE TRAINING USING
LABELED DATASETS**

**SYNCHRONOUS
CLOCKING**

**PARALLEL
DENSE COMPUTE**



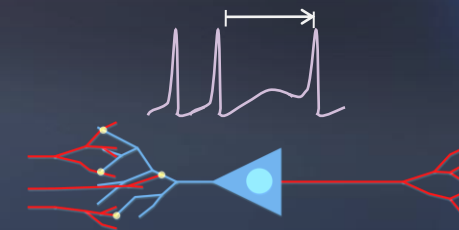
Neuromorphic Computing



**LEARN ON THE FLY THROUGH
NEURON FIRING RULES**

**ASYNCHRONOUS
EVENT-BASED SPIKES**

**PARALLEL
SPARSE COMPUTE**



Realized in Loihi

KEY PROPERTIES

Compute and memory integrated
to spatially embody programmed networks

Temporal neuron models (LIF)
to exploit temporal correlation

Spike-based communication
to exploit temporal sparsity

Sparse connectivity
for efficient dataflow and scalability

On-chip learning
without weight movement or data storage

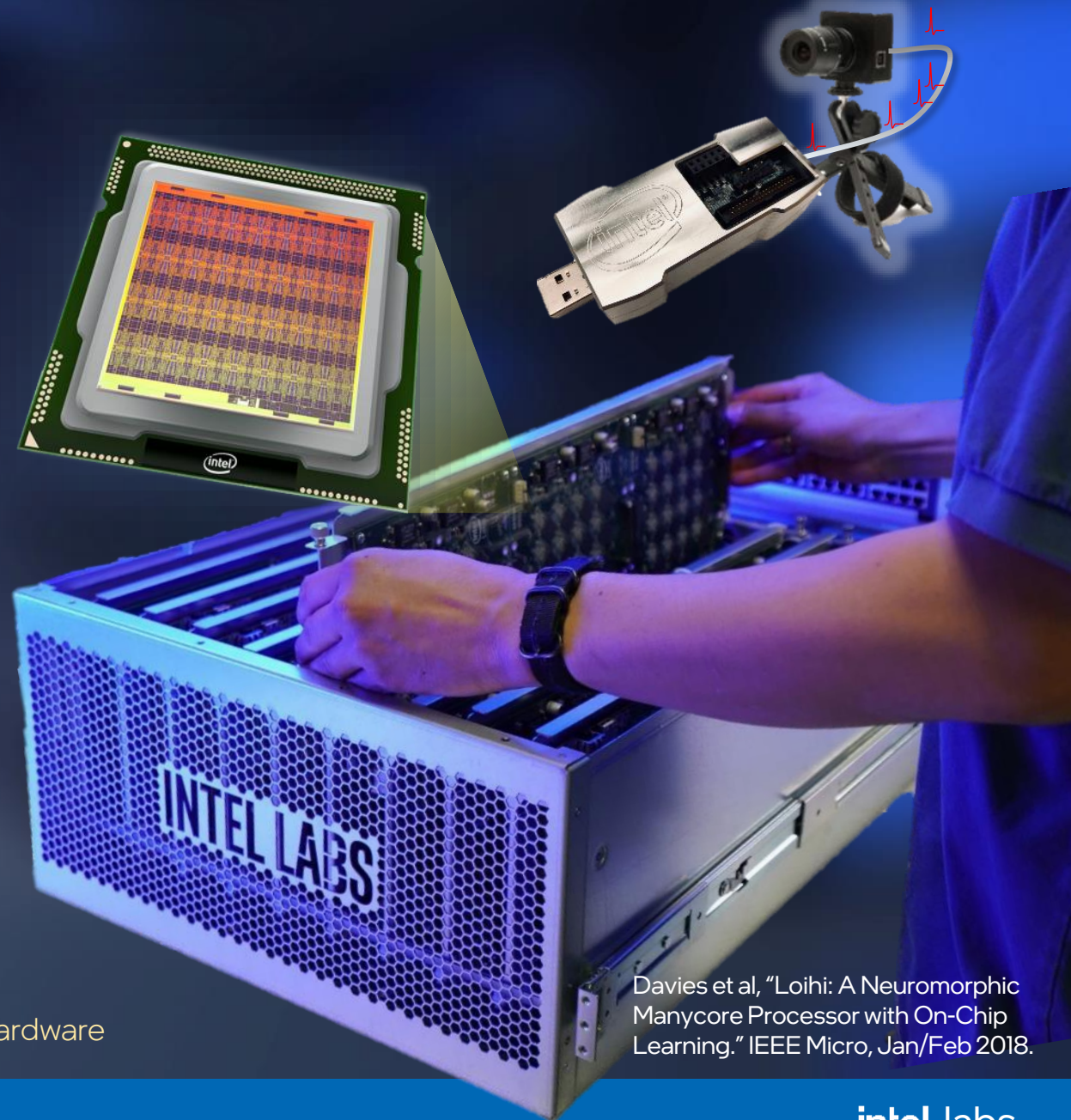
Digital asynchronous implementation
for power efficiency, scalability, and fast prototyping

Yet...

No floating-point numbers
No multiply-accumulators
No off-chip DRAM

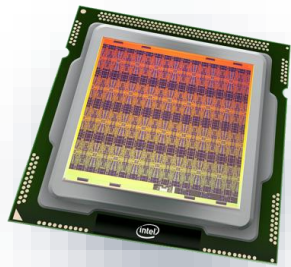
Fundamental to
deep learning hardware

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



Significant progress over three years of Loihi research

Loihi
NxSDK
INRC



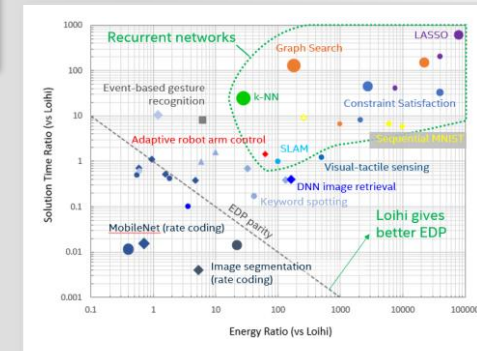
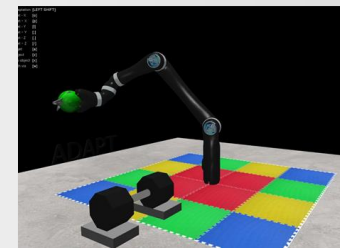
Asynchronous
Design



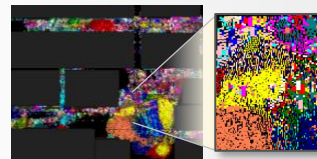
Intel Neuromorphic Research Community

Collaborating to Accelerate the Research

INRC includes over 150 groups



CIO JOURNAL
Intel to Release Neuromorphic-Computing System
Pohoiki Springs, an experimental system to be rolled out this month, mimics the way human brains work to do computations faster with less energy
THE WALL STREET JOURNAL

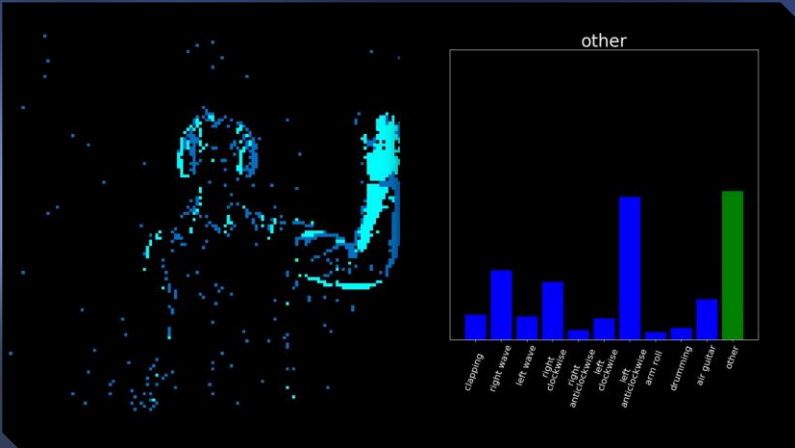


nature machine intelligence
Volume 2 Issue 3, March 2020
Neuromorphic olfaction

Neuro-inspired algorithm for odor recognition and learning demonstrated on Loihi, able to learn 3000x more data efficiently compared to DNN solution.

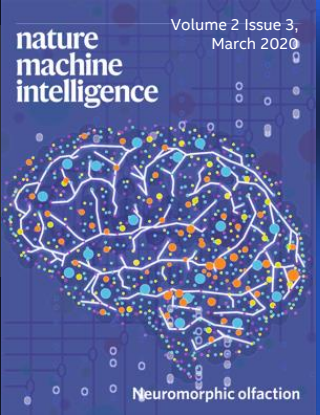


Loihi Has Confirmed the Value of This Direction

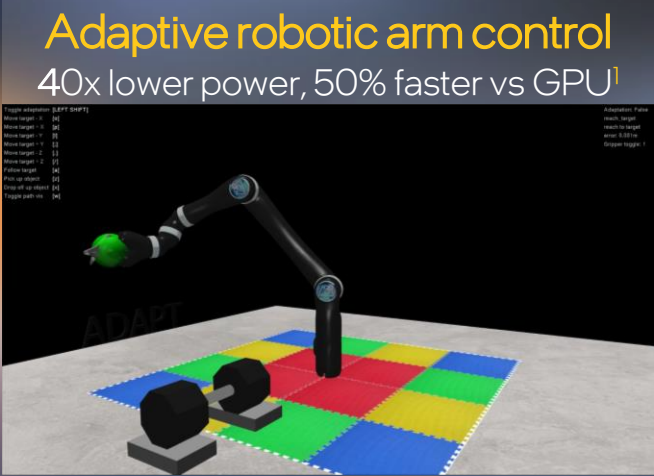
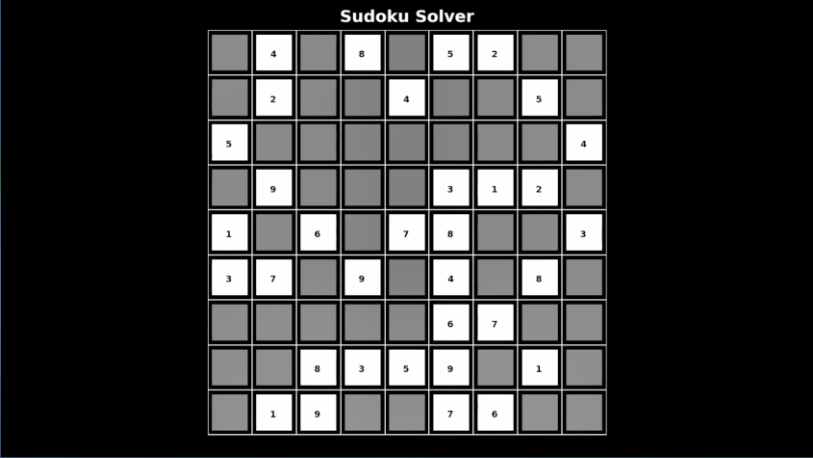


