Making Computing More Brain-Like

New Tools for a New Era of Neuromorphic Computing

Mike Davies, Garrick Orchard
Neuromorphic Computing Lab

August 4, 2022

Flash Memory Summit – Neuromorphic Computing Session

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.

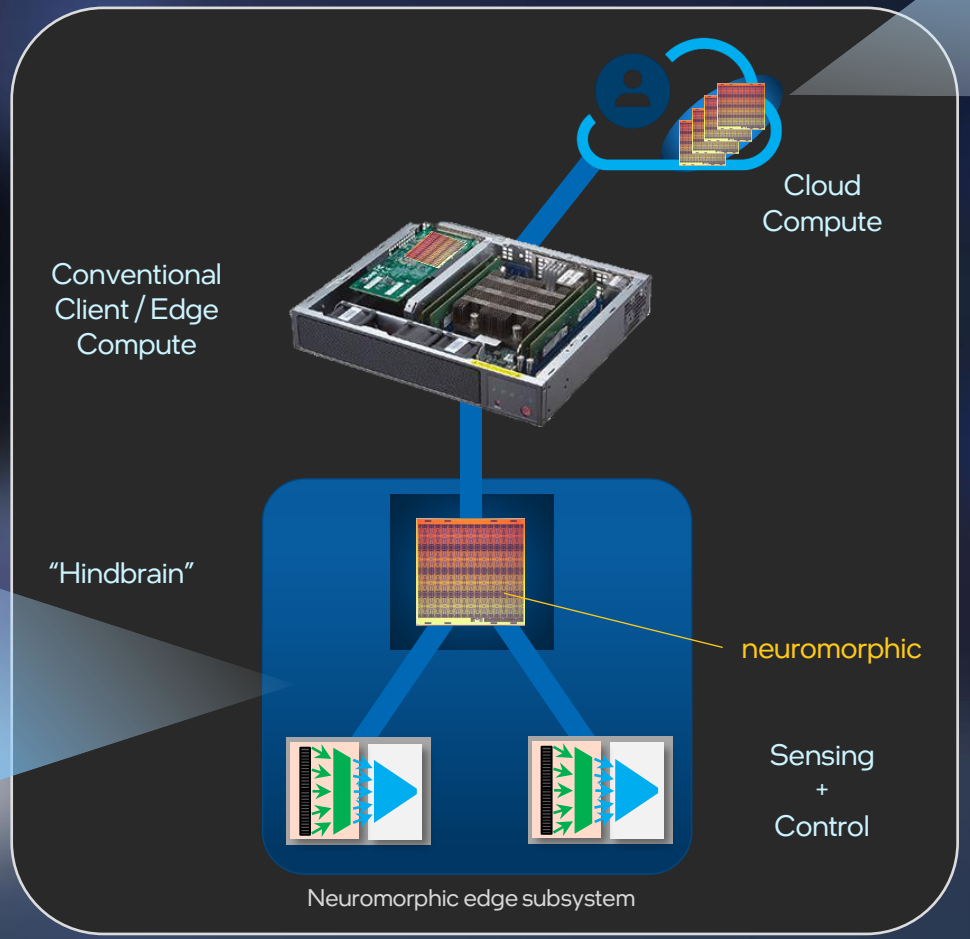
Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

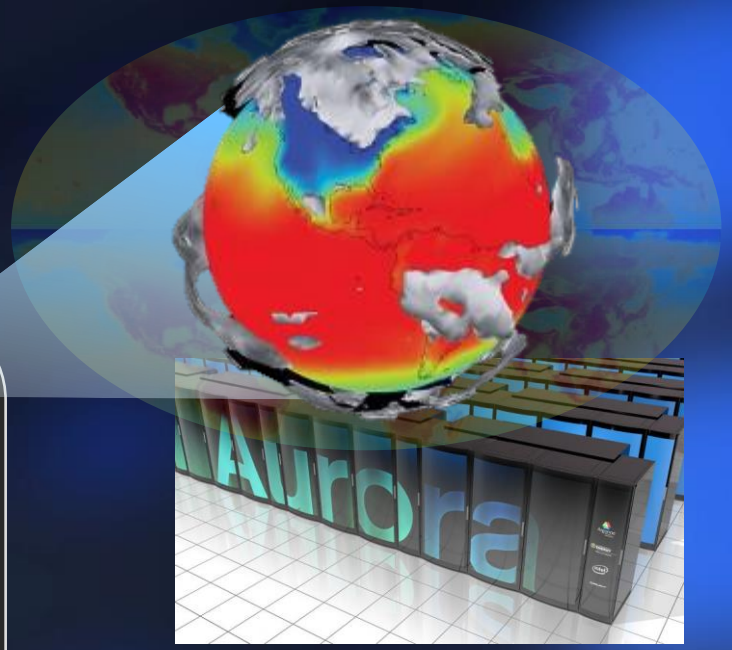
Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

