New Tools for a New Era of Neuromorphic Computing

Mike Davies
Director, Neuromorphic Computing Lab

Neuro Inspired Computational Elements Conference - NICE 2022

March 30, 2022

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.

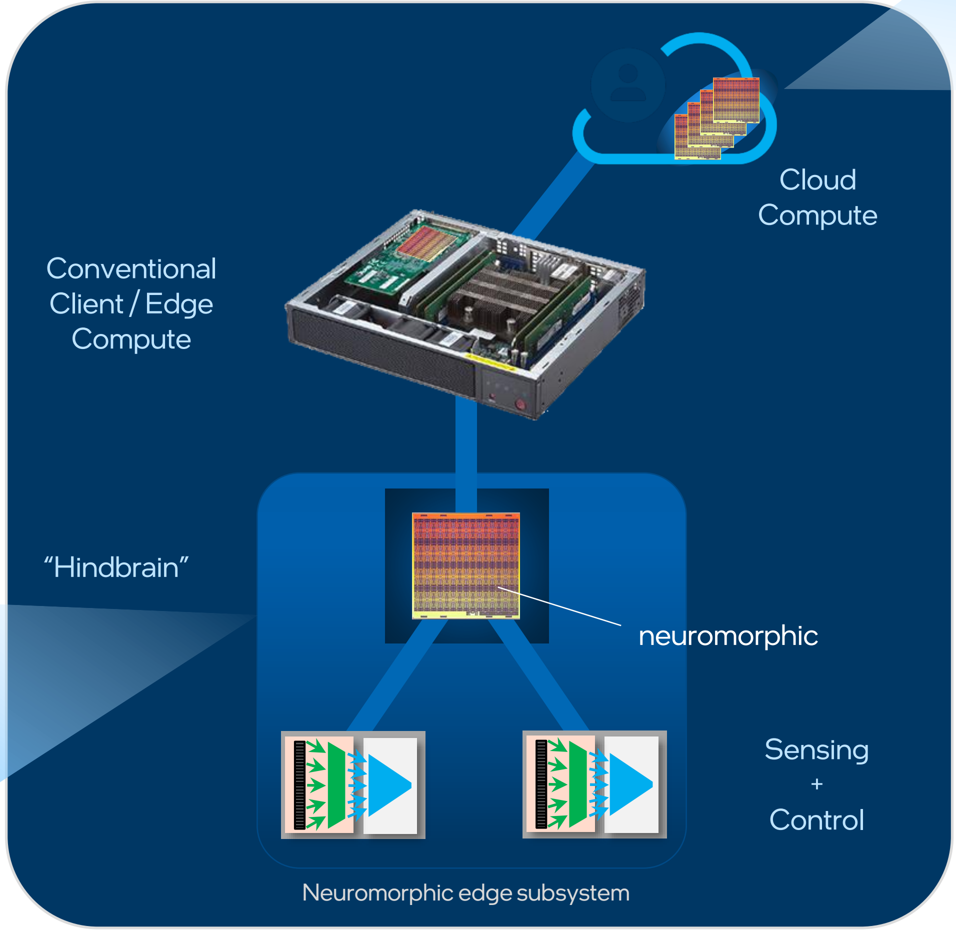
Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

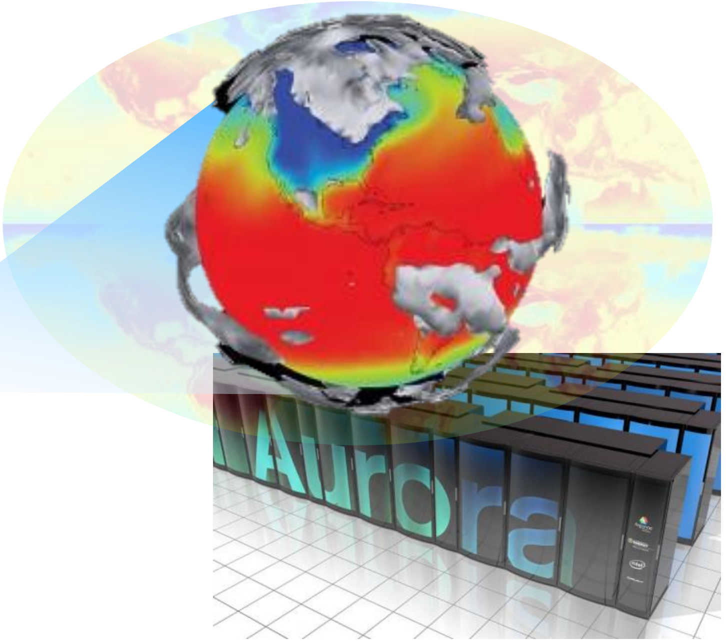
Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

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



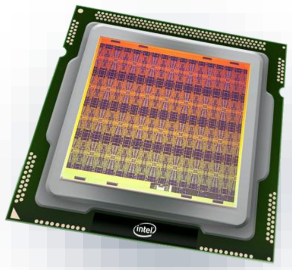


Achieve brain-like efficiency, speed, adaptability, and intelligence

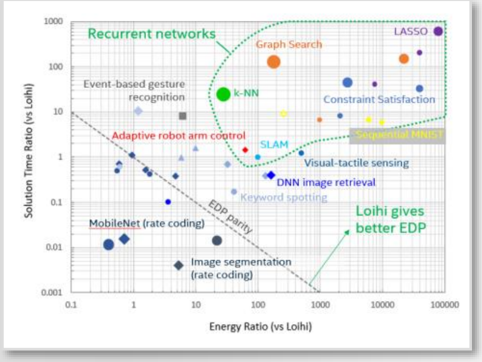
Significant progress over three years of Loihi research

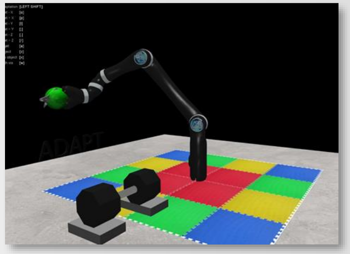
Loihi

NxSDK



Asynchronous Design



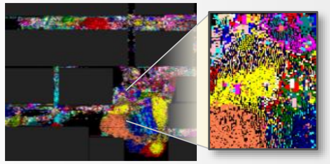


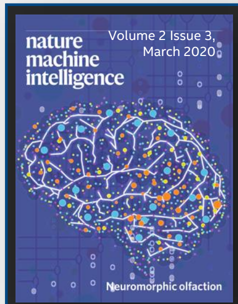
CIO JOURNAL

Intel to Release Neuromorphic-Computing System

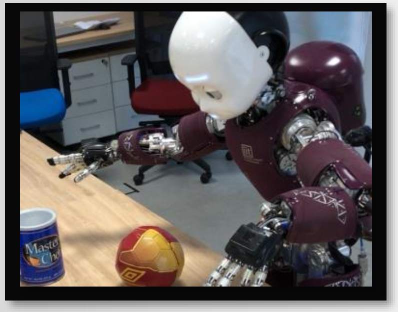
Pohoiki Springs, an experimental system to be rolled out this month, mimics the way human brains work to do computations faster with less energy

THE WALL STREET JOURNAL.