Gesture recognition + learning
 Loihi + DAVIS 240C camera
 60 mW total power, 15 mW dynamic¹

Olfaction-inspired odor recognition and learning
 3000x more data efficient learning than a deep autoencoder



Combinatorial optimization (CSP, SAT, ILP, QP)
 2,800x lower energy and 44x faster vs CPU¹

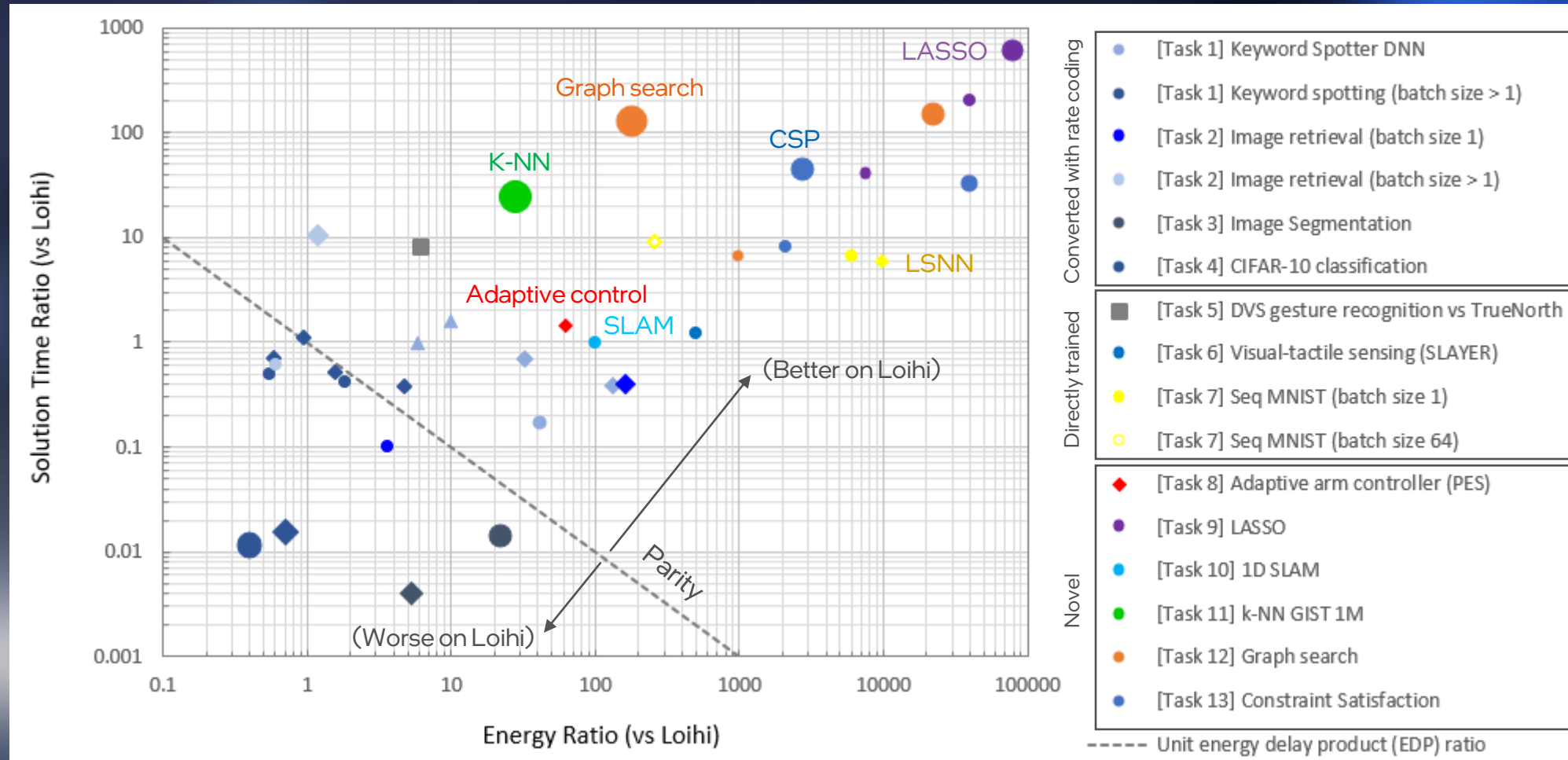


¹ M. Davies et al, "Advancing Neuromorphic Computing With Loihi: A Survey of Results and Outlook," Proc. IEEE, 2021. Results may vary.

For the right workloads, orders of magnitude gains in latency and energy efficiency are achievable

Reference architecture

- CPU (Intel Core/Xeon)
- ◆ GPU (Nvidia)
- ▲ Movidius (NCS)
- TrueNorth

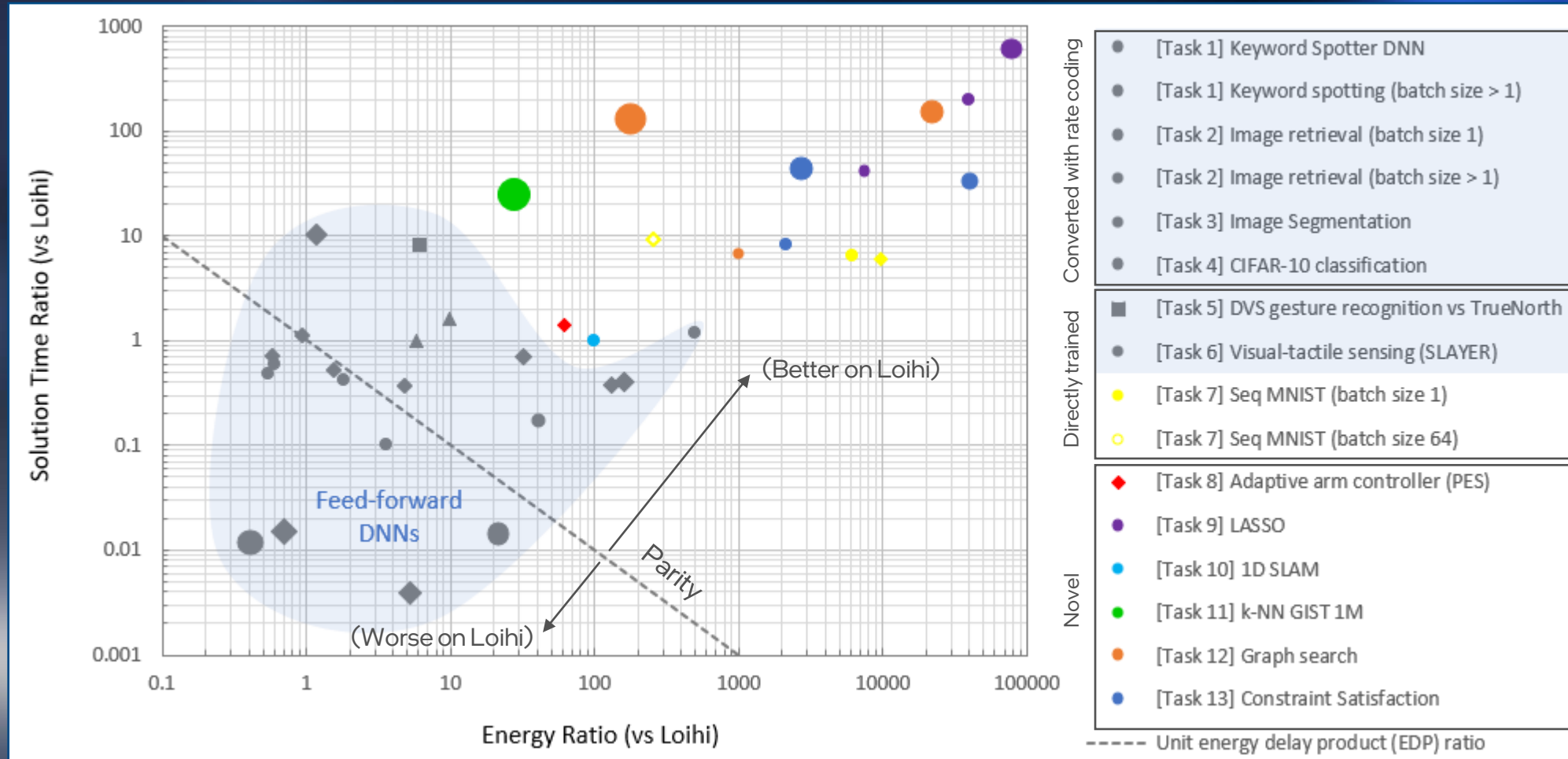


M. Davies et al, "Advancing Neuromorphic Computing With Loihi: A Survey of Results and Outlook," Proc. IEEE, 2021. Results may vary.