Research Vision

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

Integrate neuromorphic intelligence into computing products at all scales



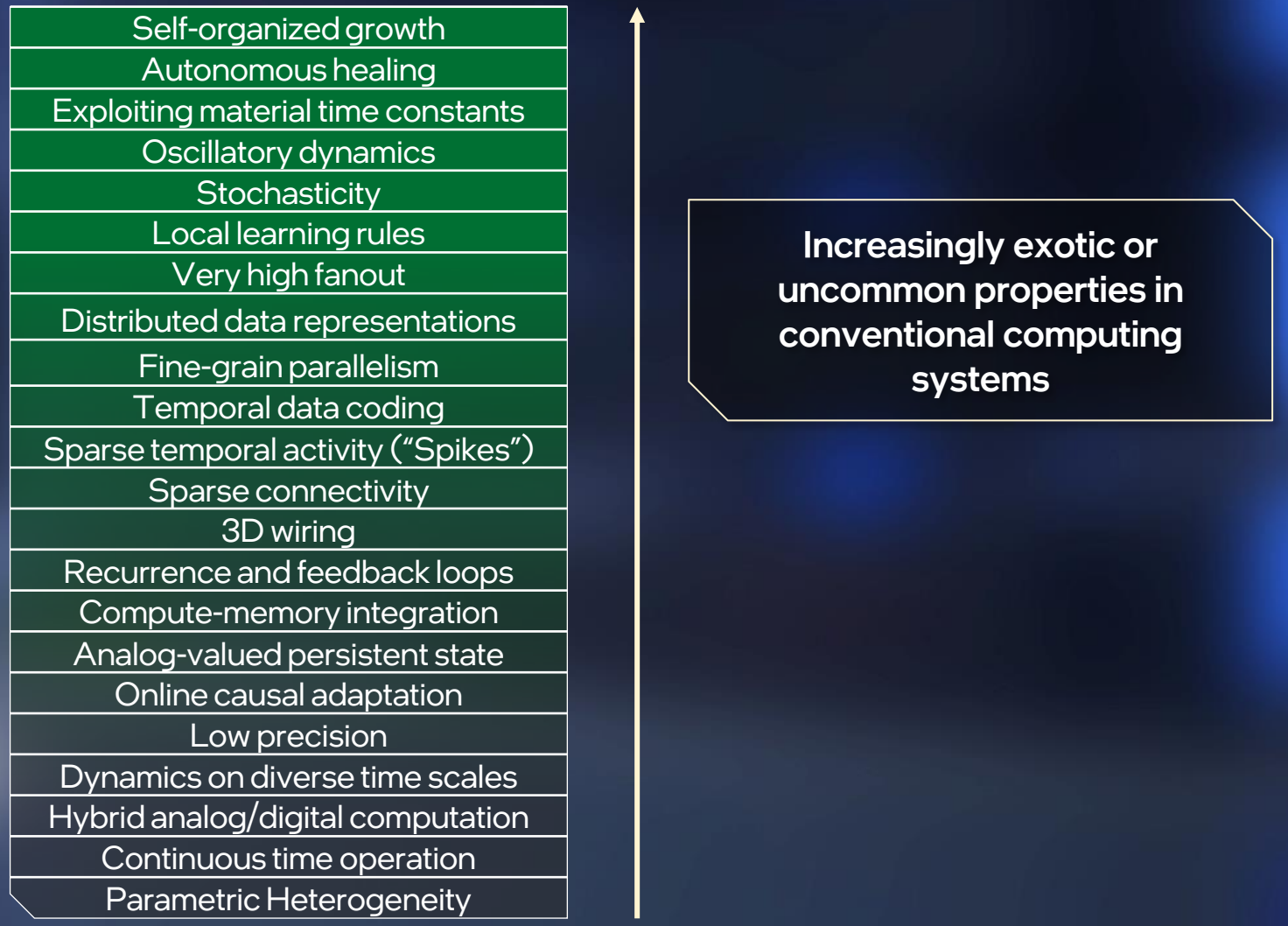


Achieve brain-like efficiency, speed, adaptability, and intelligence

Deliver gains of 10^4 or higher in energy-delay-product*

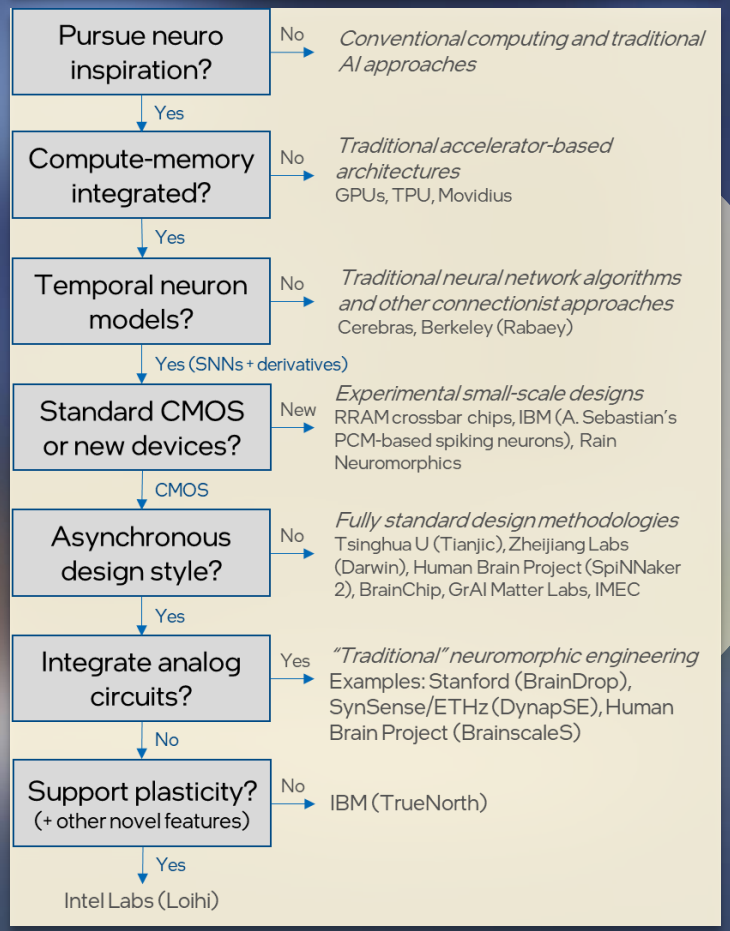
* Combined latency and energy efficiency metric

The brain offers a huge space of design exploration



Calls for iterative architecture-algorithms co-design

Neuro-Inspired Silicon



Rigorous Benchmarking

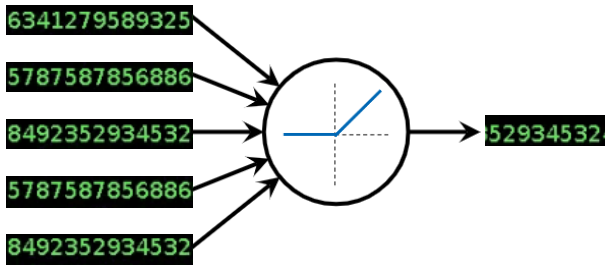
Novel Neuro-Inspired Algorithms

Paradigm	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

Dynamics at the neuron level

Artificial Neuron (Stateless)

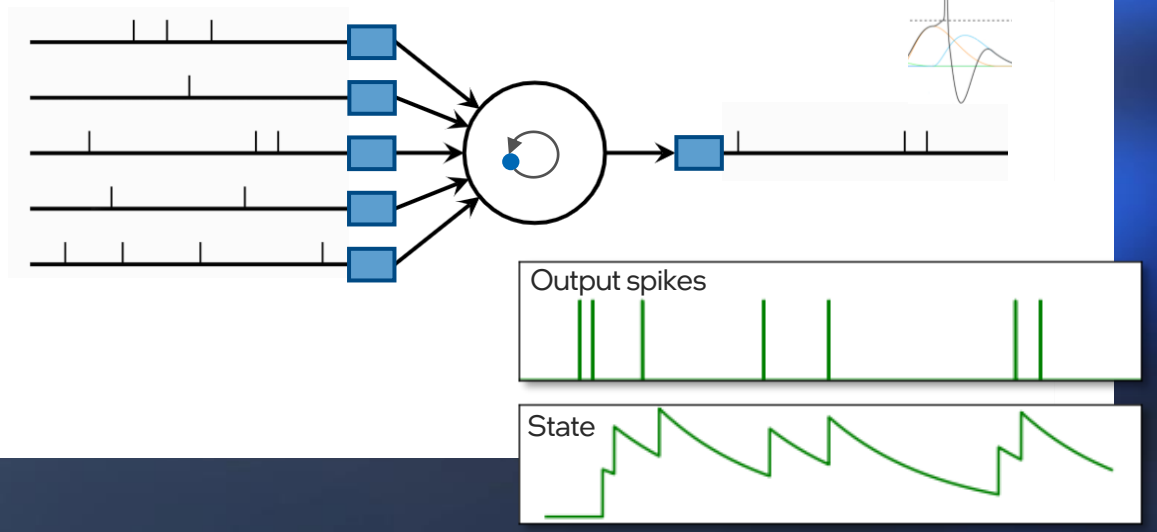
$$u_i = \sum_j w_{ij} f(u_j) + b_i$$



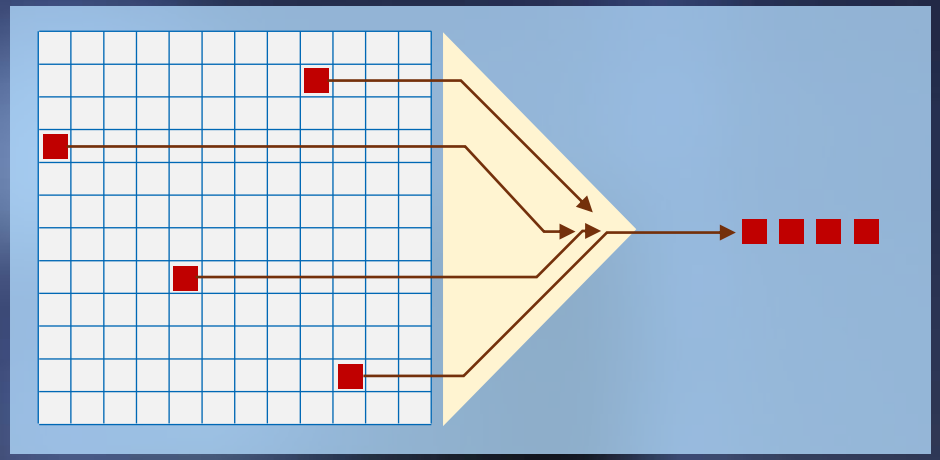
Spiking Neuron (Nonlinear Filter)