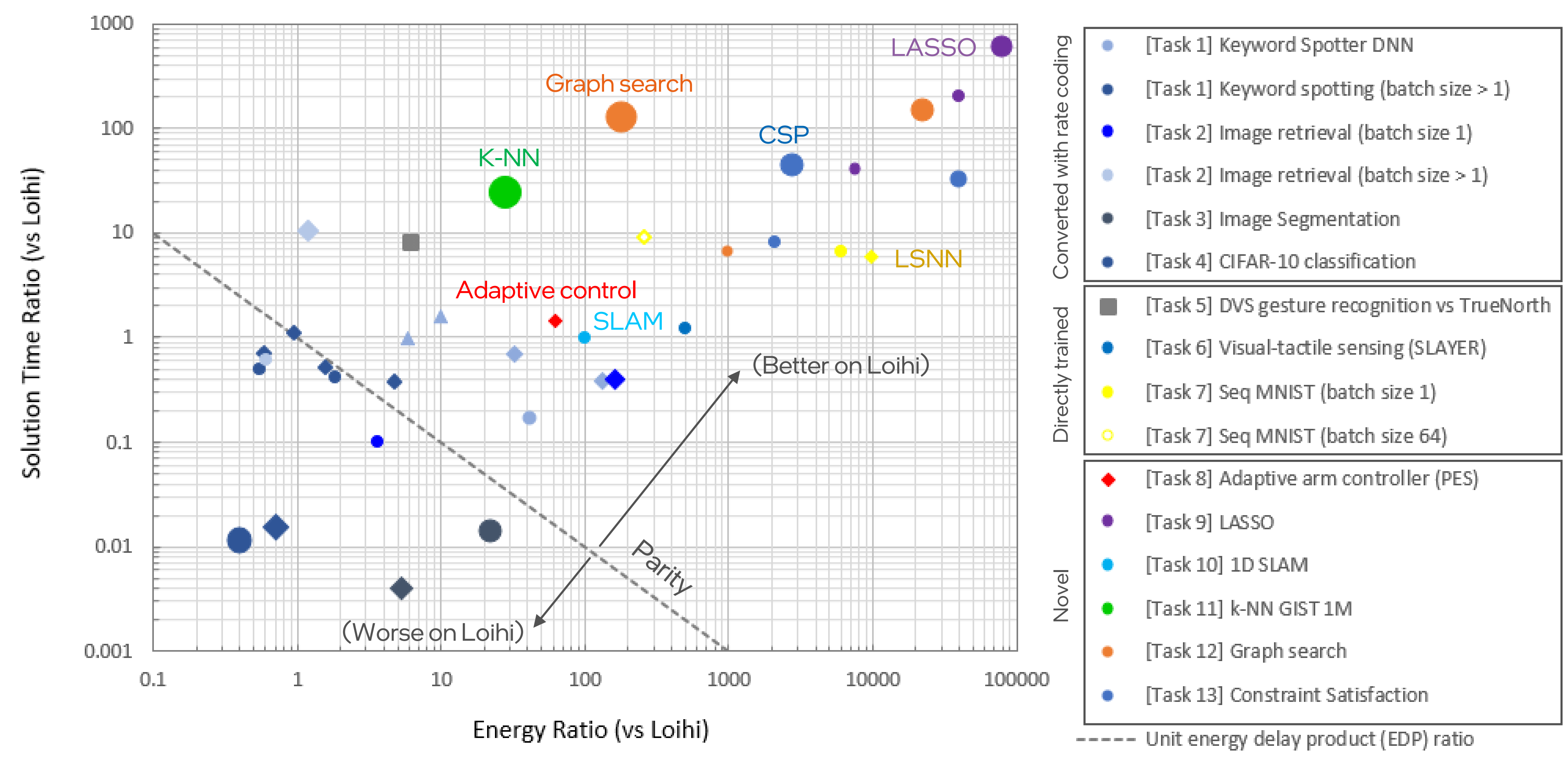
Neuro-inspired algorithm for odor recognition and learning demonstrated on Loihi, able to learn 3000x more data efficiently compared to DNN solution.



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

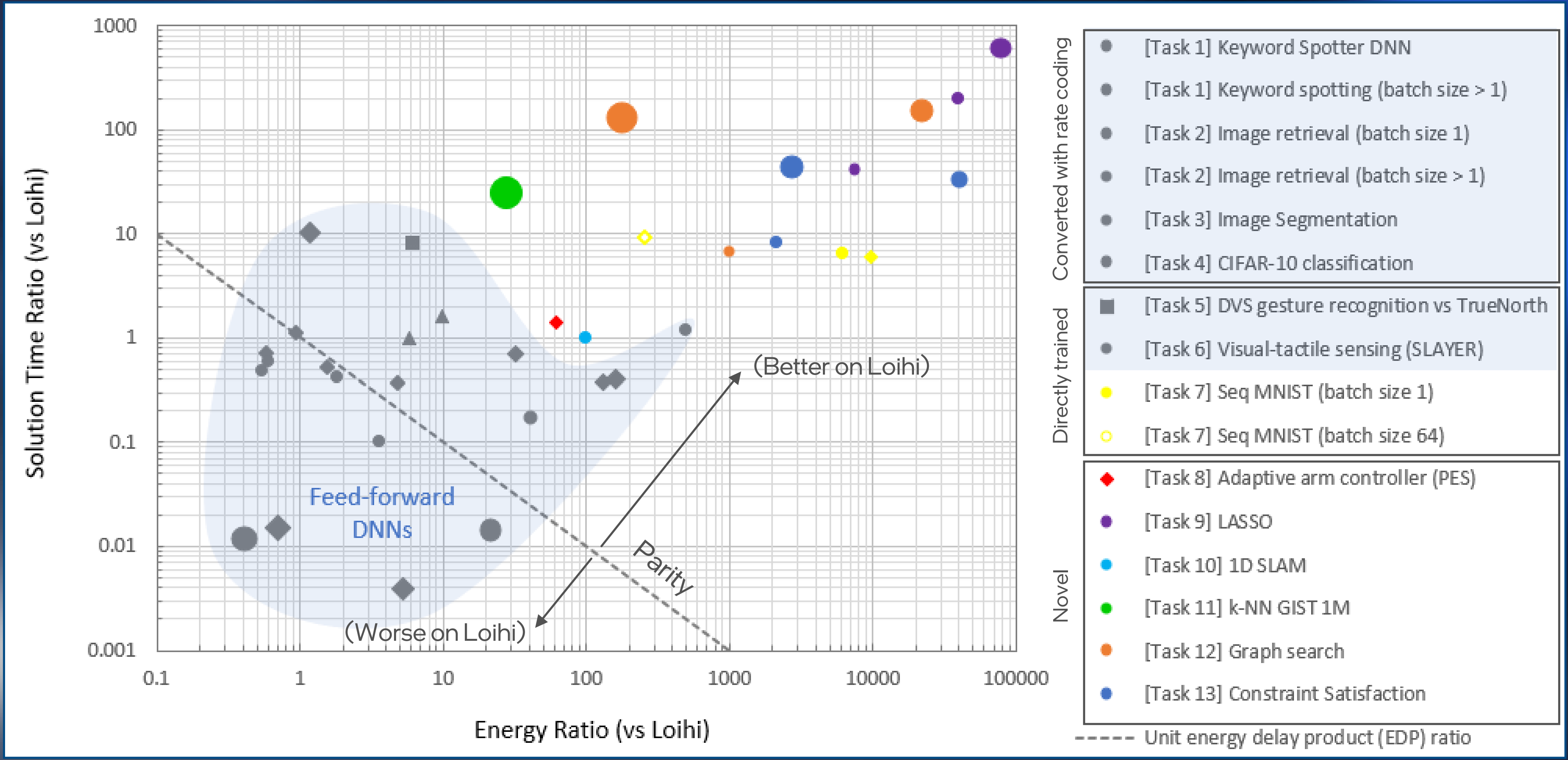


M. Davies et al, "Advancing Neuromorphic Computing With Loihi: A Survey of Results and Outlook," Proc. IEEE, 2021. Results may vary.