Standard feed-forward deep neural networks give the **least** compelling gains (if gains at all)

Reference architecture

- CPU (Intel Core/Xeon)
- ◆ GPU (Nvidia)
- ▲ Movidius (NCS)
- TrueNorth

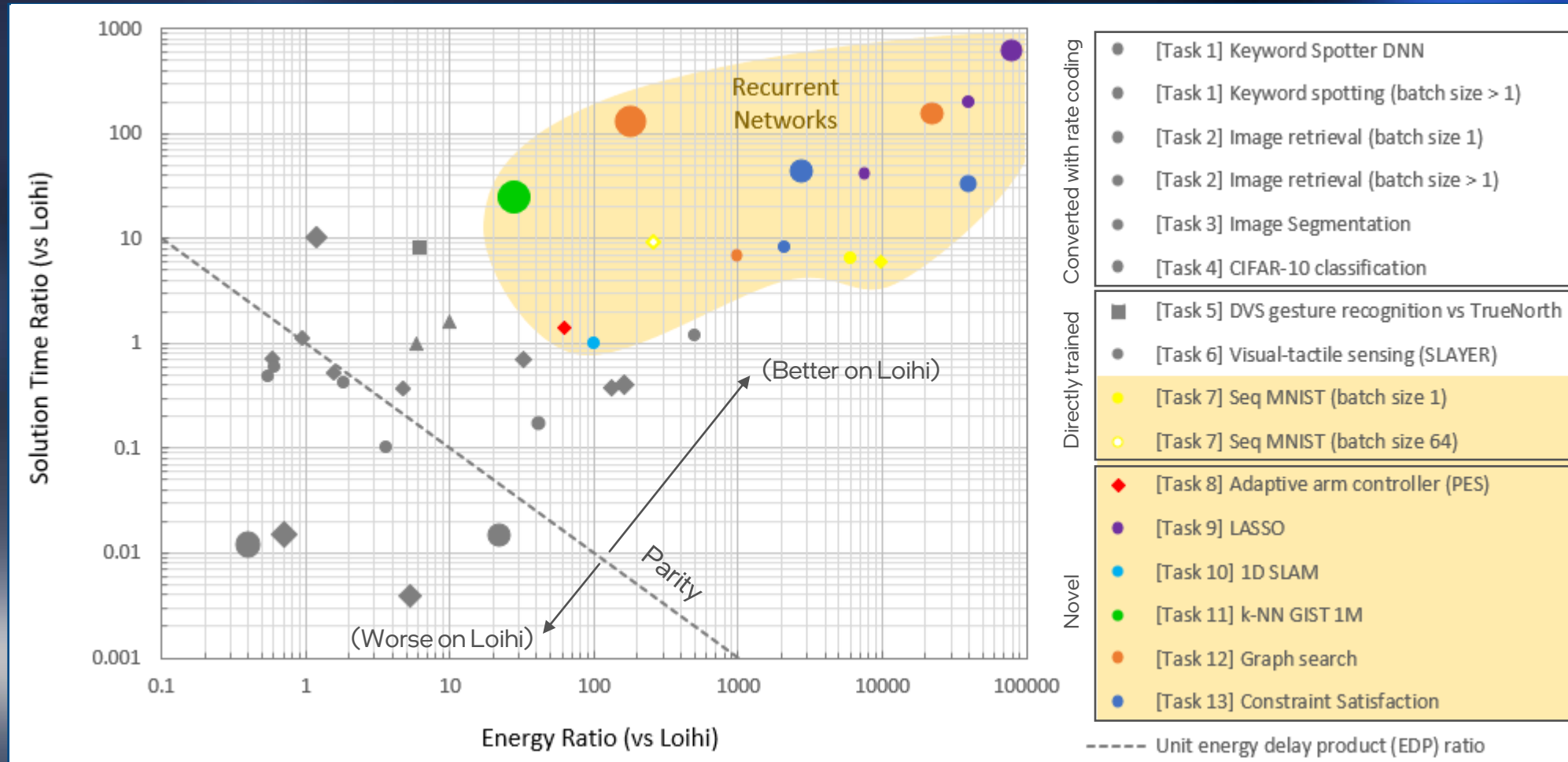


M. Davies et al, "Advancing Neuromorphic Computing With Loihi: A Survey of Results and Outlook," Proc. IEEE, 2021. Results may vary.

Recurrent networks with novel bio-inspired properties give the **best** gains

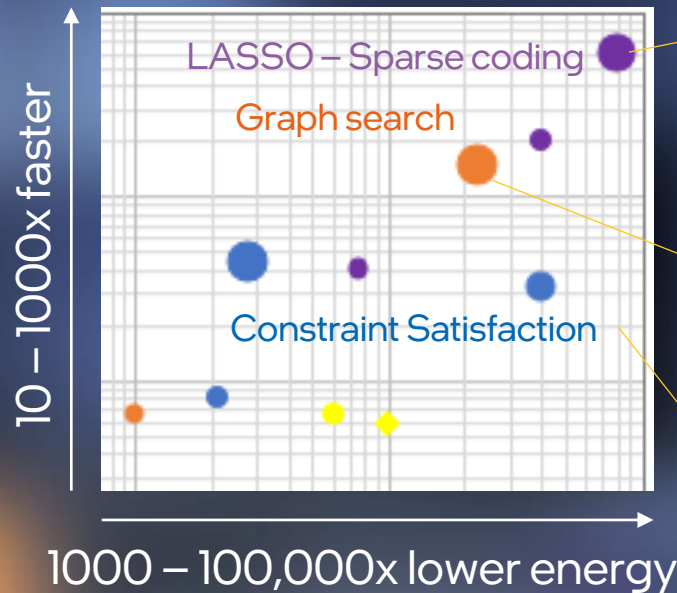
Reference architecture

- CPU (Intel Core/Xeon)
- ◆ GPU (Nvidia)
- ▲ Movidius (NCS)
- TrueNorth



M. Davies et al, "Advancing Neuromorphic Computing With Loihi: A Survey of Results and Outlook," Proc. IEEE, 2021. Results may vary.

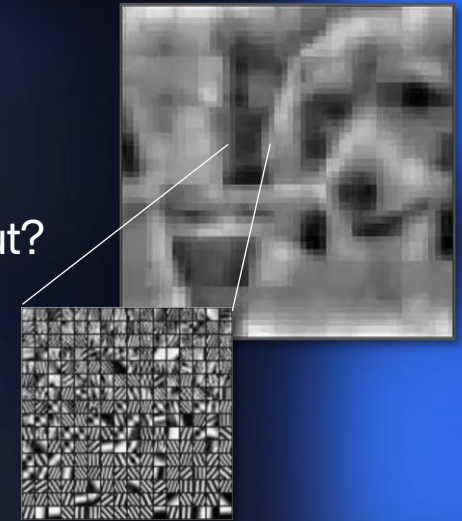
Zooming in on the best examples: Optimization problems



What features best explain the sensory input?

$$\underset{z}{\operatorname{argmin}} \|x - Dz\|_2^2 + \lambda \|z\|_1$$

Input Reconstruction Sparse regularization



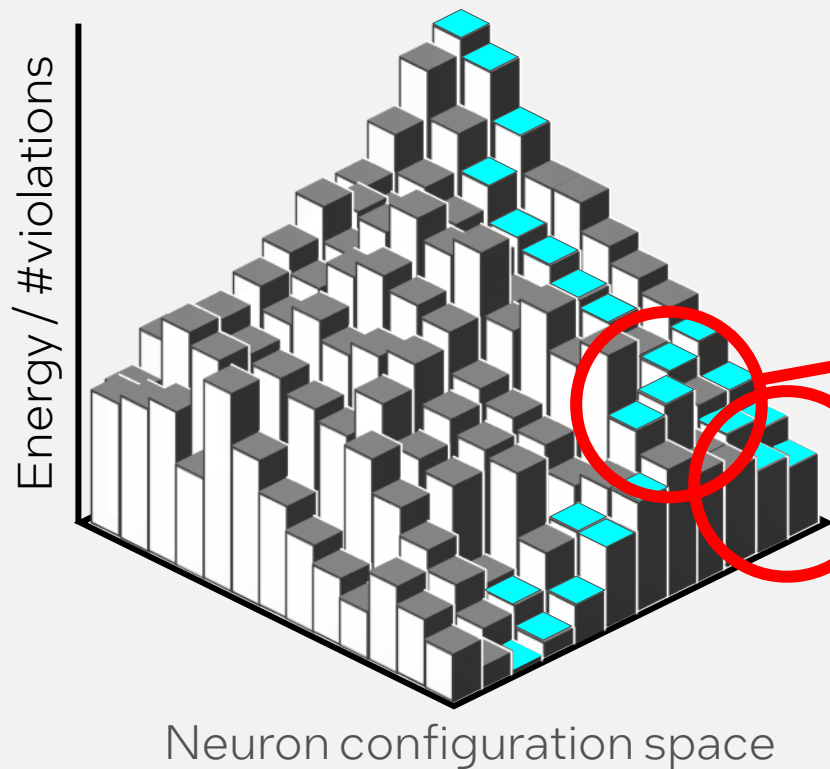
What is the shortest path to my goal?



What is the shortest path while visiting each waypoint exactly once?



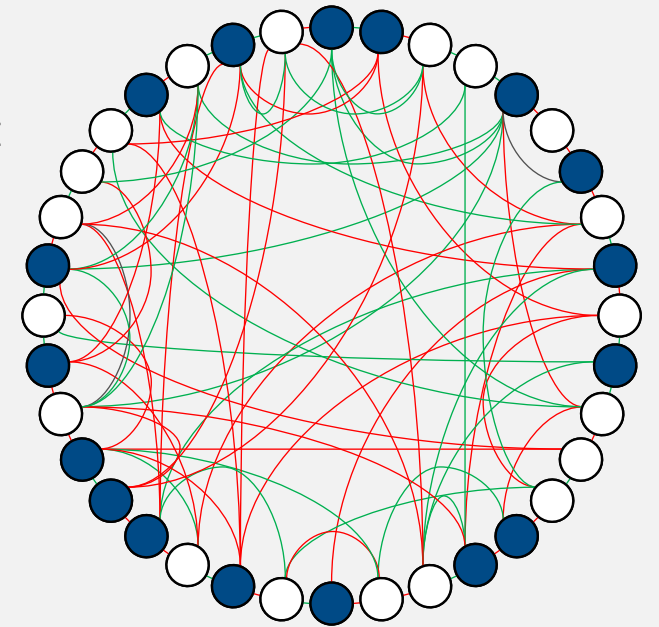
SNNs efficiently optimize via stochastic gradient descent



Neural dynamics descend the gradient

Local minimum
escaped by stochastic
spiking dynamics

Efficient descent due to massively
parallel, asynchronous neuromorphic
computing architecture

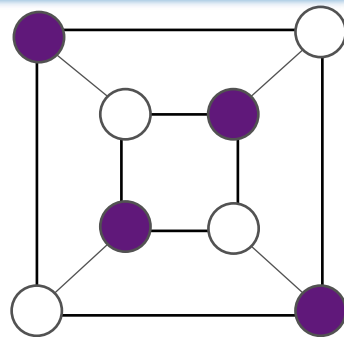


Loihi outperforms leading optimization solvers by orders of magnitude

QUBO (Maximum Independent Set)