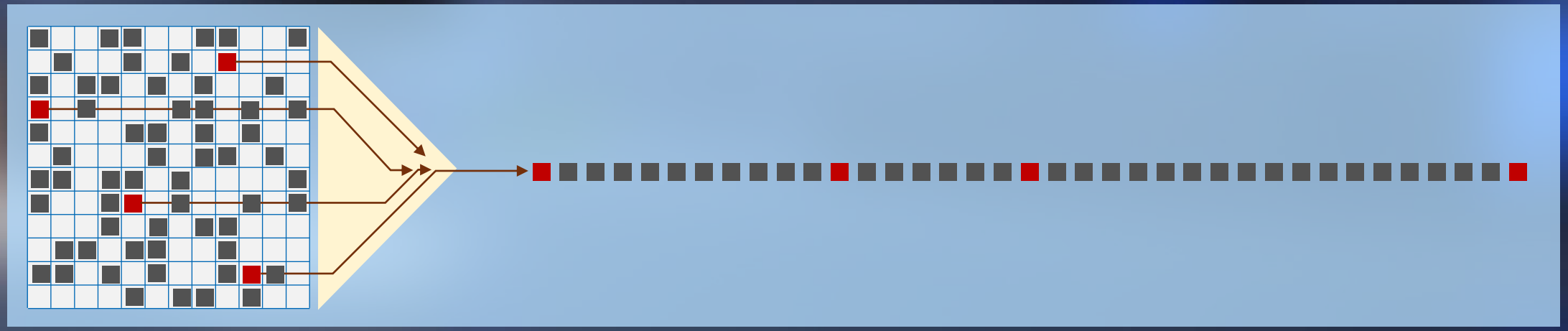
$$u_i(t) = \sum_j w_{ij} \left(\delta_j(t) * \alpha_u(t) \right) + b_i$$

$$\tau \dot{v}_i(t) = (-v_i(t) + u_i(t)) - V_{thr} \delta_i(t)$$

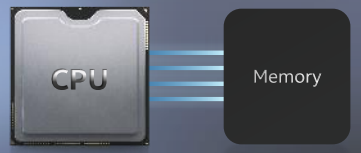
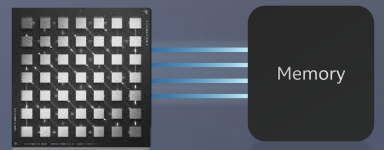
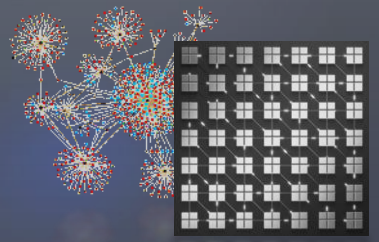
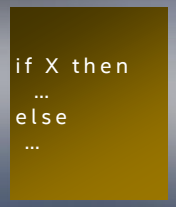
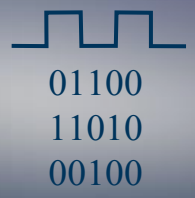
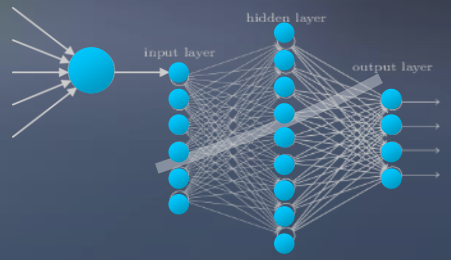
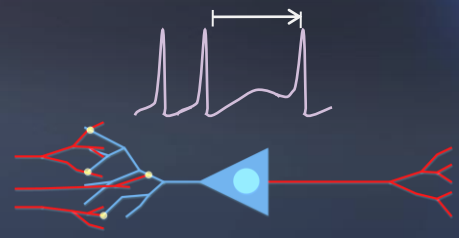


Sparse, asynchronous communication is fast + efficient





Leads us to a new class of computer architecture

Standard Computing	Parallel Computing	Neuromorphic Computing
		
PROGRAMMING BY ENCODING ALGORITHMS	OFFLINE TRAINING USING LABELED DATASETS	LEARN ON THE FLY THROUGH NEURON FIRING RULES
SYNCHRONOUS CLOCKING	SYNCHRONOUS CLOCKING	ASYNCHRONOUS EVENT-BASED SPIKES
SEQUENTIAL THREADS OF CONTROL	PARALLEL DENSE COMPUTE	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

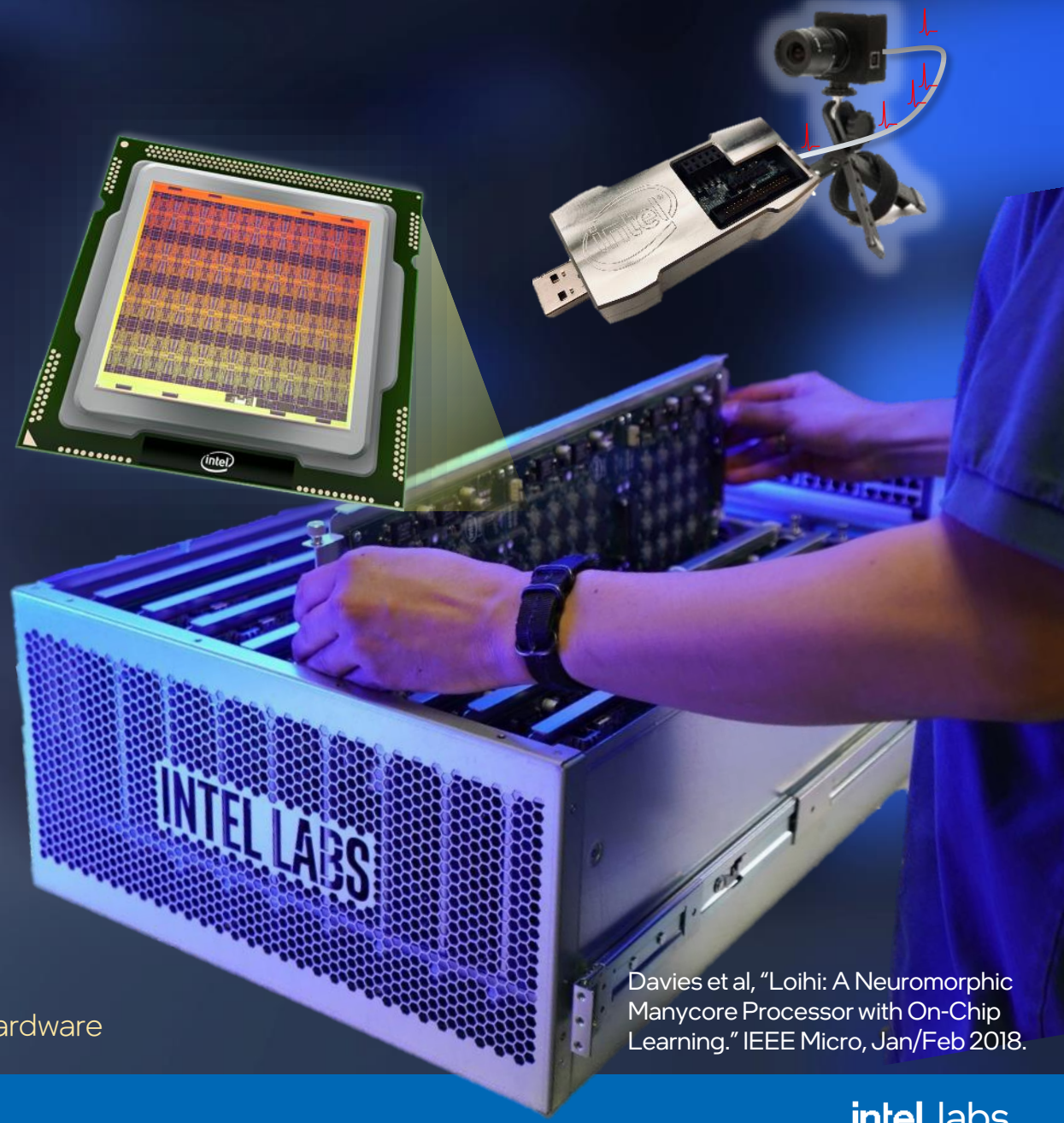
On-chip learning
without weight movement or data storage

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



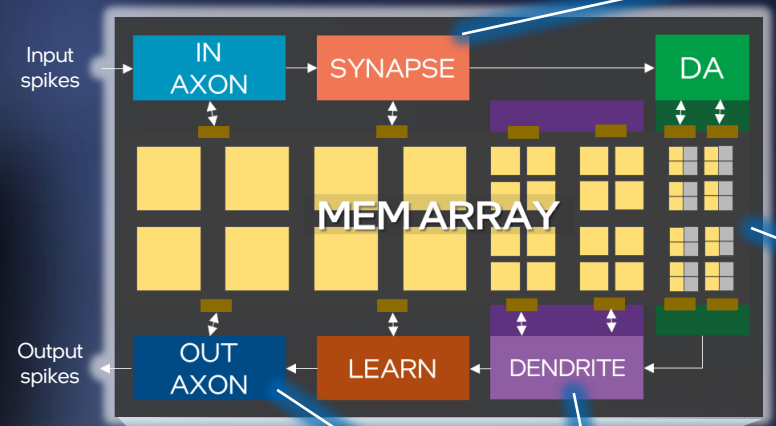
Davies et al, "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning." IEEE Micro, Jan/Feb 2018.