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

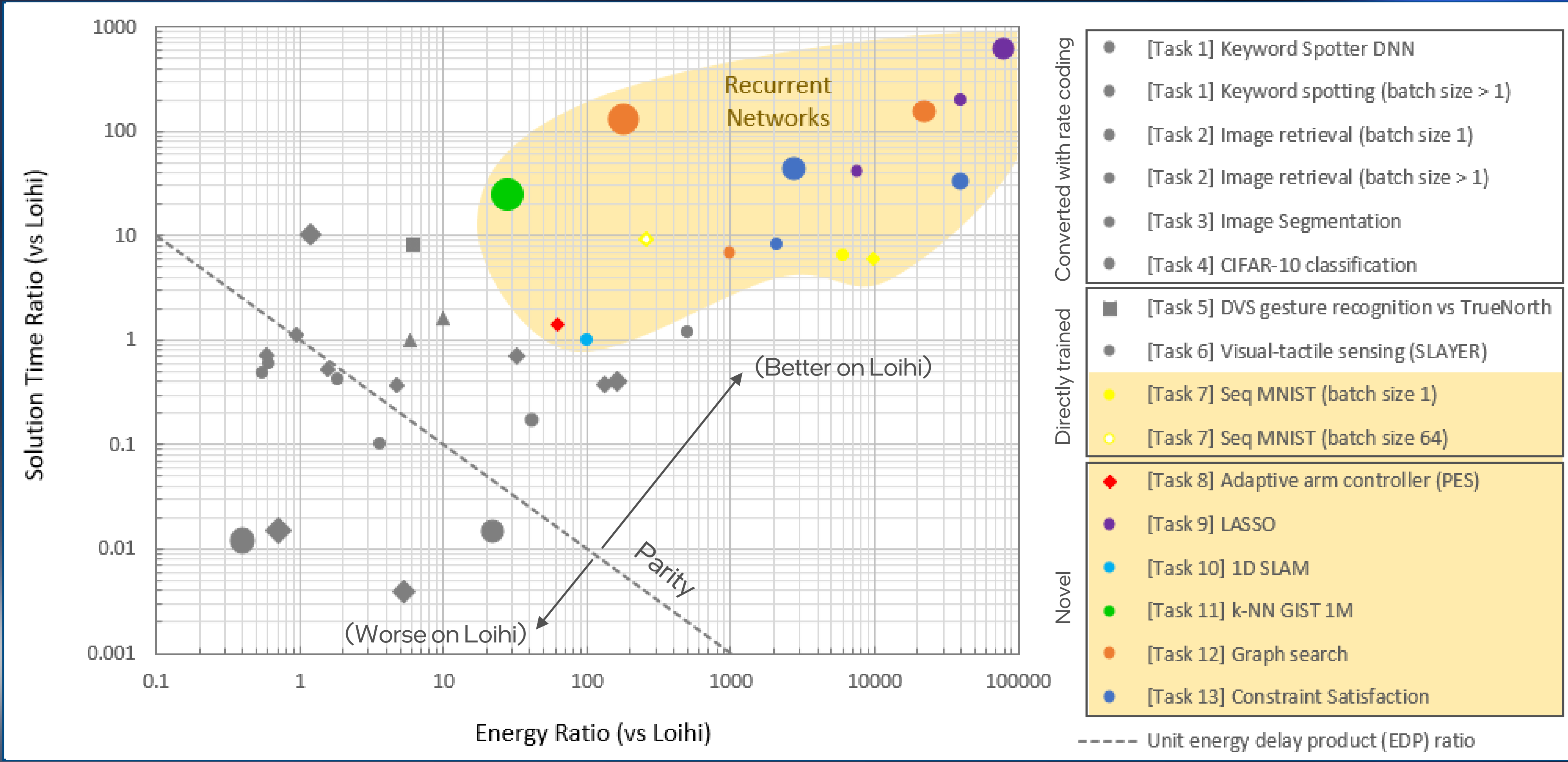


M. Davies et al, "Advancing Neuromorphic Computing With Loihi: A Survey of Results and Outlook," Proc. IEEE, 2021. Results may vary.

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

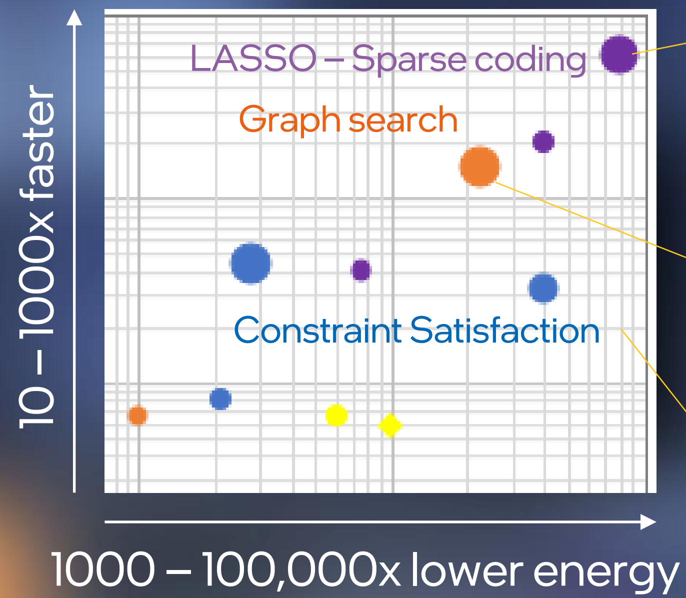
Reference architecture

- 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



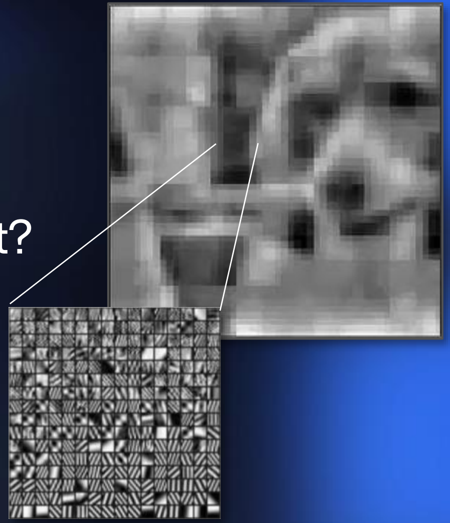
What features best explain the sensory input?

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

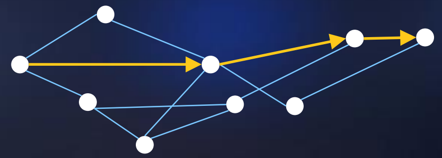
Input

Reconstruction

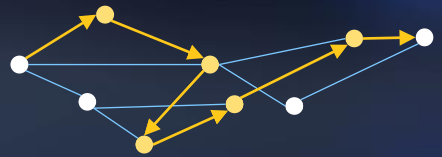
Sparse regularization



What is the shortest path to my goal?



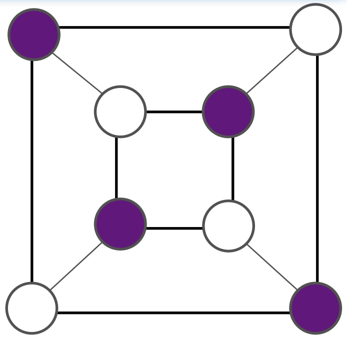
What is the shortest path while visiting each waypoint exactly once?