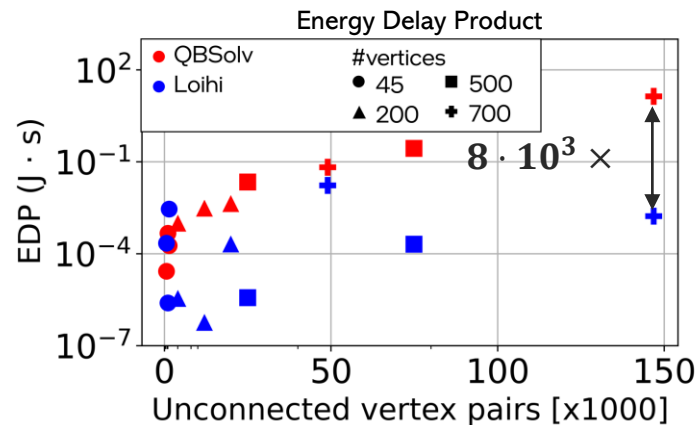
Workload:

Find largest set of unconnected vertices



Relevance:

- Target of SOTA quantum annealing approaches
- NP hard



Integer Linear Programming (Train Scheduling)

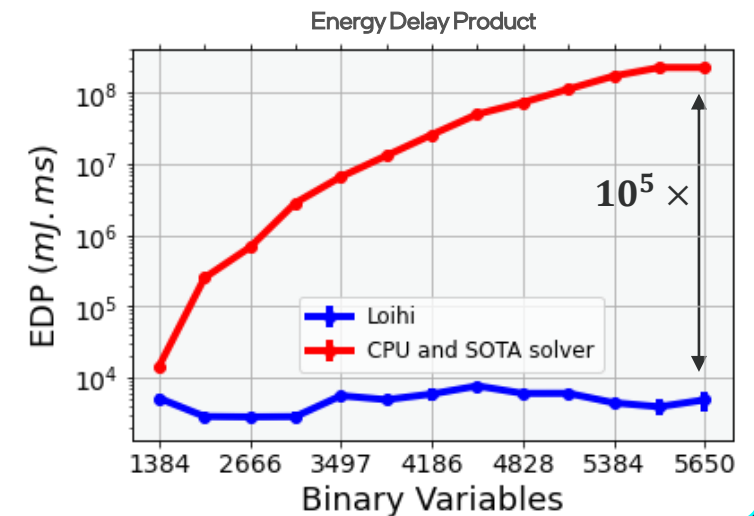
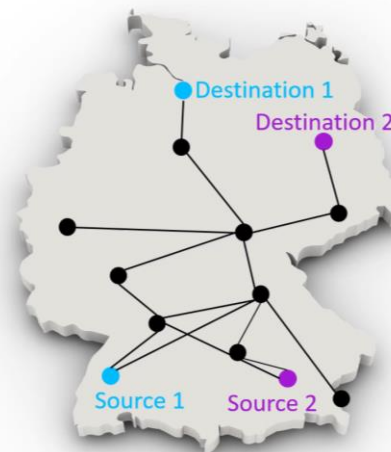
In collaboration with:

Workload:

Find the largest possible set of route assignments, given customer requests and railway, time and train constraints.

Relevance:

- Large-scale, real-world use case
- Applicable to resource allocation in warehouses and production lines.



Loihi: Nahuku board running NxSDK 0.95 with an Intel Core i7-9700K host with 128GB RAM, running Ubuntu 16.04.6 LTS

QUBO-QBSolv/CPU: benchmarks ran on an Intel Xeon CPU E5-2699 v3 @ 2.30GHz with 32GB DRAM (<https://github.com/dwavesystems/qbsolv>)

ILP-CPU: Xeon-based commercial cloud service as used operationally by DB. Solver runtime was measured; energy consumption estimated based on a 100W TDP estimate.

Performance results are based on testing as of September 2021 and may not reflect all publicly available security updates. Results may vary.

Generalizing neuromorphic optimization

Example Applications



Logistics

Train scheduling
Route optimization
Supply chain design
Job-shop scheduling
Flight gate assignment

CSP
QUBO
MILP
CSP
QUBO



Scientific computing

Prototype design
Material design
Particle jet reconstruction
Molecule structure prediction

MILP
LP
QUBO
QUBO



Robotics & AI

Trajectory optimization
Coordinating mobile robots
Model predictive control
Image compression

QP
MIQP
QP
CSP

Optimization Problem Class

	Problem	Domain	Constraints	Cost
CSP	constraint satisfaction problems	\mathbb{Z}^n	$\geq, =, \dots$	Constant
ILP	integer linear programming	\mathbb{Z}^n	$\geq, =$	Linear
LP	linear programming	\mathbb{R}^n	$\geq, =$	
MILP	mixed-integer linear programming	$\mathbb{Z}^n \cup \mathbb{R}^n$	$\geq, =$	Nonlinear: Quadratic
QUBO	quadratic unconstrained binary optimization	$\{0,1\}^n$	/	
QP	quadratic programming	\mathbb{R}^n	$\geq, =$	
MIQP	mixed-integer quadratic programming	$\mathbb{Z}^n \cup \mathbb{R}^n$	$\geq, =$	



Available on Loihi

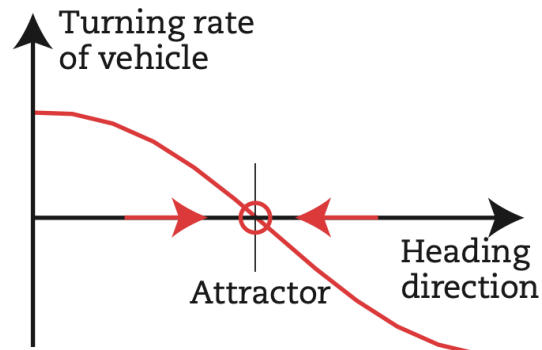
Work in progress

= Equality constraints

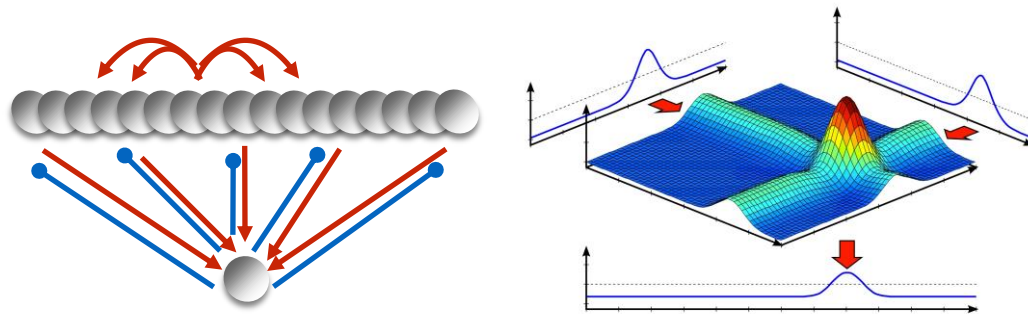
\geq Inequality constraints

Programming with Attractor Networks: Dynamic Neural Fields

Neural dynamics designed to create attractor states



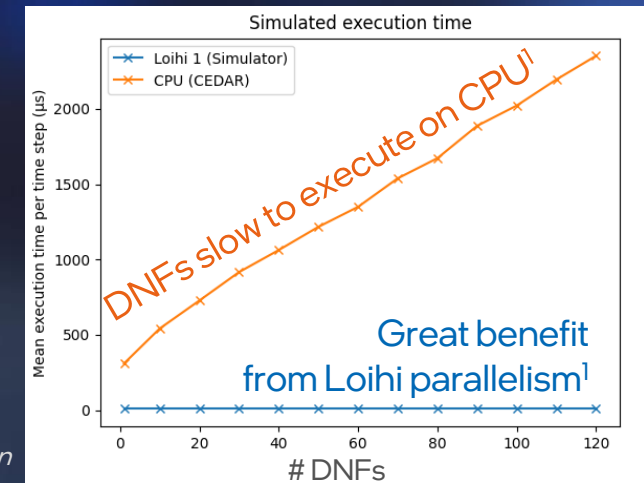
$$\tau \dot{\phi}(t) = -\phi(t) + A(t)$$



Attractor Networks as State Variables

- Memory – capture and hold inputs
- Attention – ignore distractors
- Defined state transitions

Execution time



¹See backup for characterization details. Results may vary.

Into a New Era of Neuromorphic Computing

Computational value is proven
(using today's manufacturing tech)

Motivates a new computational paradigm
(cheap, continuous optimization)

Many successful learning algorithms
(albeit shallow so far, not deep)

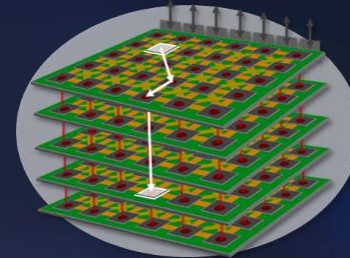
Properties of suitable applications:

- Power constrained
- Latency constrained
- Process real-time signals
- Slowly evolving structure
- Benefit from shallow online learning
- Apply deep learning for offline training

Outlook to Commercial Value

Scaled up systems

- Acceleration for datacenter optimization workloads
- Recommendation systems
- Scientific computing, HPC



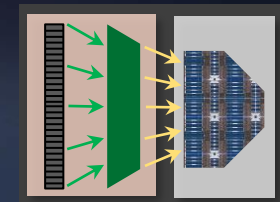
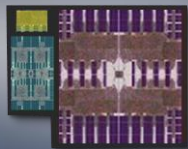
Intelligent Extreme Edge Co-Processors

- Aerospace and robotics devices
- Scene awareness and localization
- Model predictive control
- Navigation and planning
- Consumer devices (longer term)