Second generation Loihi neuromorphic core

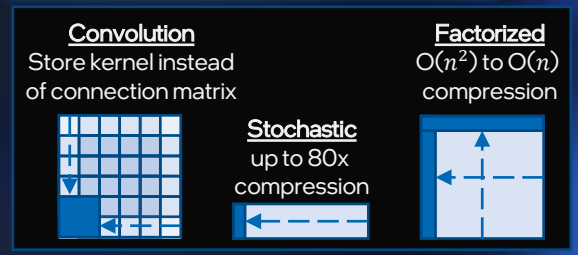
Generalized Spikes
Spikes carry int8 magnitudes for greater workload precision

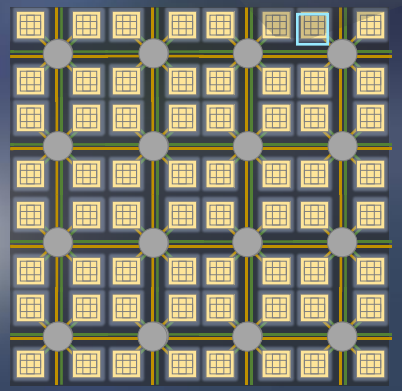
Programmable Neurons
Neuron models described by microcode instructions

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

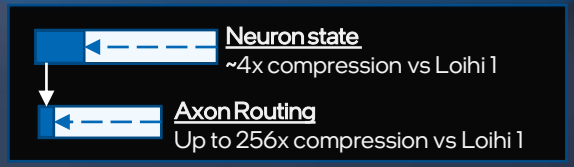


Better Synaptic Compression

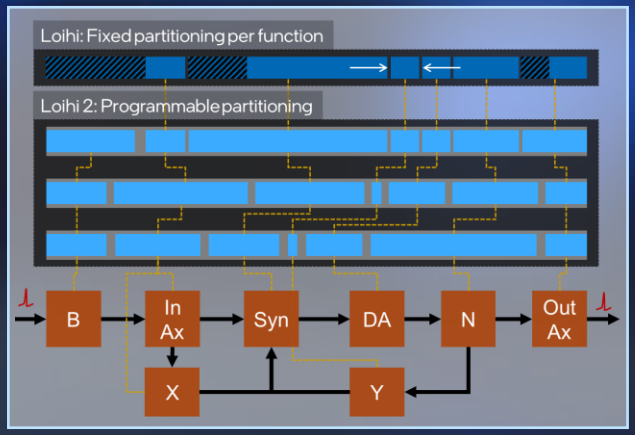




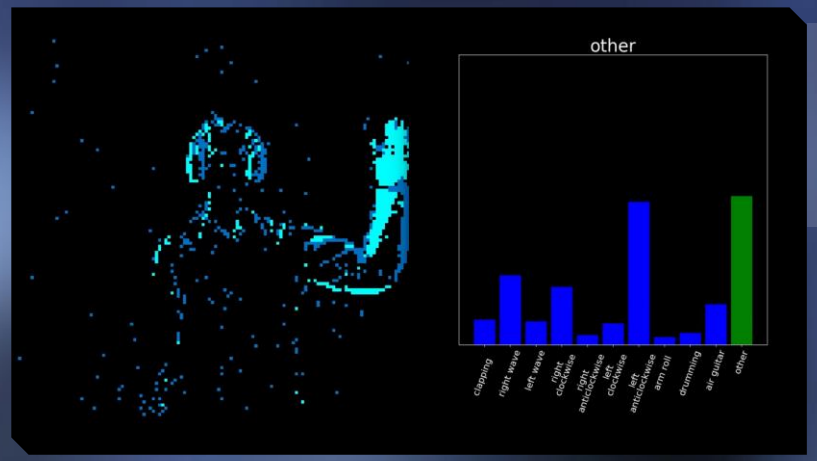
Better Neuron and Routing State Compression



Better Utilization of Core Memory
Highly ported centralized async memory array provides resource allocation flexibility



Loihi Has Confirmed the Value of This Direction

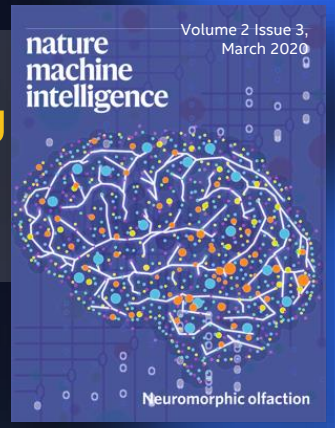


Gesture recognition + learning

Loihi + DAVIS 240C camera
60 mW total power, 15 mW dynamic

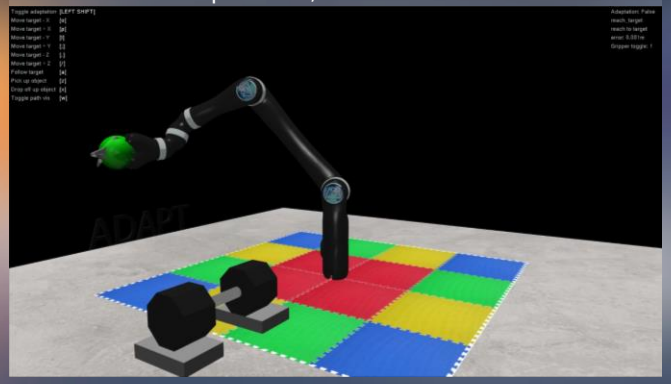
Olfaction-inspired odor recognition and learning

3000x more data efficient learning than a deep autoencoder



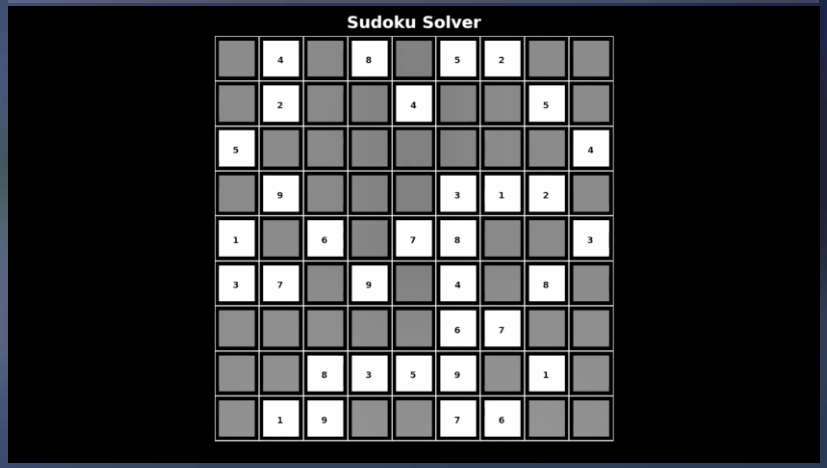
Adaptive robotic arm control

40x lower power, 50% faster vs GPU



Combinatorial optimization (CSP, SAT, ILP, QP)