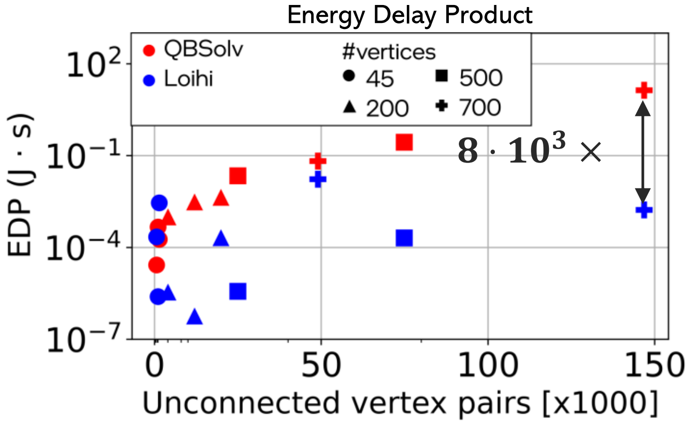
Loihi outperforms leading optimization solvers by orders of magnitude

QUBO
(Maximum Independent Set)

Workload:
Find largest set of unconnected vertices



Relevance:
- Target of SOTA quantum annealing approaches
- NP hard

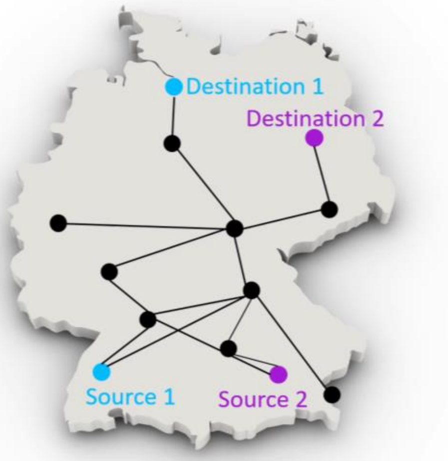


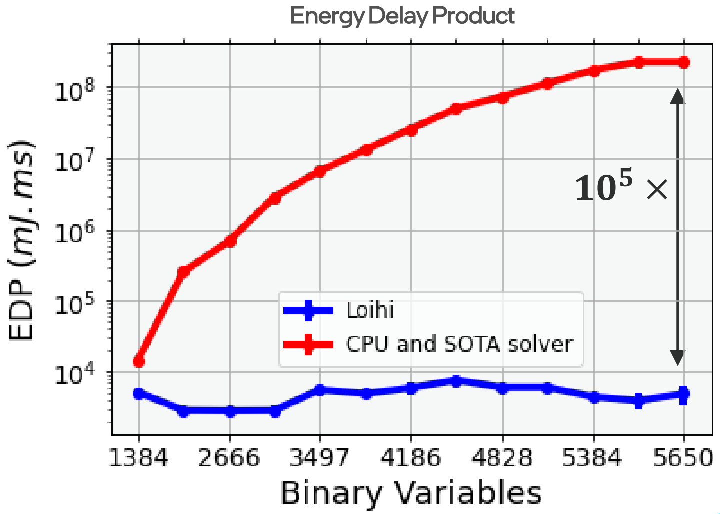
Integer Linear Programing
(Train Scheduling)

In collaboration with: DB

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: Commercial solver running on Linux64 with 16 processor cores.

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

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

Prototype design	MILP
Material design	LP
Particle jet reconstruction	QUBO
Molecule structure prediction	QUBO

Trajectory optimization	QP
Coordinating mobile robots	MIQP
Model predictive control	QP
Image compression	CSP

Optimization Problem Class

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

Available on Loihi
Work in progress
= Equality constraints
≥ Inequality constraints

Into a New Era of Neuromorphic Computing

Computational value is proven
using today's manufacturing tech

Embrace online optimization
a powerful computational primitive

Many successful learning algos
yet all shallow so far, not deep

Yet still many challenges...

Properties of suitable applications:

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

Challenges and headwinds

High cost due to on-chip memory integration

Algorithms and Programming models

Software convergence

A greatly improved Loihi 2 chip

Programmable Neurons
Neuron models described by microcode instructions

Generalized Spikes
Spikes carry integer magnitudes for greater workload precision

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

10x Faster
2-10x faster circuits[2] and design optimizations speed up workloads by up to 10x[3]

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 link[4], >10x compression[5] with standard interfaces

[2] Based on silicon characterization of Loihi 1 and a combination of silicon and pre-silicon simulation estimates for Loihi 2.

[3] Based on simulation modeling of a 9-layer Sigma-Delta Neural Network implementation of the PilotNet DNN inference workload compared to a rate-coded SNN implementation on Loihi 1.

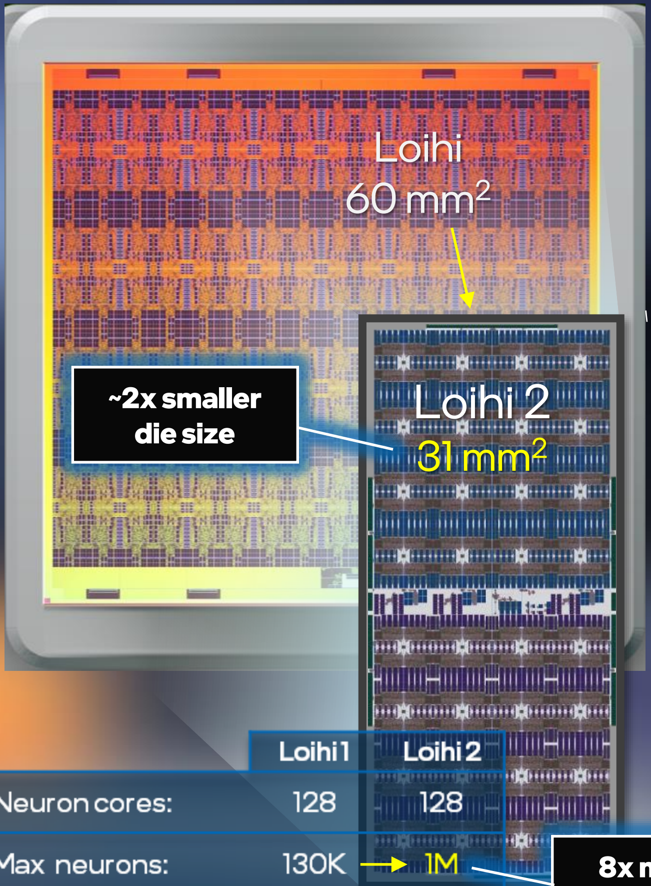
[4] Based on pre-silicon circuit simulations.

[5] Based on a 7-chip Locally Competitive Algorithm workload analysis.