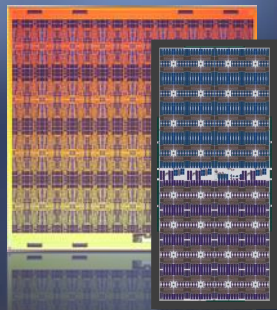


Specialized Designs

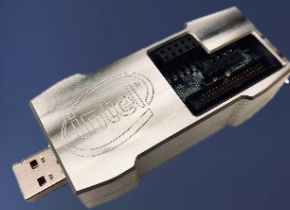
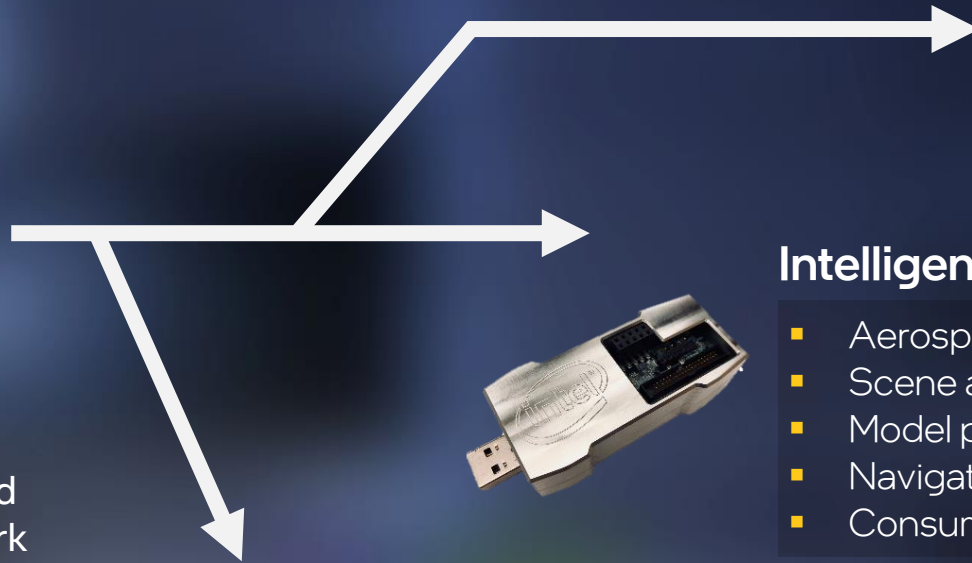
- Audio and other signal processing functions in SoCs
- Sensor integration (e.g. event-based cameras, electronic skins)
- Wireless signal processing and channel optimization
- IP and embedded accelerators for Intel Foundry customers



Today:



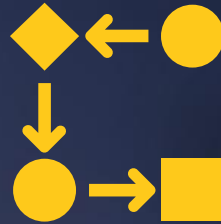
General-purpose research chips and software framework



Challenges and headwinds



High cost due to on-chip
memory integration



Algorithms and
Programming models



Software
convergence

A greatly improved Loihi 2 chip

10x Faster

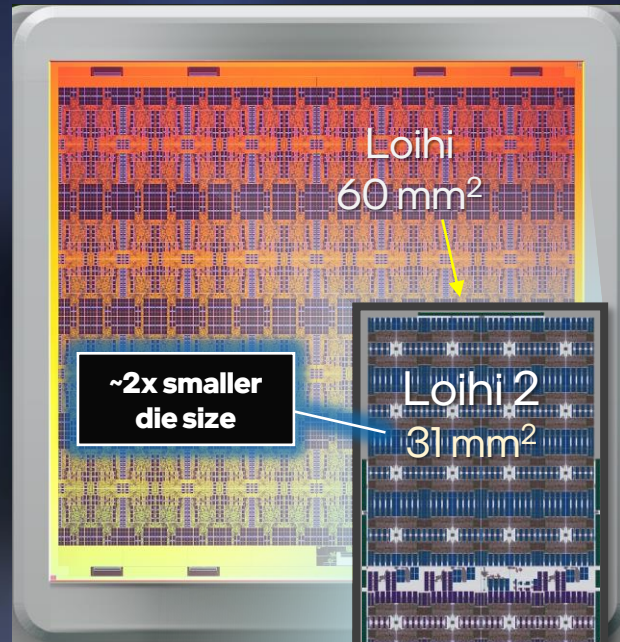
2-10x faster circuits² and design optimizations speed up workloads by up to 10x³

8x More Neurons

Up to 1 million neurons per chip with up to 80x better synaptic utilization, in 1.9x smaller die

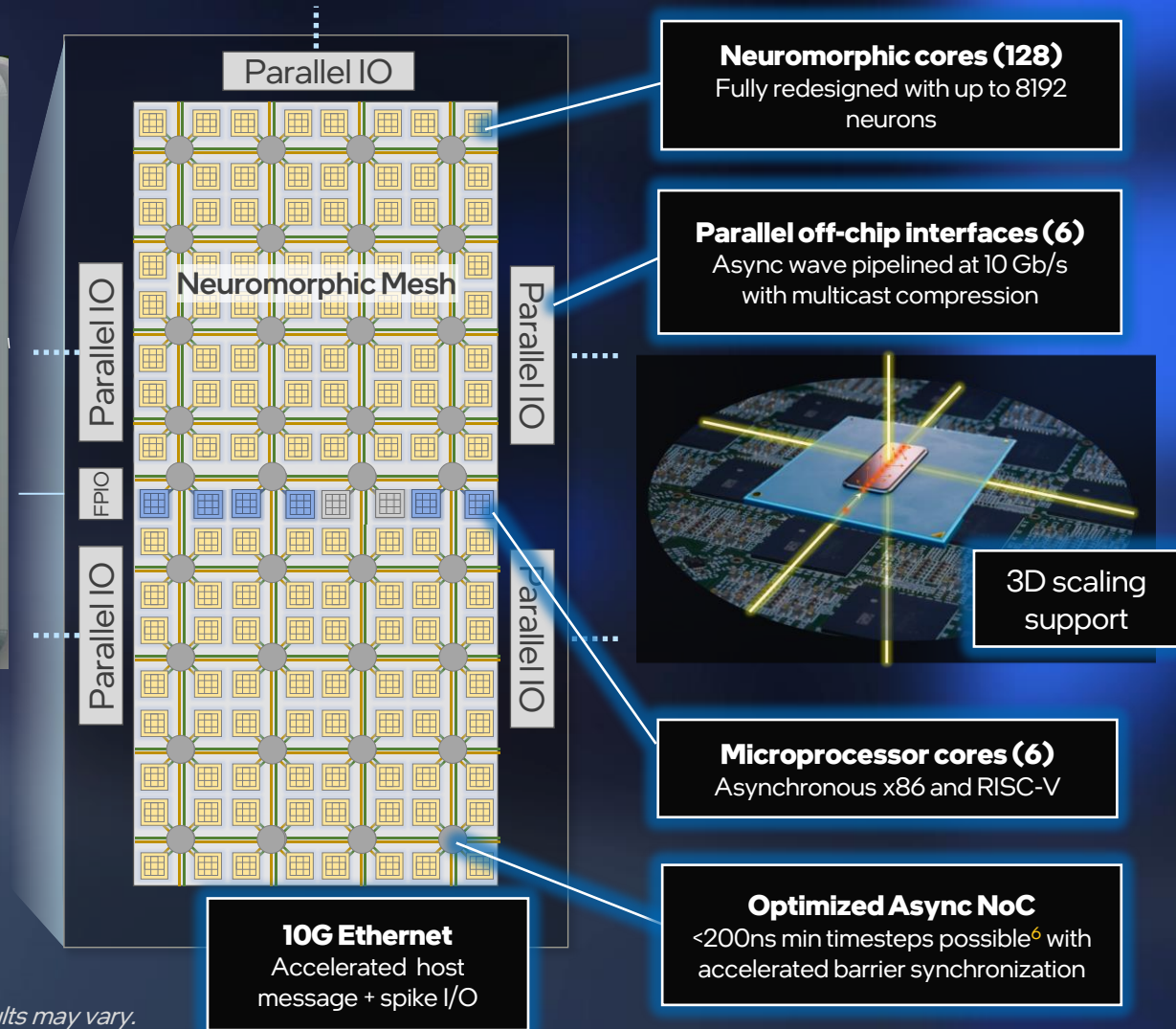
Better Scaling and Integration

3D scaling with 4x more bandwidth per link⁴, >10x compression⁵ with standard interfaces



	Loihi1	Loihi2
Neuron cores:	128	128
Max neurons:	130K	1M
Max synapses:	128M	123M
Max μ P cores:	3	6

^{2,3,4,5,6} See backup for characterization details.. Results may vary.



Generalized and optimized neuromorphic core

Generalized Spikes

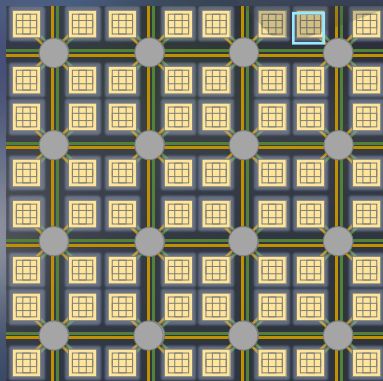
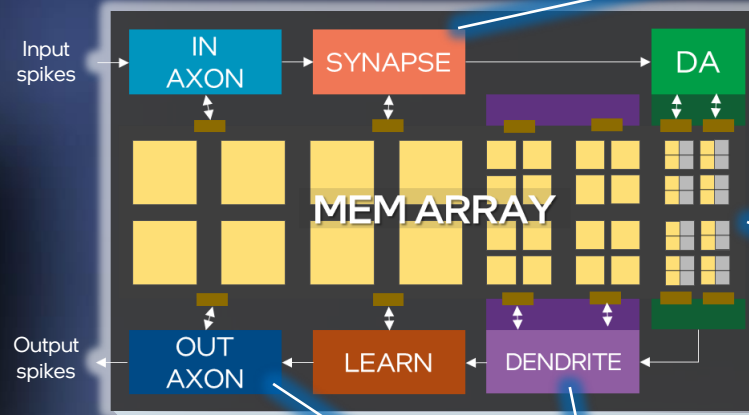
Spikes carry int8 magnitudes for greater workload precision

Programmable Neurons

Neuron models described by microcode instructions

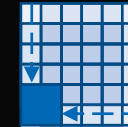
Enhanced Learning

Support for powerful new "three factor" learning rules from neuroscience



Better Synaptic Compression

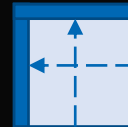
Convolution
Store kernel instead of connection matrix