2,800x lower energy and 44x faster vs CPU



Scene understanding

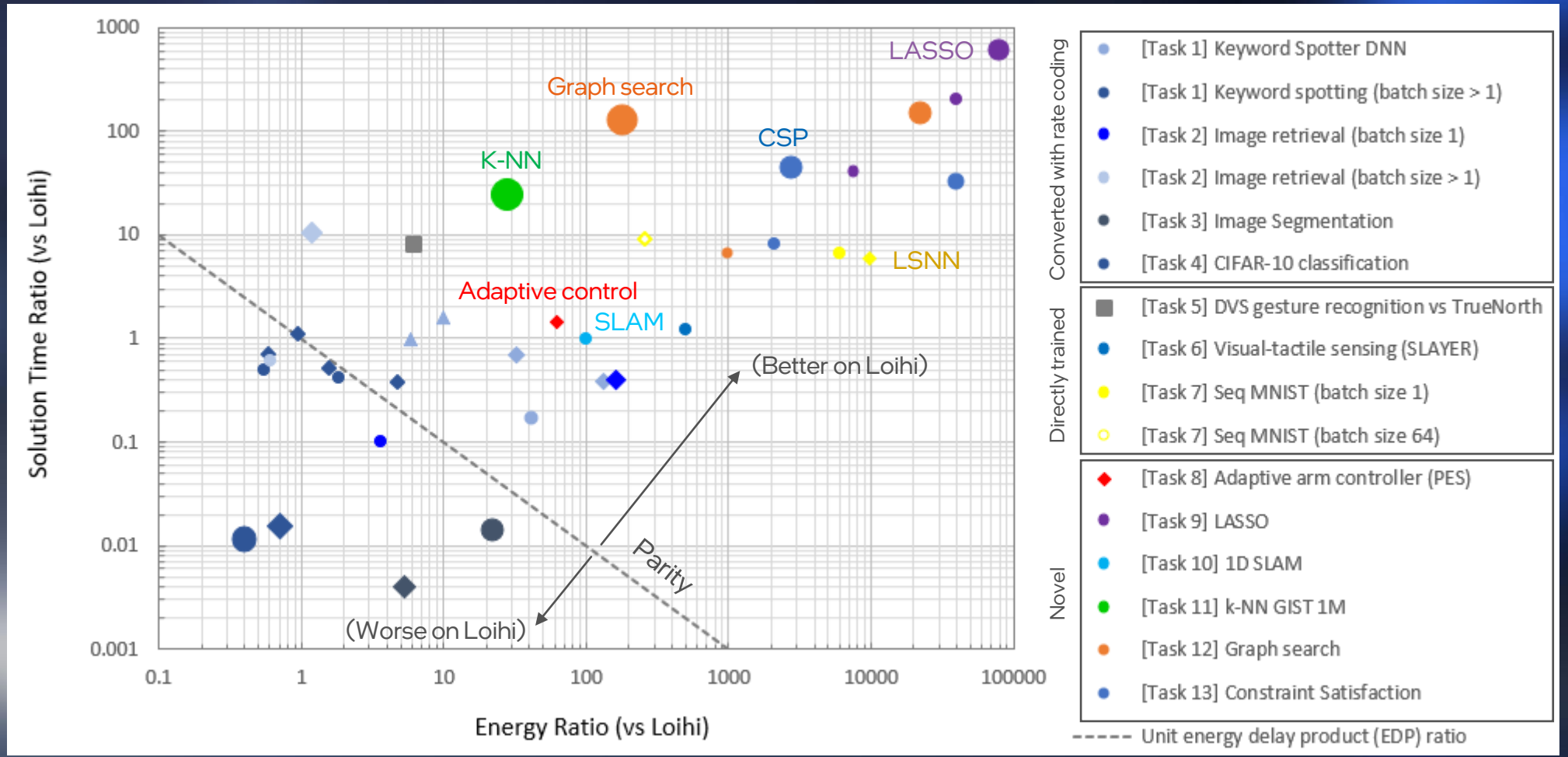
Integrated behaviors: Object recognition, tracking, learning
100x lower power SLAM vs CPU

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

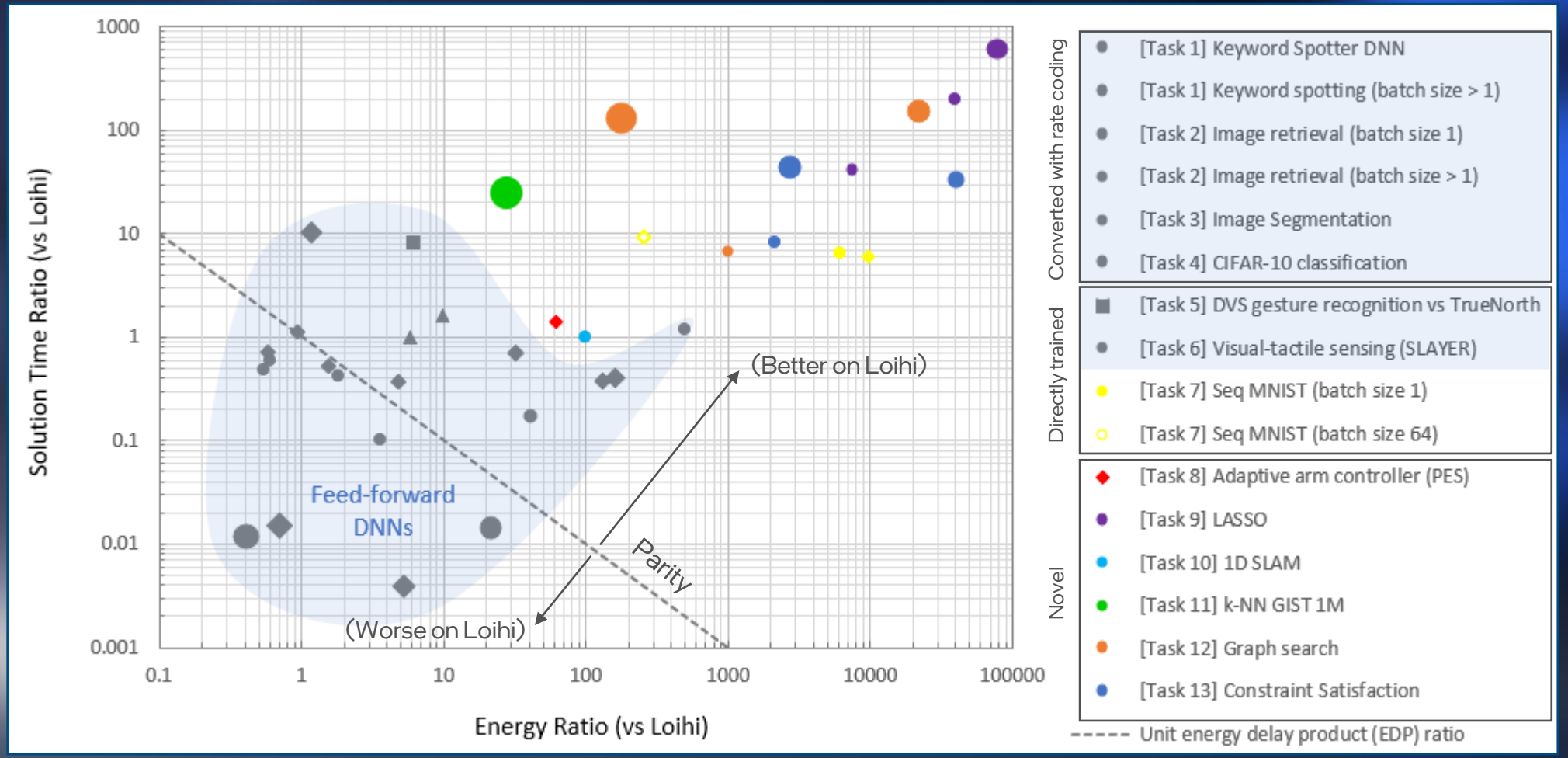


M. Davies et al, "Advancing Neuromorphic Computing With Loihi: A Survey of Results and Outlook," Proc. IEEE, 2021. Results may vary.