See backup for analysis details. Results may vary.

https://download.intel.com/newsroom/2021/new-technologies/neuromorphic-computing-loihi-2-brief.pdf

More Resources, Better Packing, Greater Density

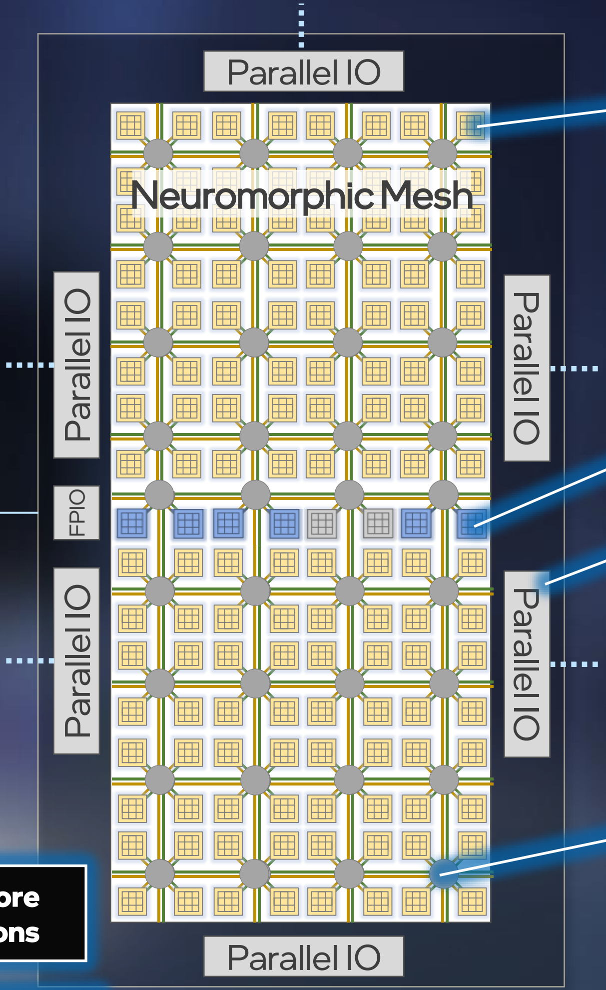


~2x smaller die size

	Loihi 1	Loihi 2
Neuron cores:	128	128
Max neurons:	130K	→ 1M
Max synapses:	128M	123M
Max μP cores:	3	→ 6

8x more neurons

2x more processors



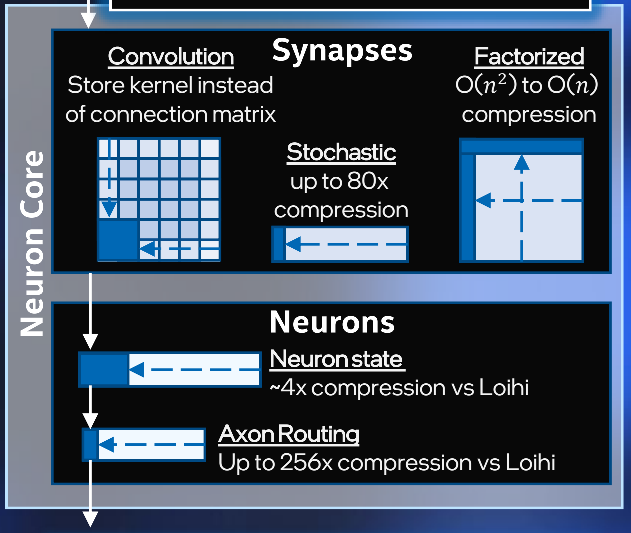
Neuromorphic core (128)
Programmable neuron model
Programmable learning
Up to 128kB synaptic memory
Up to 8192 neurons
Asynchronous design

Microprocessor cores (6)
Efficient spike-based communication
Data encoding/decoding
Network configuration

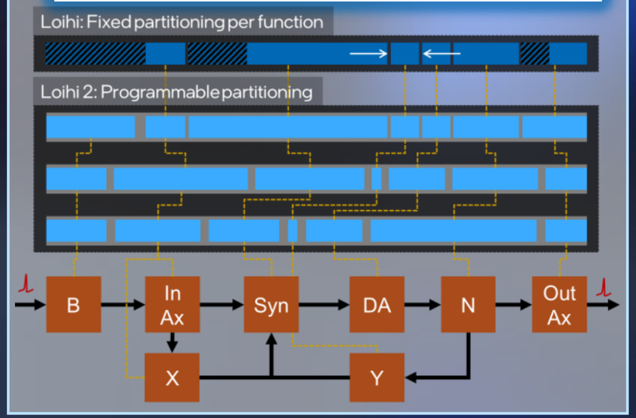
Parallel off-chip interfaces (6)
Faster chip-to-chip links, 3D scaling,
Support for standard synchronous protocols and event-based vision sensors

Low overhead NoC fabric
8x16-core 2D mesh
Scalable to 1000s of cores
Dimension order routed
Two physical fabrics
Acceleration for handshaking between cores

Better Synaptic Compression

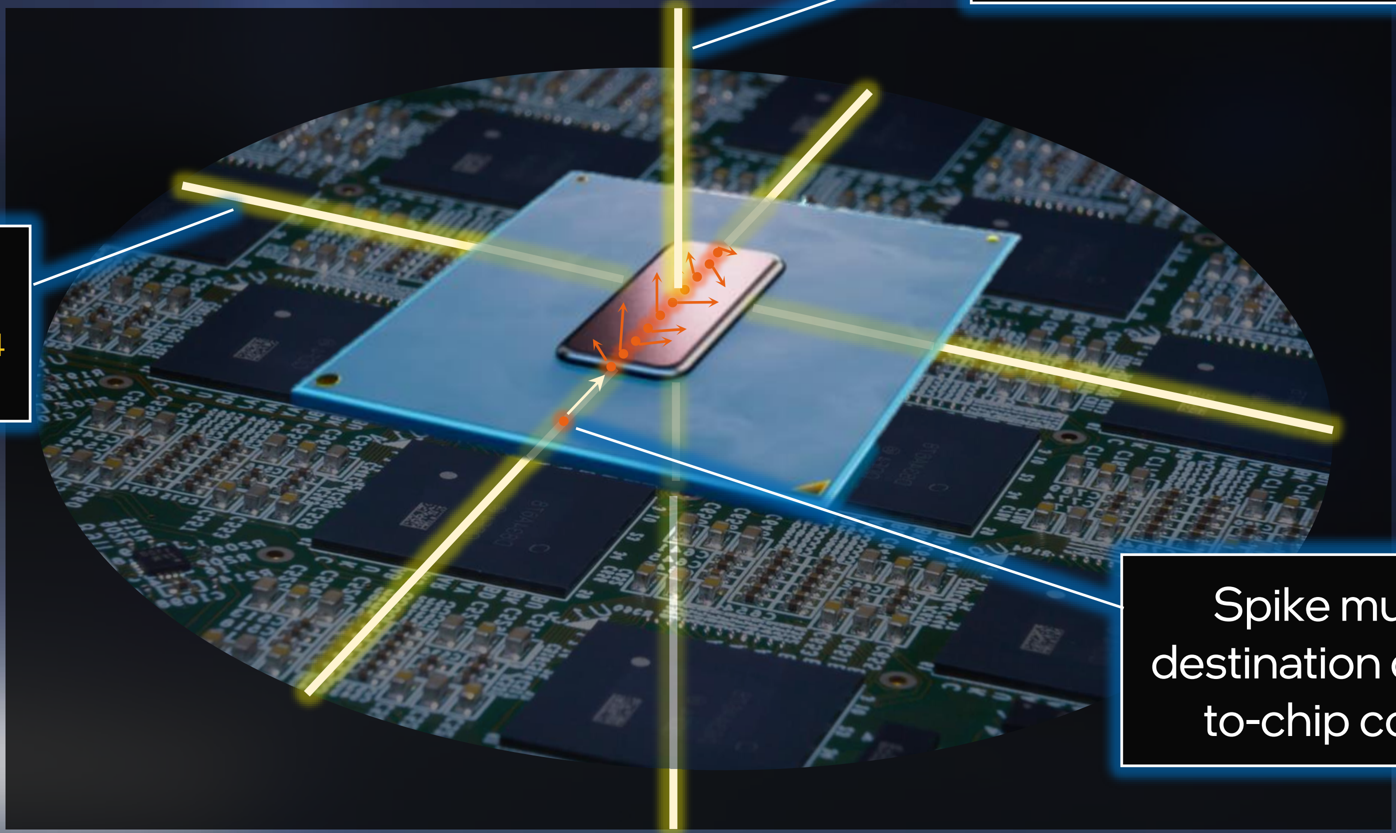


Better Utilization of Core Memory



Leading scalability

Radix-6 mesh routing for scaling in three dimensions



4x more bandwidth per chip-to-chip link[4]