Stochastic
up to 80x compression

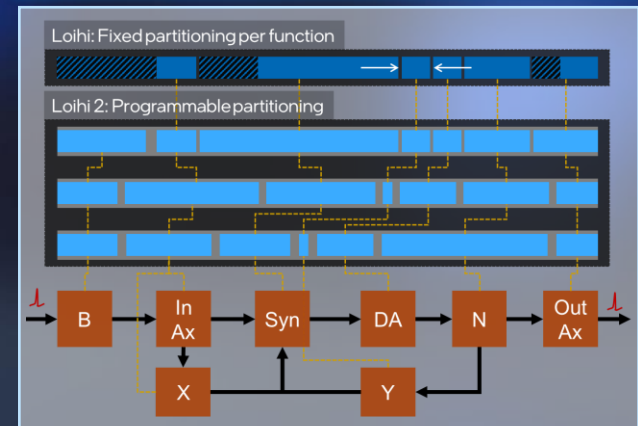


Factorized
 $O(n^2)$ to $O(n)$ compression



Better Utilization of Core Memory

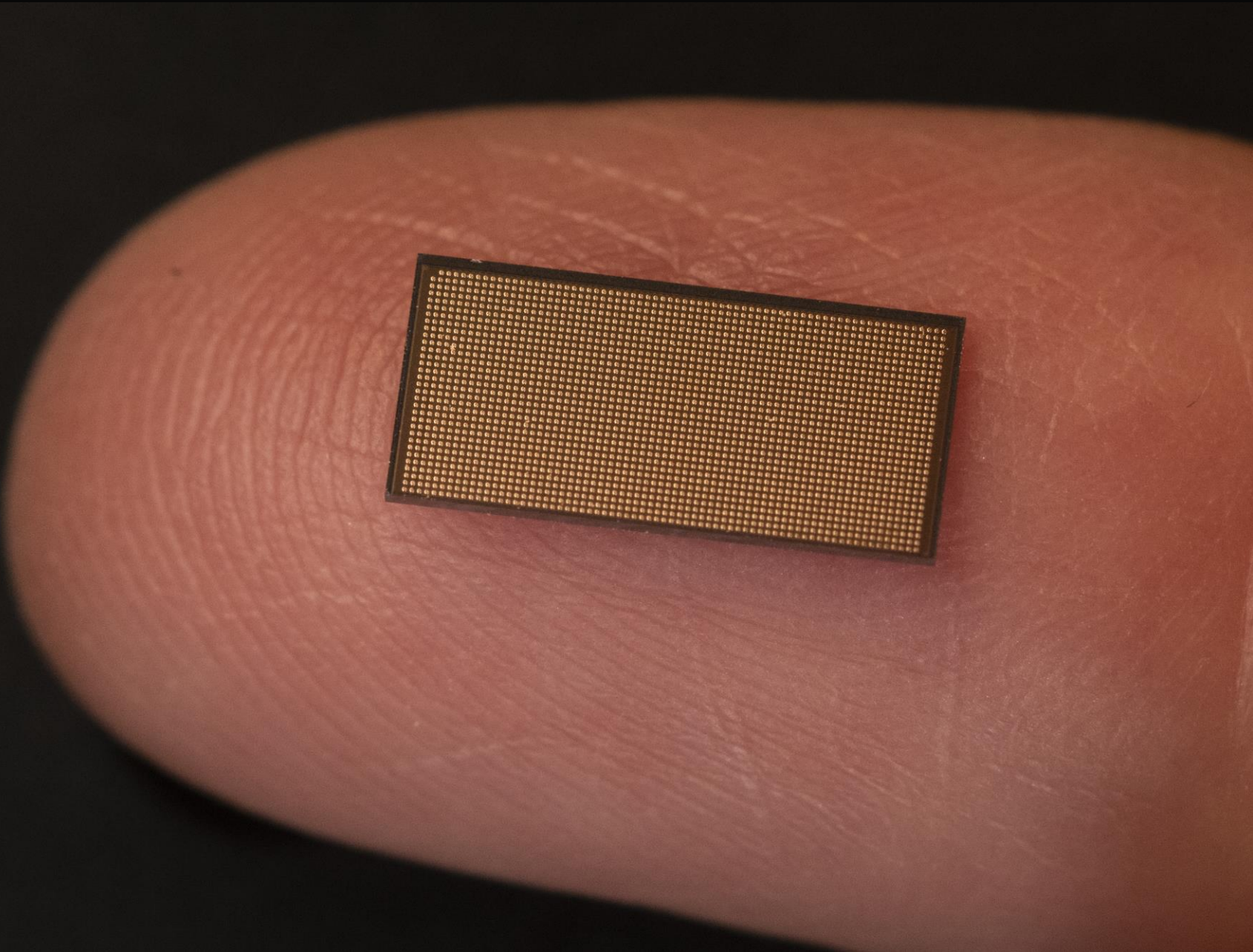
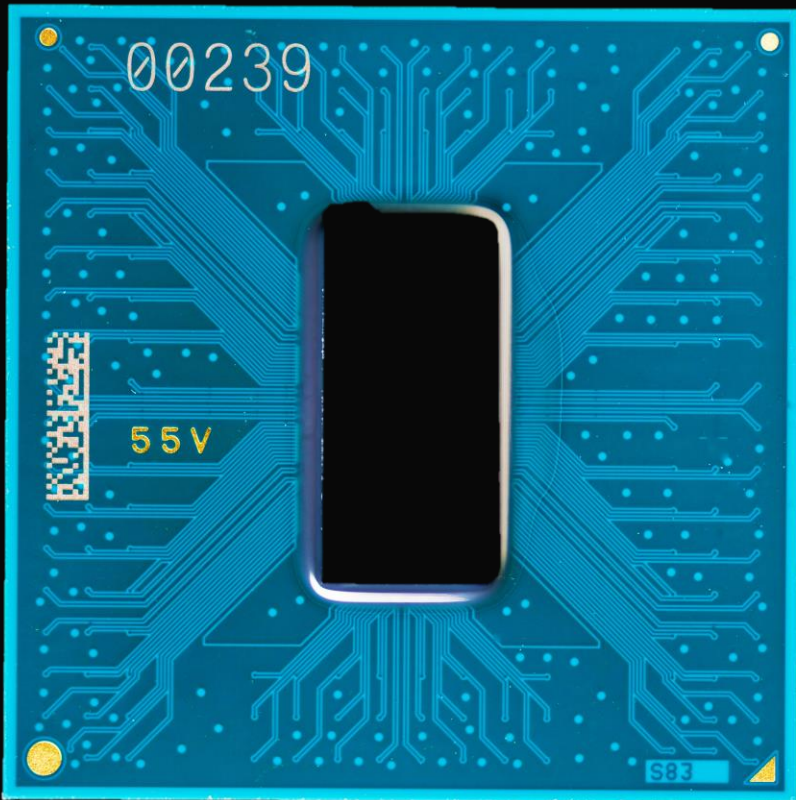
Highly ported centralized async memory array provides resource allocation flexibility



Better Neuron and Routing State Compression

Neuron state
~4x compression vs Loihi 1

Axon Routing
Up to 256x compression vs Loihi 1



Loihi 2 systems and characterization



Oheo Gulch
Single-chip system



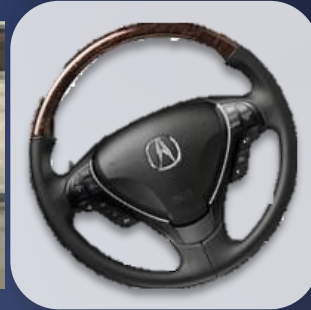
Kapoho Point
Stackable 8-chip board

Selected chip measurements

	Loihi 1 ⁷	Loihi 2 ⁶	Improvement
Neuron update time (ns)	9.6	4.4	2.2x faster
Synaptic Op time (ns)	4.0	0.66	6x faster
Minimum timestep (us)	1.57	0.19	8.3x faster
Neuron update energy (pJ)	70	56	25% lower
Synaptic Op energy (pJ)	21	7.8	2.7x lower

^{6,7} See backup for characterization details.. Results may vary.

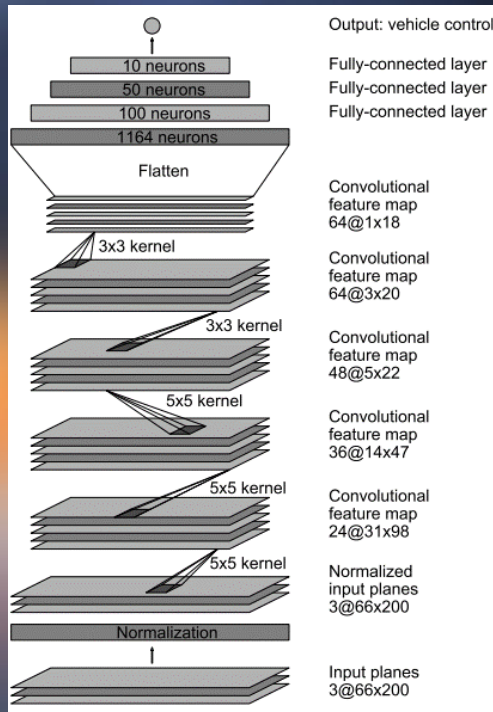
Deep Learning with Loihi 2



PilotNet: Predict steering wheel angle from dashboard video

Bojarski, Mariusz et al. "End to end learning for self-driving cars." *arXiv preprint arXiv:1604.07316* (2016).

Loihi 2 greatly improves Loihi 1's weakest results
(Feed-forward DNNs)