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

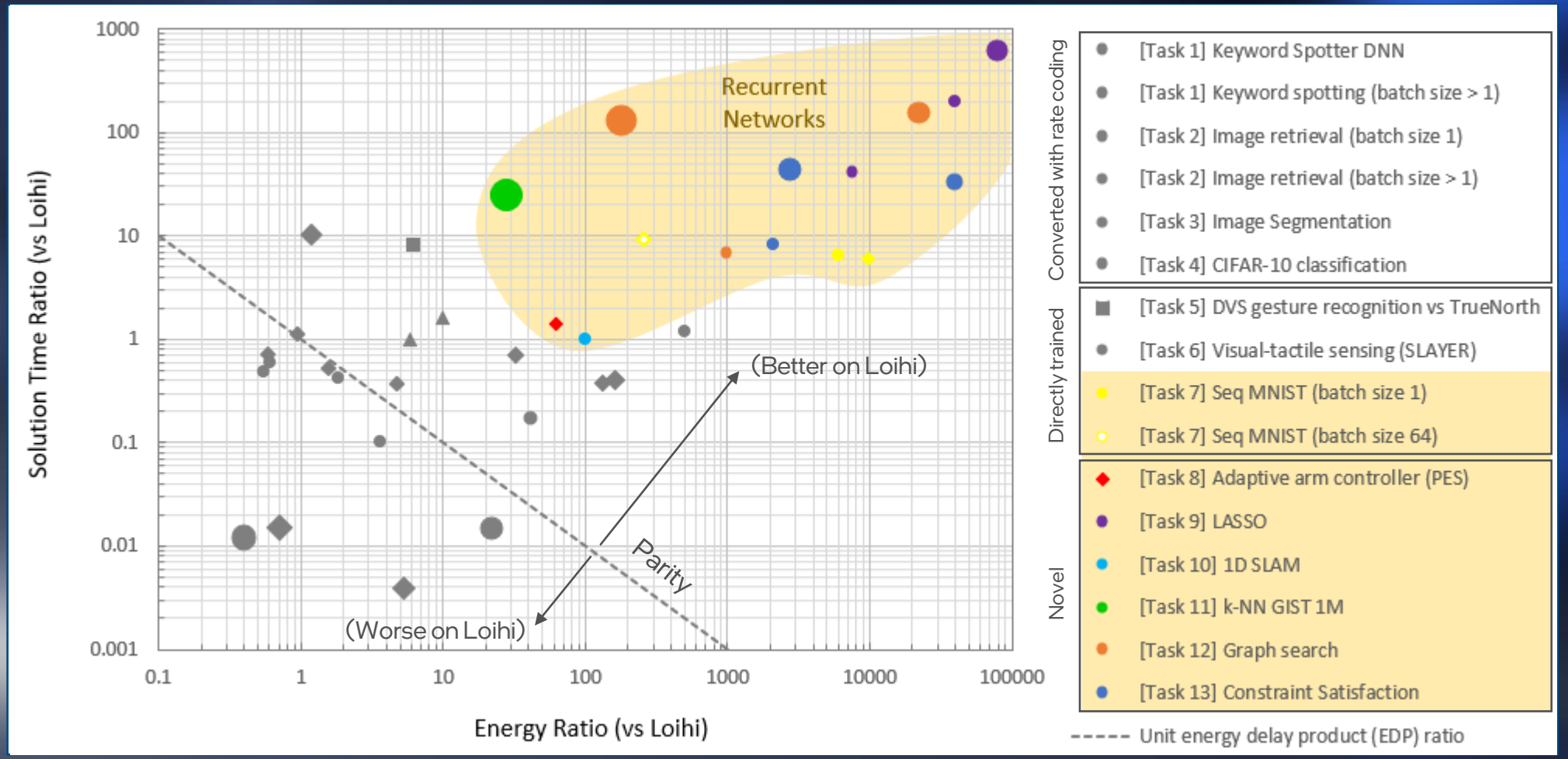


M. Davies et al, "Advancing Neuromorphic Computing With Loihi: A Survey of Results and Outlook," Proc. IEEE, 2021. Results may vary.

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

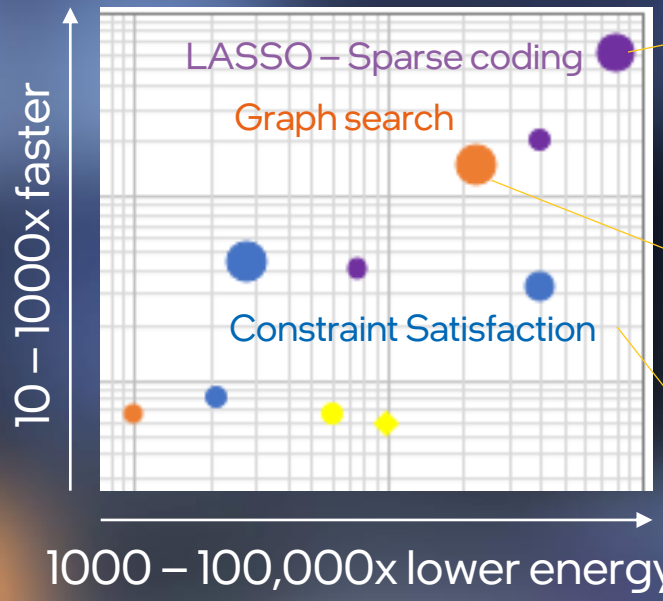
Reference architecture

- 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.

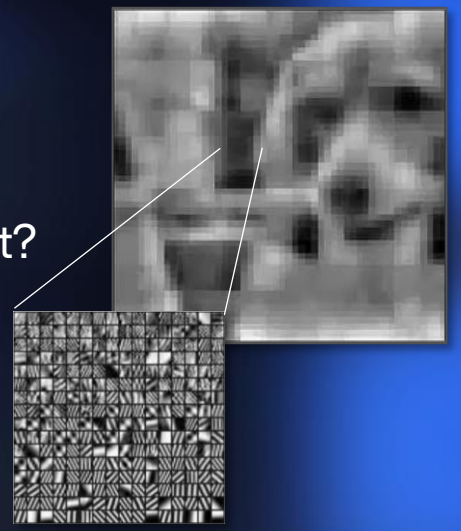
Zooming in on the best examples: Optimization problems



What features best explain the sensory input?

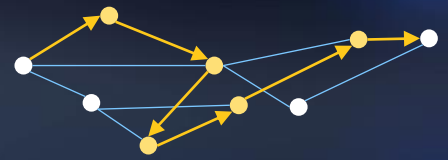
$$\underset{z}{\operatorname{argmin}} \|x - Dz\|_2^2 + \lambda \|z\|_1$$

Input Reconstruction Sparse regularization



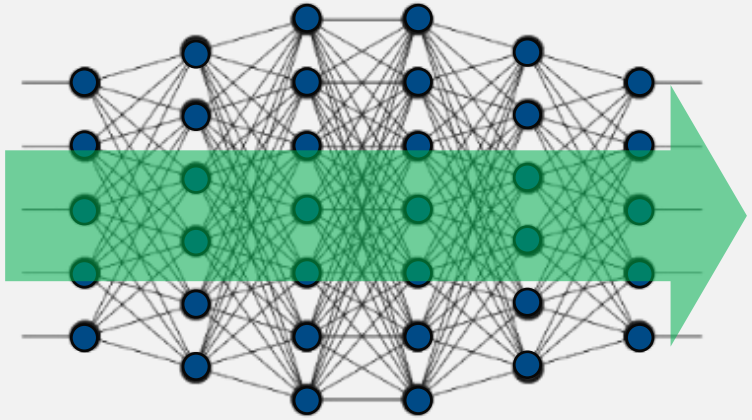
What is the shortest path to my goal?