- ~10 Gb/s
- Wave pipelined
- Single-ended

Spike multicast support on destination chips to reduce chip-to-chip congestion by >10x[5]

[4,5] *See backup for analysis details. Results may vary.*

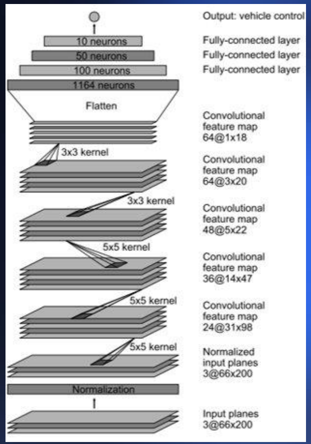
Loihi 2 versus Loihi 1: First silicon measurements

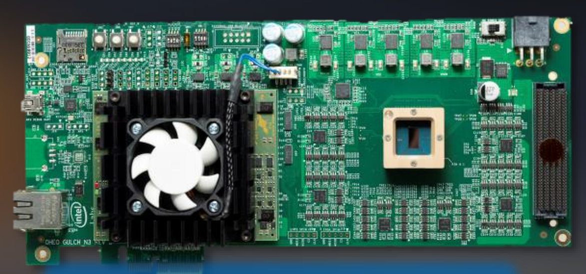
Chip measurements

	Loihi 1*	Loihi 2**	Improvement
Neuron updates 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

Feed-forward backprop-trained SNN

PilotNet convolutional network





Oheo Gulch

Single-chip system

Kapoho Point

Stackable 8-chip board

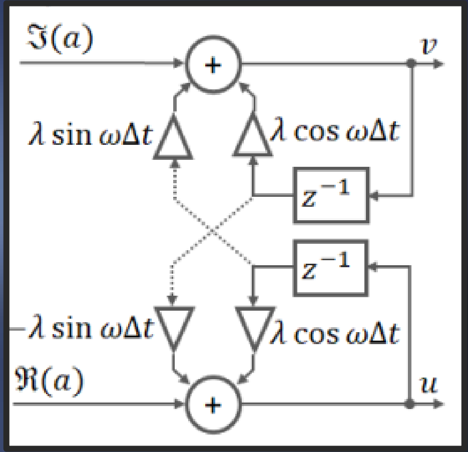
	Loihi 1*	Loihi 2**	Improvement
Resources (cores)	368	70	5.3x less
Speed (frames-per-sec)	101	610	6x faster
Energy (uJ)	14	2	7x lower
Energy-Delay-Product	140	3.6	39x better

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

* Measurements were obtained on Nahuku 32 board ncl-ghrd-01 using NxSDK v1.0.0

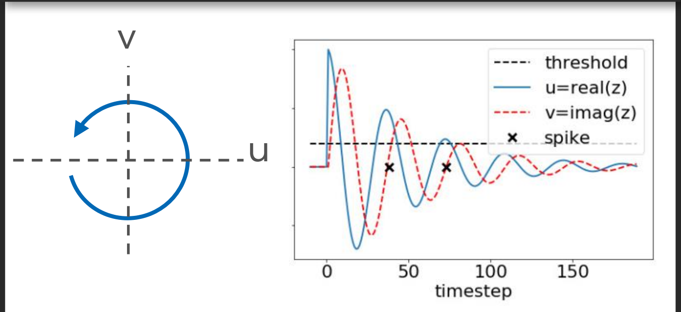
** Measurements were obtained on Oheo Gulch FMC board ncl-og-06 using an internal version of NxSDK advanced from v1.0.0

An example new direction: Resonate-and-Fire neurons



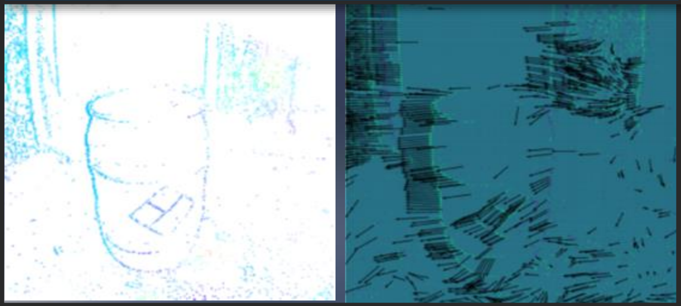
Resonate and Fire neurons compute optical flow for event-cameras with higher accuracy and 90x fewer ops than leading DNN solution

Resonate & Fire Neuron Behavior

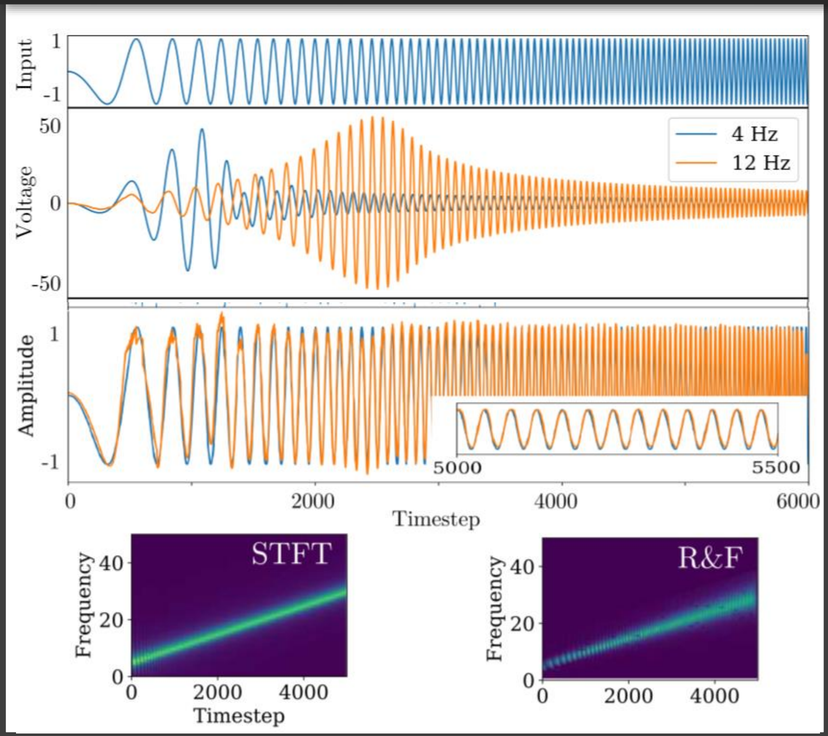


$$z_k[t] = \lambda e^{i\omega\Delta t} z_k[t-1] + a_k[t]$$

Optical Flow for Event-Cameras



Spectral Decomposition of an Audio Chirp



50x sparser output than conventional Short Time Fourier Transform

G. Orchard et al, "Efficient Neuromorphic Signal Processing with Loihi 2" IEEE International Workshop on Signal Processing Systems, Coimbra, Portugal, Oct 2021