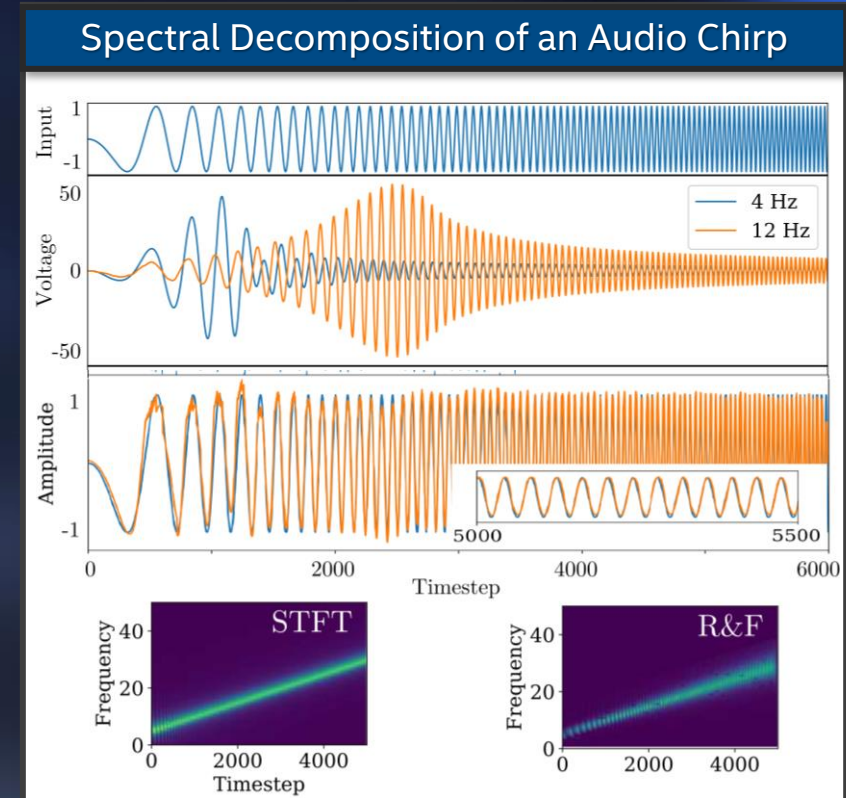
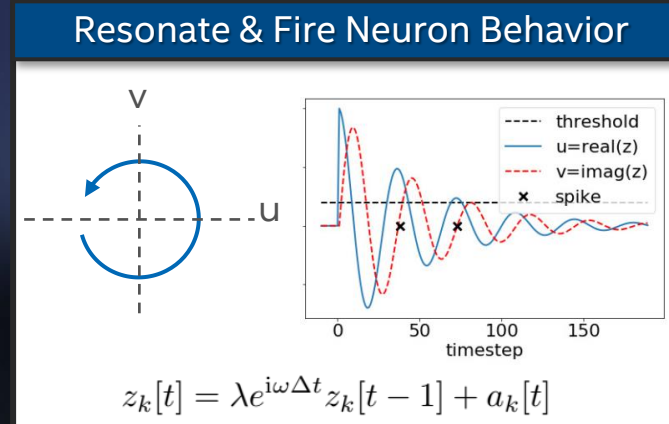
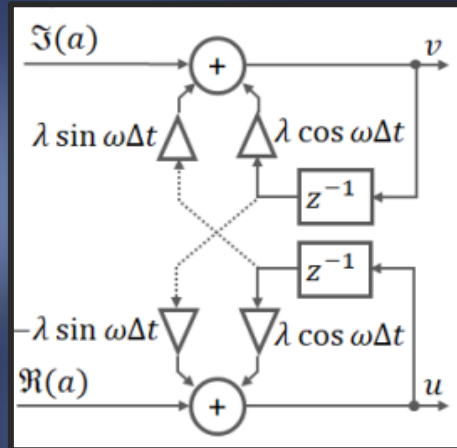
9-layer
Convolutional
network

	Loihi 1 LIF SNN ⁷	Loihi 2 LIF SNN ⁶		Loihi 2 - SDNN ³	
Mean-Square-Error	0.049	0.049	1	0.037	32% lower
Neuron cores	368	70	5x smaller	70	5x smaller
Latency (ms)	15.5	2.56	6x faster	1.22	9-12x faster
Throughput (fps)	808	4877		7404	
TOPS (DNN equiv)	0.05	0.166	6.5x better	0.25	15x more efficient
Energy (uJ/frame)	1770	270		120	
TOPS/W (DNN equiv)	0.02	0.13		0.28	

LIF SNNs are rate-coded and direct trained. SDNN is a sigma-delta coded ReLU network
All networks are trained with lava-dl. Unbatched data processing

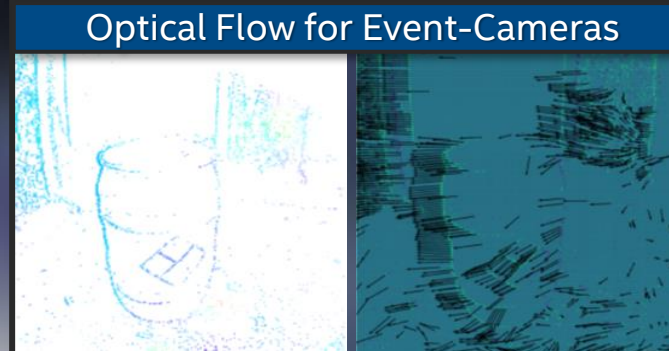
^{3,6,7} See backup for characterization details.. Results may vary.

Other new Loihi 2 networks: Resonate-and-Fire neurons



50x sparser output than conventional Short Time Fourier Transform

Resonate and Fire neurons compute optical flow for event-cameras with higher accuracy and 90x fewer ops than leading DNN solution





a new software framework for neuromorphic computing

Event-based communication
between simple parallel processes

Multi-Paradigm

Multi-Abstraction

Multi-Platform

Open source with permissive licensing of all core components

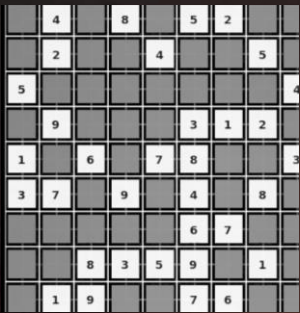
Today's SW for neuromorphic computing

	TensorFlow	PyTorch	Nengo	PyNN	NxSDK	BRIAN	ROS	Lava
Asynchronous message passing	✗	✗	✗	✗	✗	✗	✓	✓
CPU and GPU support	✓	✓	✓	✗	✗	✓	✓	✓
HW acceleration	✓	✓	✓	✓	✓	✗	✗	✓
Direct Backprop	✓	✓	✗	✗	✗	✗	✗	✓
Behavioral abstraction	✗	✗	✓	✗	✗	✗	✗	✓
Spiking neuron modeling	✗	✗	✓	✓	✓	✓	✗	✓
Permissive open source licensing	✓	✓	✗	✗	✗	✗	✓	✓

See <https://github.com/lava-nc>

Multi-Paradigm

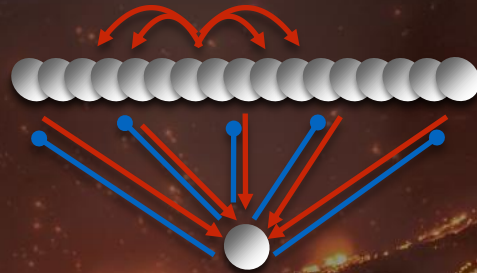
Optimization



LCA, Stochastic SNNs
LASSO, QP,
CSP, ILP, QUBO

+ model learning

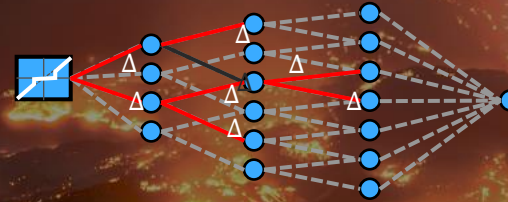
Neural Attractors



Dynamic Neural Fields,
Continuous Attractor NNs,
WTA

+ associative learning

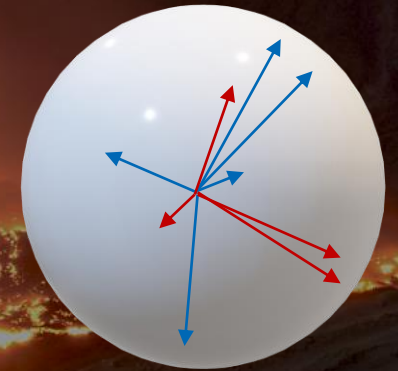
Deep Learning



ANN->SNN rate-coded conversion,
Directly trained SNN ConvNets
Sigma-Delta Neural Networks
TTFS- and Phase-coded SNNs

+ gradient learning

Vector Symbolic

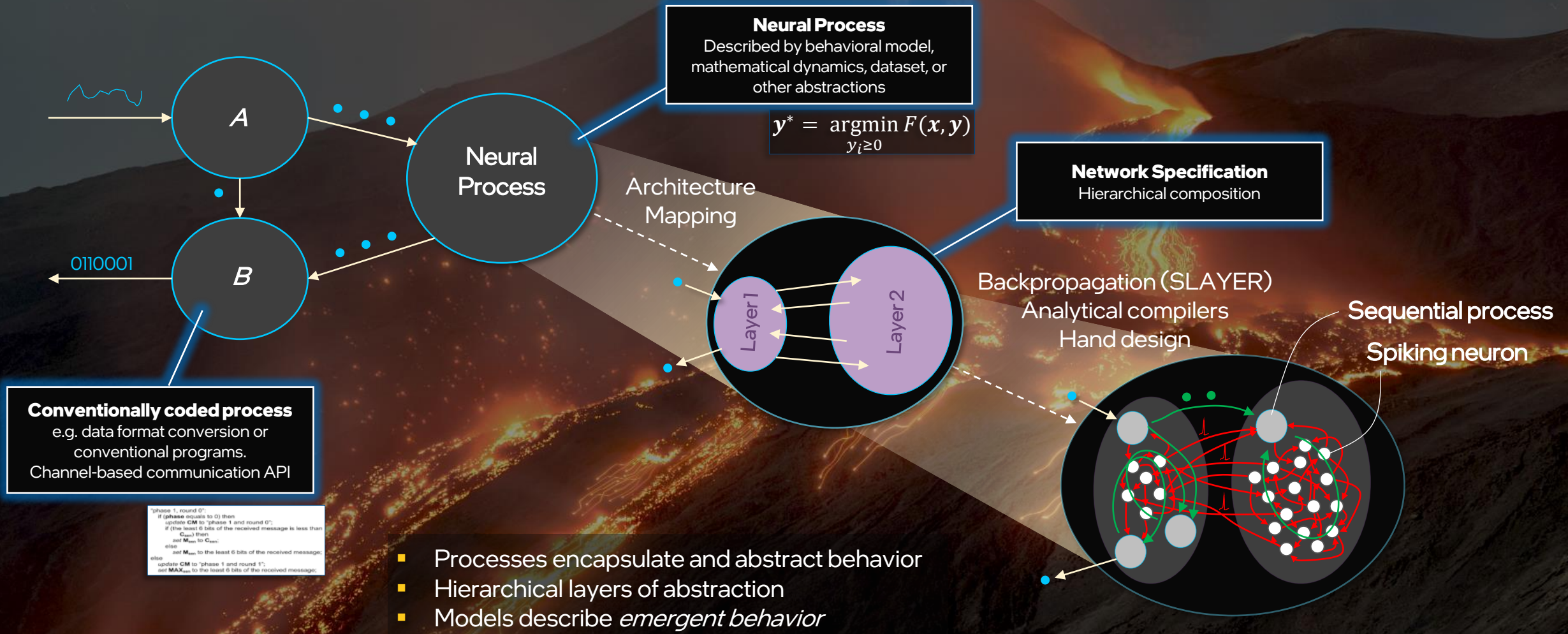


HRRs, MAPs,
Sparse Block Codes,
Associative Memories,
Resonator Networks

+ HD learning

Many others to come: NEF, Reservoir Computing, STICK, Equilibrium Propagation, evolutionary, ...