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



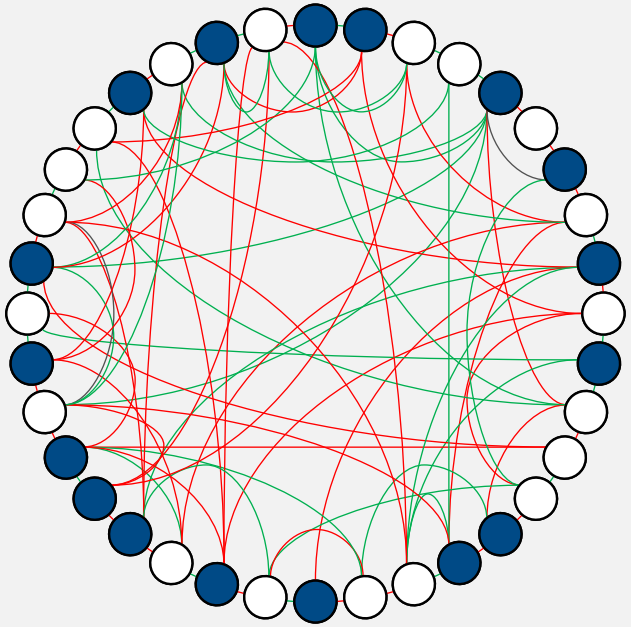
Key insight: Neuromorphic networks efficiently optimize solutions via stochastic gradient descent

Conventional deep neural networks



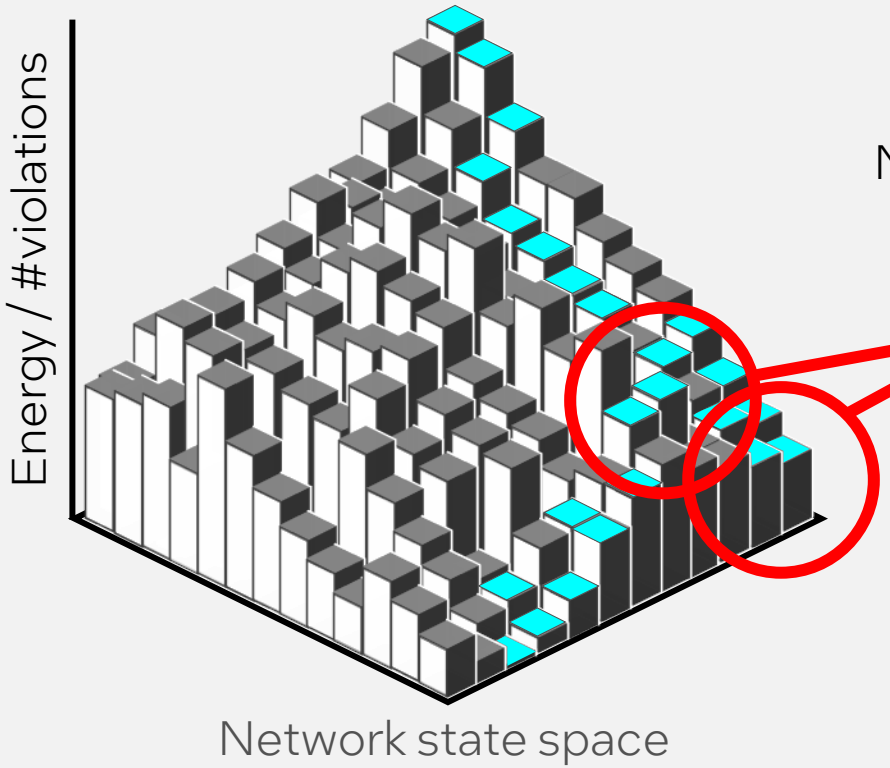
Single input produces single inference result

Neuromorphic networks



Network continually visits different candidate solution states

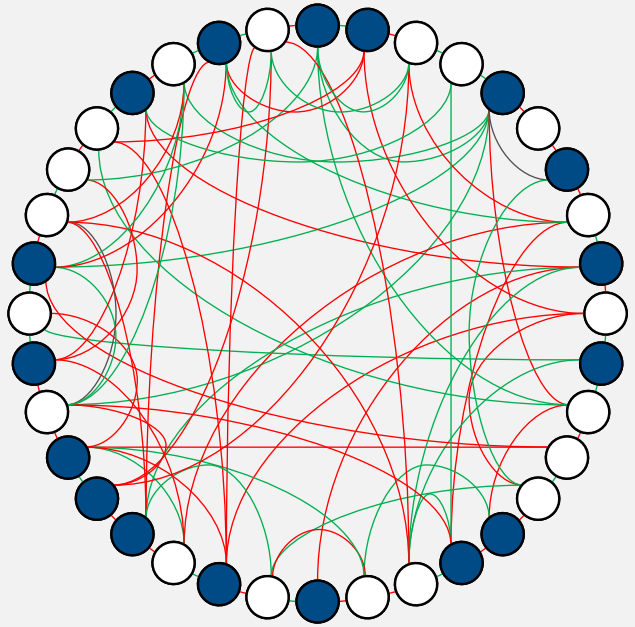
Key insight: Neuromorphic networks efficiently optimize solutions via stochastic gradient descent



Neural dynamics descend the gradient

Local minimum escaped by stochastic spiking dynamics

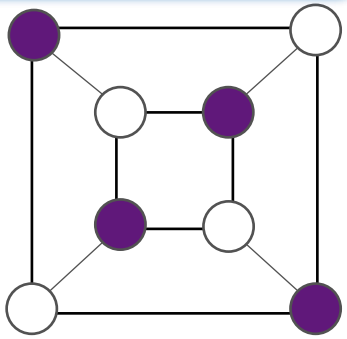
Efficient descent due to massively parallel, asynchronous neuromorphic computing architecture



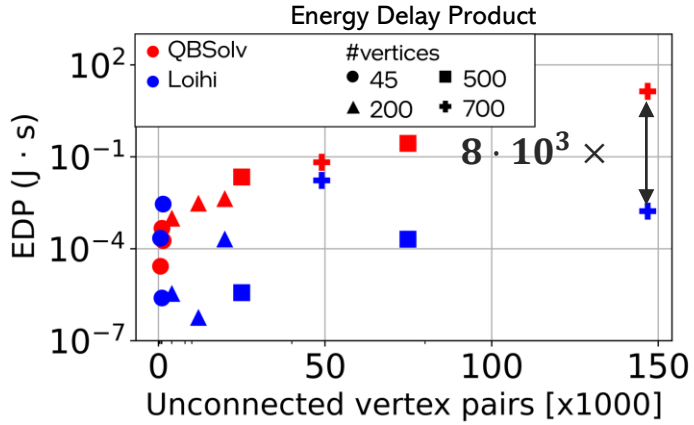
Loihi outperforms leading optimization solvers by orders of magnitude

QUBO (Maximum Independent Set)

Workload:
Find largest set of unconnected vertices



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

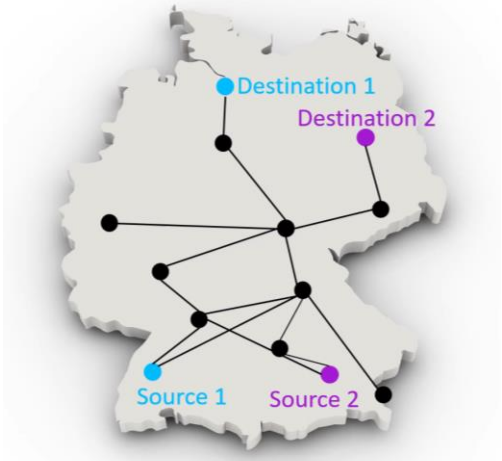


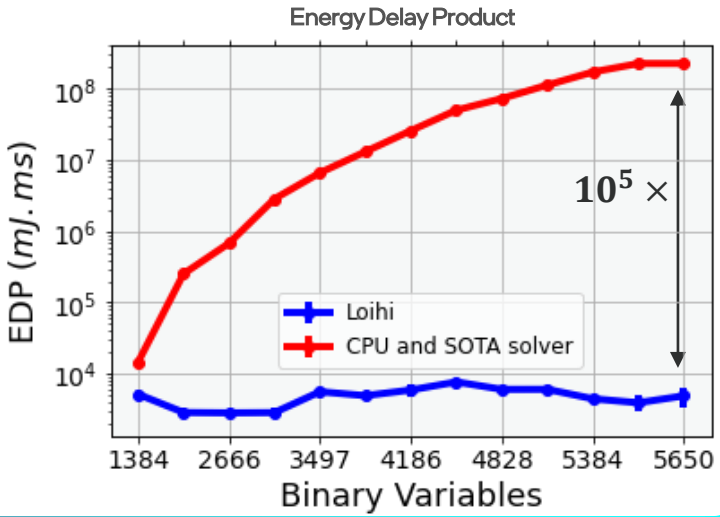
Integer Linear Programing (Train Scheduling)

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: 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.