LAVA a new software framework for neuromorphic computing

Event-based communication
between simple parallel processes

Multi-Paradigm

Multi-Abstraction

Multi-Platform

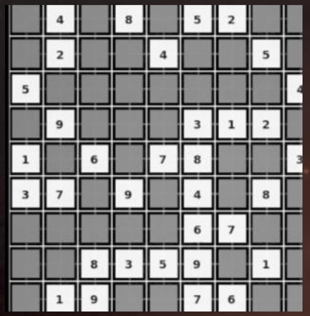
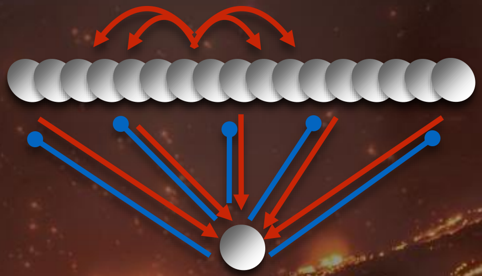
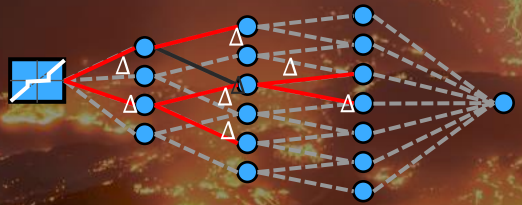
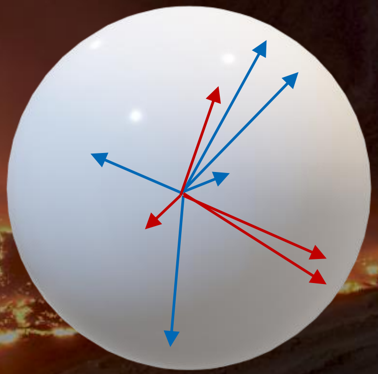
Open source with permissive licensing of all core components

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	☑	☑	☒	☒	☒	☒	☑	☑

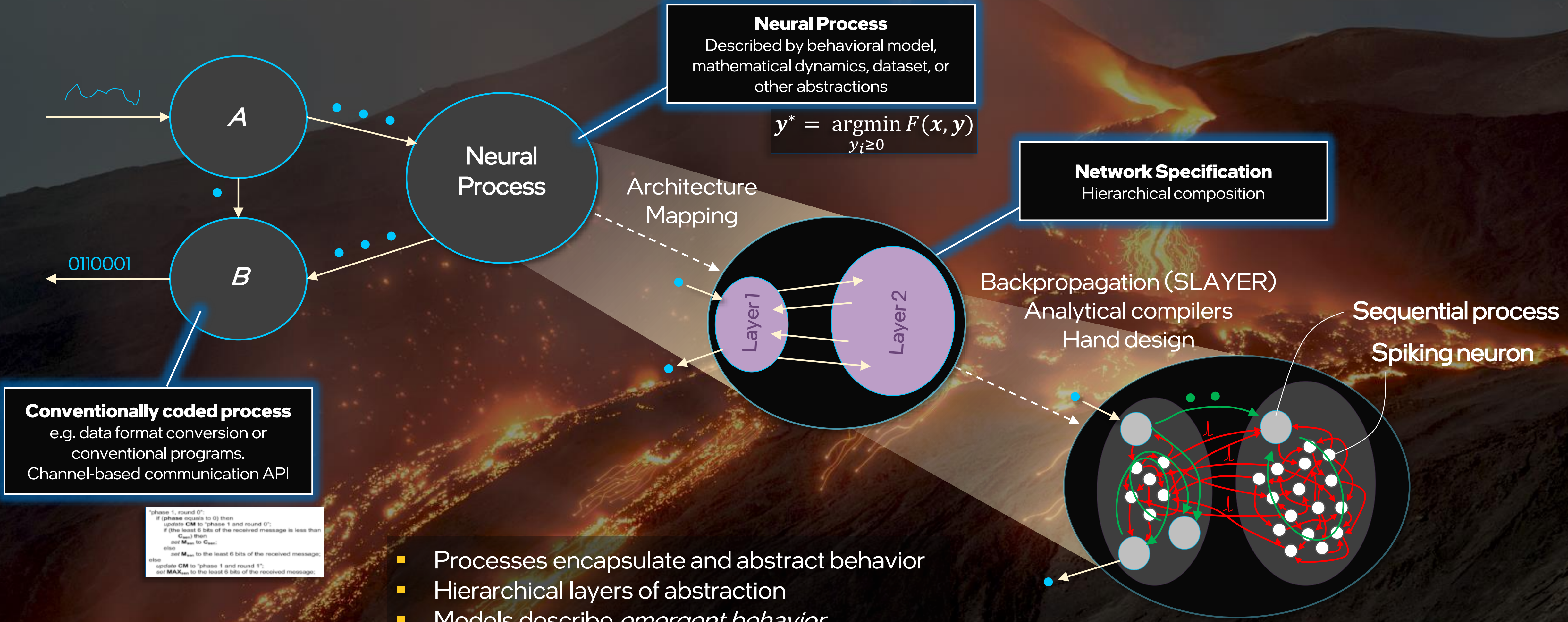
See https://github.com/lava-nc

Multi-Paradigm

Optimization	Neural Attractors	Deep Learning	Vector Symbolic
			
LCA, Stochastic SNNs LASSO, QP, CSP, ILP, QUBO	Dynamic Neural Fields, Continuous Attractor NNs, WTA	ANN->SNN rate-coded conversion, Directly trained SNN ConvNets Sigma-Delta Neural Networks TTFS- and Phase-coded SNNs	HRRs, MAPs, Sparse Block Codes, Associative Memories, Resonator Networks
+ model learning	+ associative learning	+ gradient learning	+ HD learning