Multi-Abstraction



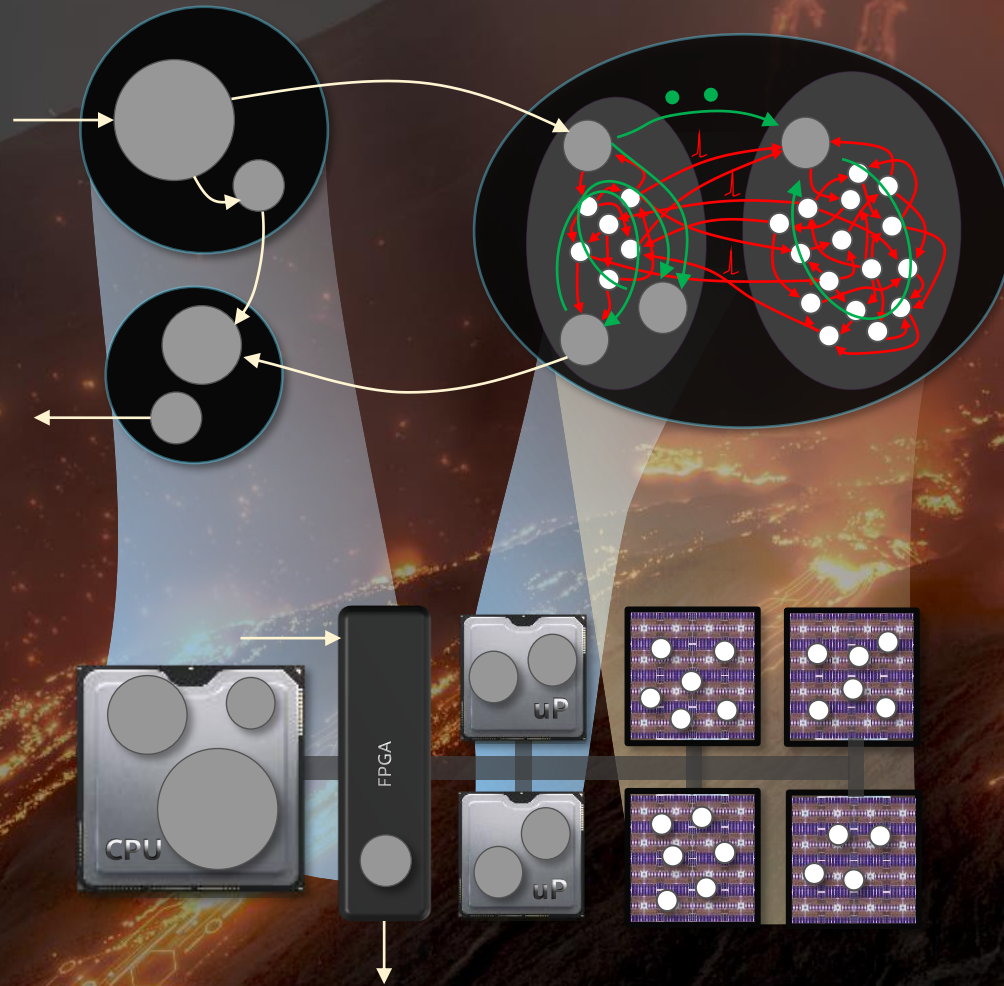
- Processes encapsulate and abstract behavior
- Hierarchical layers of abstraction
- Models describe *emergent behavior*

Multi-Platform

Abstraction Layer



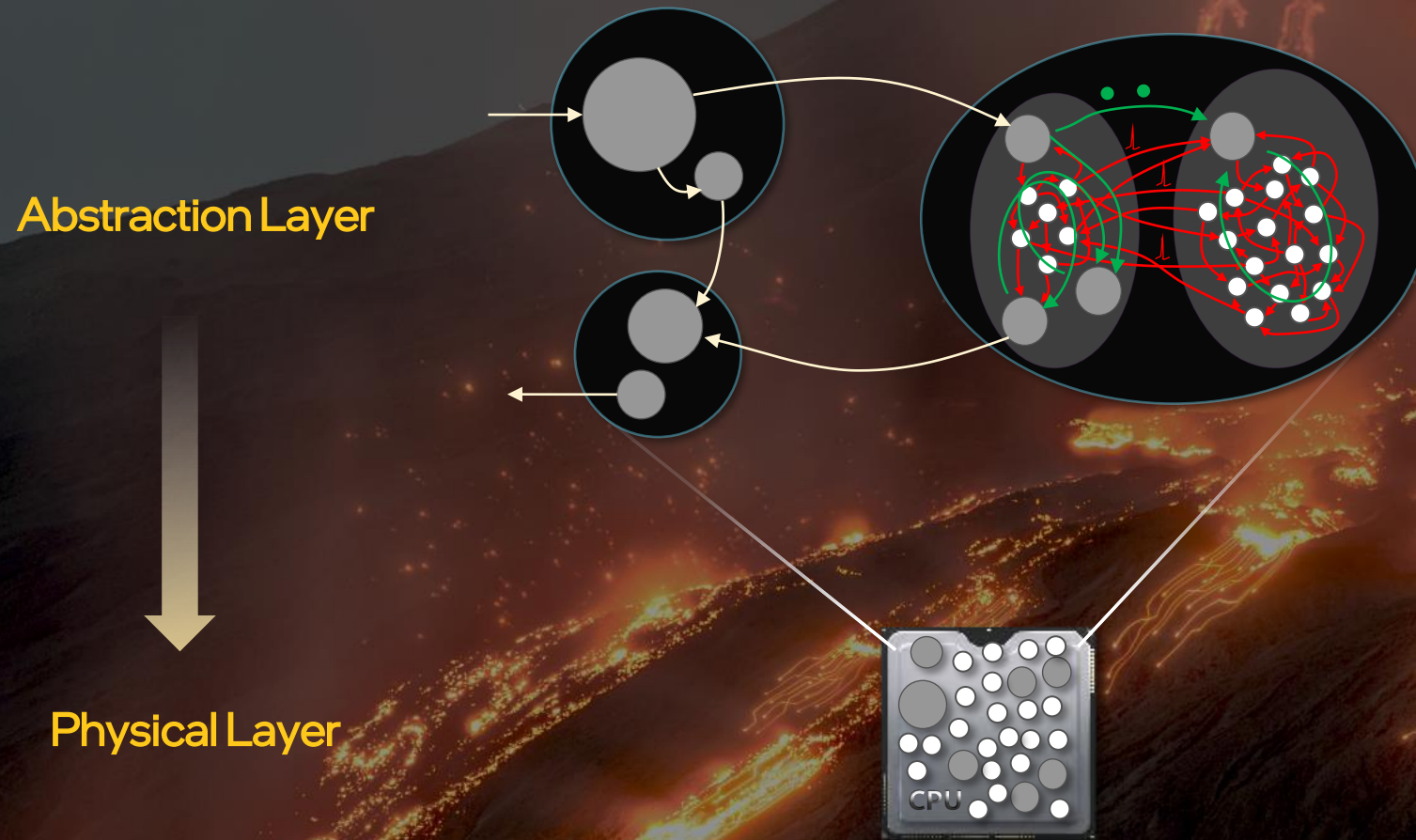
Physical Layer



- Heterogeneous system architecture
- Multi-backend execution + profiling
- Fast compilation and execution
- Performant real-time operation

- CPU
- GPU
- FPGA
- Loihi 1
- Loihi 2
- Others...

CPU-only Execution for Exploration and Prototyping



- Heterogeneous system architecture
- Multi-backend execution + profiling
- Fast compilation and execution
- Performant real-time operation

- CPU
- GPU
- FPGA
- Loihi 1
- Loihi 2
- Others...

Performance Analysis Details

¹ CPU dynamic neural field measurements obtained using repo version of Cedar (<https://cedar.ini.rub.de/>) as of October 2021 running on an Intel Core i7-4720HQ CPU with four threads, 128GB RAM, with Ubuntu 18.04 OS. Loihi 1 simulation measurements obtained using a silicon-calibrated Lava profiling model (unreleased) as of September 2021. Each DNF is a 2D mesh attractor with 27x27 neurons, with one input DNF fanning out to all other DNFs operating in parallel.

² Based on comparisons between barrier synchronization time, synaptic update time, neuron update time, and neuron spike times between Loihi 1 and 2. Loihi 1 parameters measured from silicon characterization (see below); Loihi 2 parameters measured from both silicon characterization with N3B1 revision and pre-silicon circuit simulations using back-annotated timing for Loihi 2.

³ Based on Lava simulations in September, 2021 of a nine-layer variant of the PilotNet DNN inference workload implemented as a sigma-delta neural network on Loihi 2 compared to the same network implemented with SNN rate-coding on Loihi. The Loihi 2 SDNN implementation gives better accuracy than the Loihi 1 rate-coded implementation. Equivalent DNN op counts calculated from a conventional DNN implementation with the same topology and same number of 8-bit parameters.

See Bojarski, Mariusz et al. "End to end learning for self-driving cars." arXiv preprint arXiv:1604.07316 (2016).

⁴ Circuit simulations of Loihi 2's wave pipelined signaling circuits show 800 Mtransfers/s compared to Loihi 1's measured performance of 185 Mtransfers/s.

⁵ Based on analysis of 3-chip and 7-chip Locally Competitive Algorithm examples.

⁶ Loihi 1 measurements were obtained on Oheo Gulch FMC board ncl-og-06 using an internal version of NxSDK advanced from v1.0.0

⁷ Loihi 2 measurements were obtained on Nahuku 32 board ncl-ghrd-01 using NxSDK v1.0.0

The Lava performance model for both chips is based on silicon characterization in September 2021 using the Nx SDK release 1.0.0 with an Intel Xeon E5-2699 v3 CPU @ 2.30 GHz, 32GB RAM, as the host running Ubuntu version 20.04.2. Loihi results use Nahuku-32 system ncl-ghrd-04. Loihi 2 results use Oheo Gulch system ncl-og-04.

Results may vary.

Thank You!



Email inrc_interest@intel.com for more information
Visit <https://github.com/lava-nc> to get started with Lava