Into a New Era of Neuromorphic Computing

Proven computational value
(using today's manufacturing tech)

Motivates a new computational paradigm
(cheap, continuous optimization)

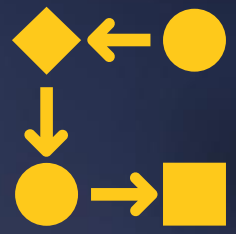
Many successful learning algorithms
(albeit shallow so far, not deep)

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 trainingc

Challenges and headwinds

High cost due to on-chip memory integration

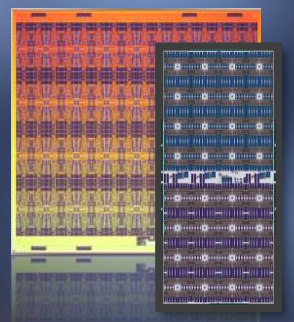


Algorithms and Programming models

Software convergence

Outlook to Commercial Value

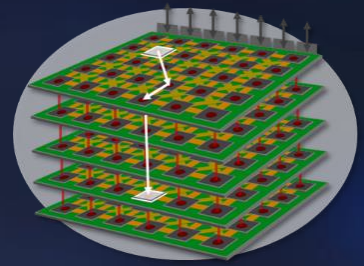
Today:

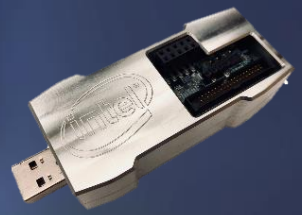


General-purpose research chips and software framework

Scaled up systems

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

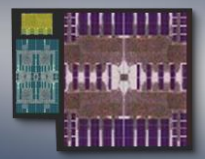




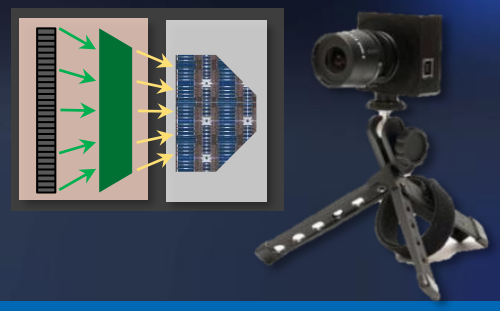
Intelligent Extreme Edge Co-Processors

- 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



Priorities for process technology innovation

- Memory capacity and integration
  - Increased on-chip memory density
  - 3D integration of memory and compute
- Low voltage, low leakage transistors
  - Reduce static power in the presence of sparsity
- Optical interconnect
  - Fast communication and synchronization across a chip or many chips
- Analog synaptic state
  - Exploit physical/analog device properties (time constants) to model time varying neuron dynamics
  - Asynchronous ADC/DAC to perturb or readout the current state

Thank You!

Email inrc_interest@intel.com for more information

Visit https://github.com/lava-nc to get started with Lava

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.

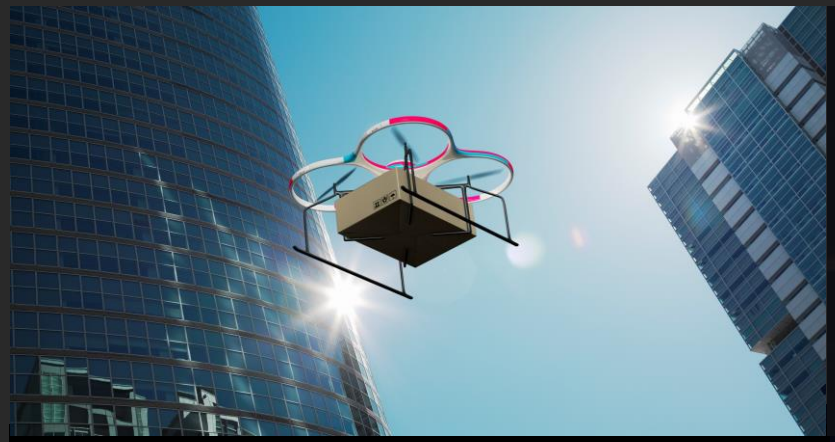
[5] Based on analysis of 3-chip and 7-chip Locally Competitive Algorithm examples.

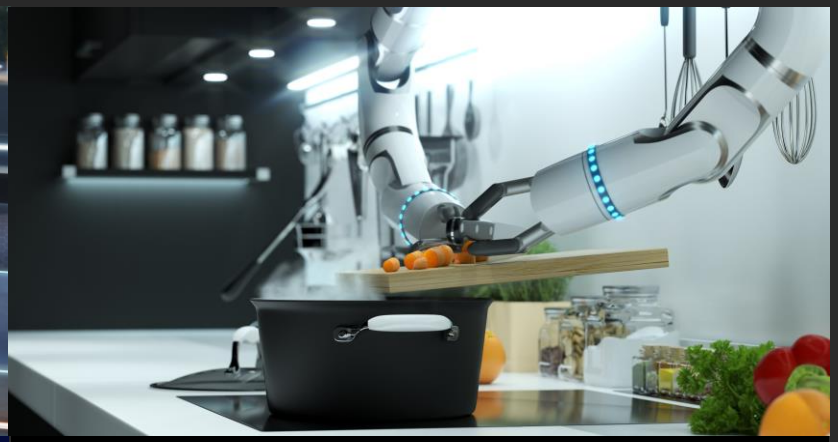
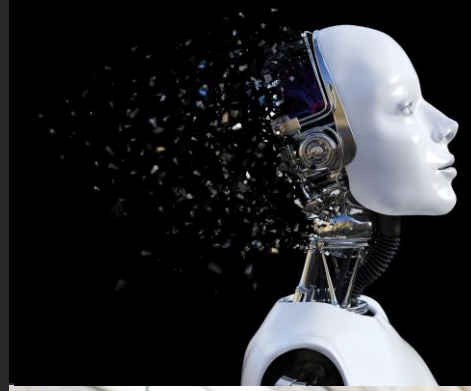
[6] Loihi 1 measurements were obtained on Oheo Gulch FMC board ncl-og-06 using an internal version of NxSDK advanced from v1.0.0

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

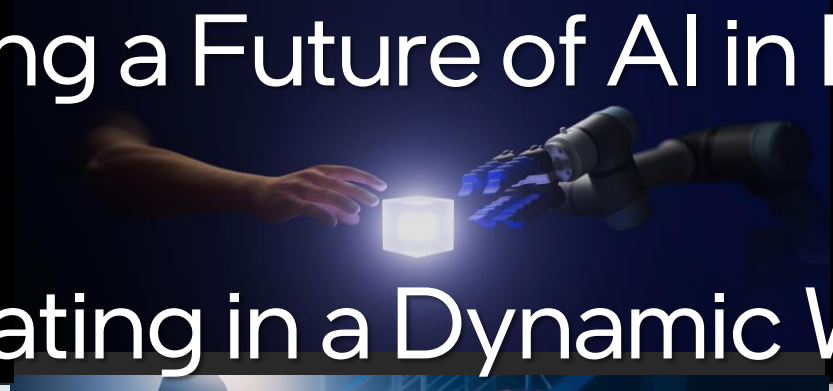
The Lava performance model for both chips is based on silicon characterization in September 2021 using the Nx SDK release 1.0.0 with an Intel Xeon E5-2699 v3 CPU @ 2.30 GHz, 32GB RAM, as the host running Ubuntu version 20.04.2. Loihi results use Nahuku-32 system ncl-ghrd-04. Loihi 2 results use Oheo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

Results may vary.



Enabling a Future of AI in Motion




Operating in a Dynamic World