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

Multi-Abstraction

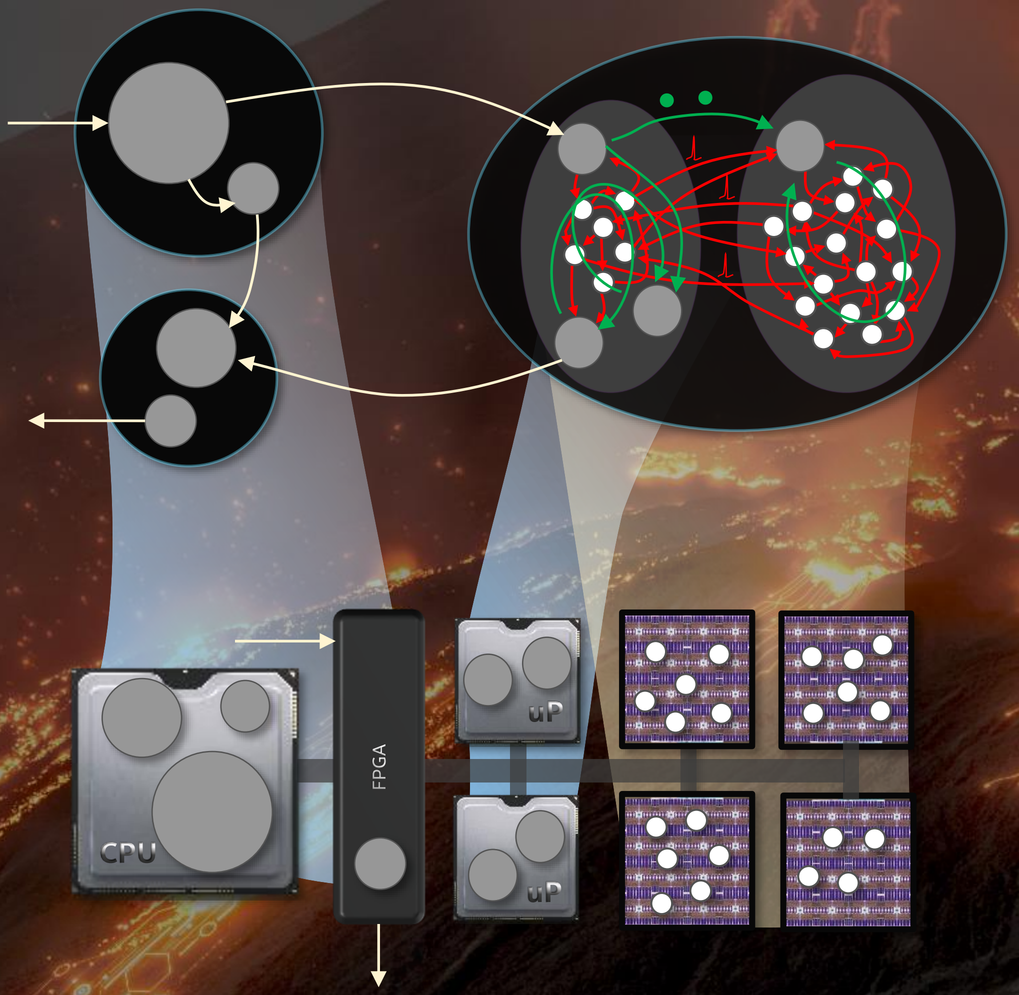


- Processes encapsulate and abstract behavior
- Hierarchical layers of abstraction
- Models describe *emergent behavior*

Multi-Platform

Abstraction Layer

Physical Layer



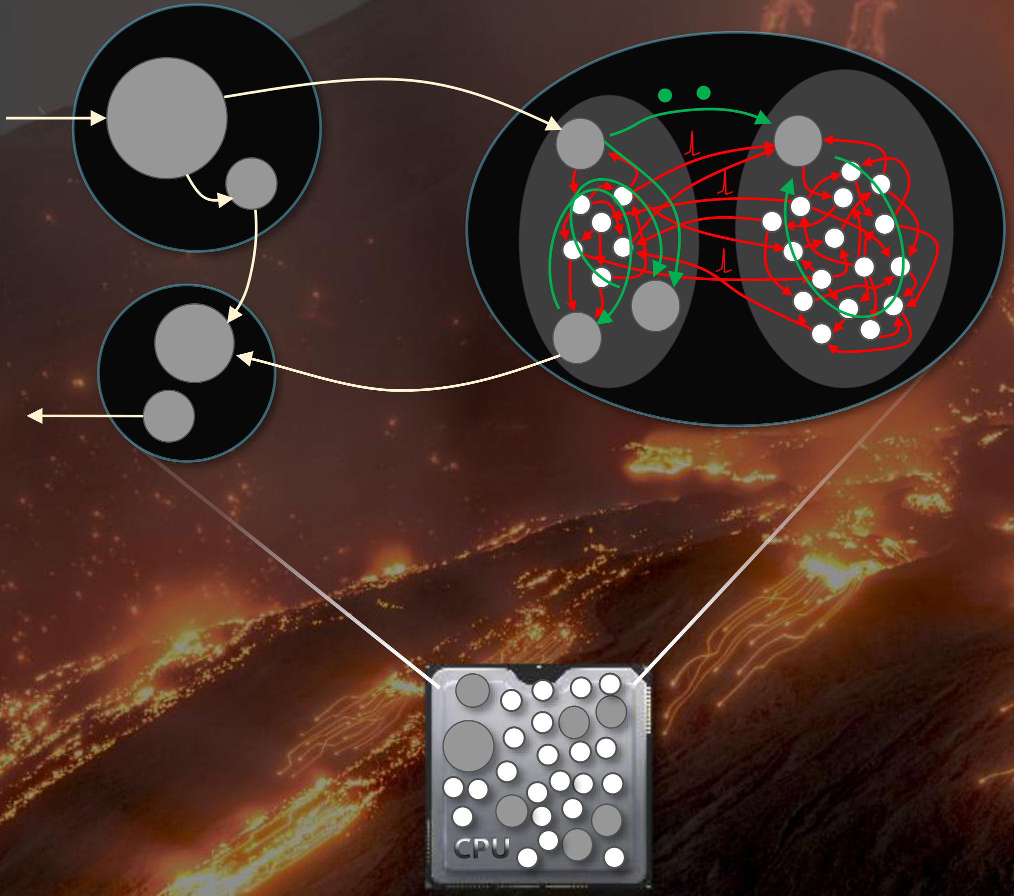
- Heterogeneous system architecture
- Multi-backend execution + profiling
- Fast compilation and execution
- Performant real-time operation

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

CPU-only Execution for Exploration and Prototyping

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...

Open and Extensible Software Stack

Lava Algo/App libraries BSD

Optimization | Deep Learning | DNF | VSA | ••

Lava Process library BSD

Future Third Party Interfaces

Nengo ROS PyNN Fugu BRIAN

Magma

Communicating Sequential Process API — BSD/Proprietary

Compiler — BSD/LGPL/Proprietary

Runtime — LGPL/Proprietary

Utilities

Profiler — BSD

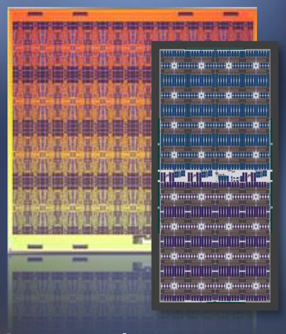
Visualizer — BSD

Float2Fixed — BSD

Validator — BSD

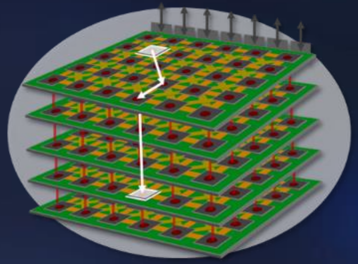
Outlook to Commercial Value

Today:



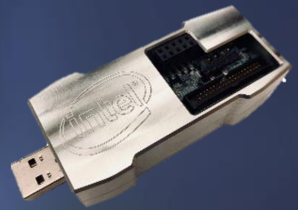
General-purpose research chips and software framework

Scaled up systems



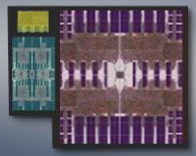
- Acceleration for datacenter optimization workloads
- Recommendation systems
- Scientific computing, HPC

Intelligent Extreme Edge Co-Processors

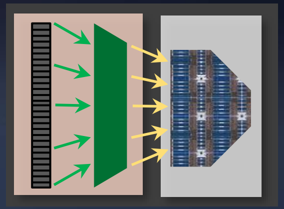


- Aerospace and robotics devices
- Scene awareness and localization
- Model predictive control
- Navigation and planning
- Consumer devices (longer term)

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



Get Involved!

Attend Tomorrow's Tutorial

Download Lava
https://github.com/lava-nc

Join the Intel Neuromorphic Research Community
e-mail inrc_interest@intel.com

Attend our Workshop April 19-22

Thank You!

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

Loihi 2 Performance Analysis Details

[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.

[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.

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.