Low-power, low-latency computing with Loihi 2

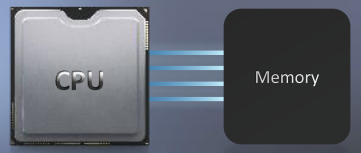
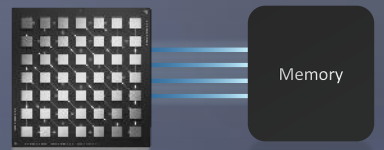
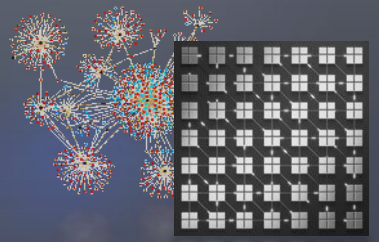
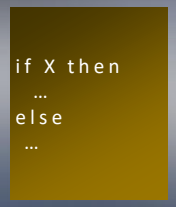
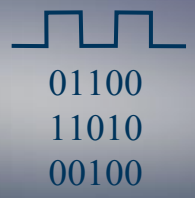
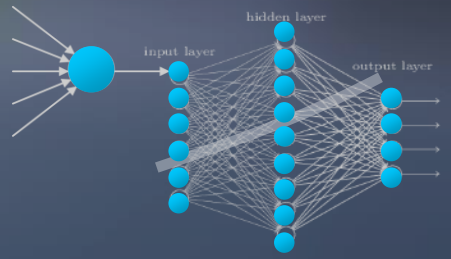
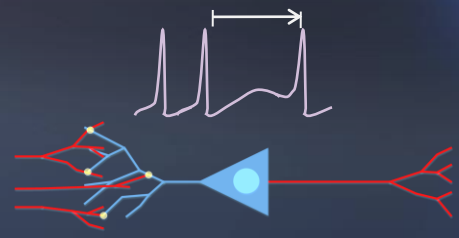
Ashish Rao Mangalore
ashish.rao.mangalore@intel.com

22nd November 2023

Agenda

- Introduction to the Intel Neuromorphic Research Chip, Loihi
- Where Loihi shines
- Current focus areas for research

A new class of computer architecture

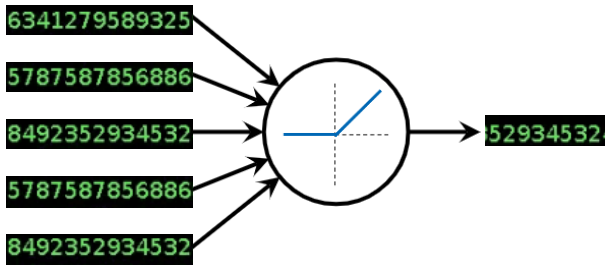
Standard Computing	Parallel Computing	Neuromorphic Computing
		
PROGRAMMING BY ENCODING ALGORITHMS	OFFLINE TRAINING USING LABELED DATASETS	LEARN ON THE FLY THROUGH NEURON FIRING RULES
SYNCHRONOUS CLOCKING	SYNCHRONOUS CLOCKING	ASYNCHRONOUS EVENT-BASED SPIKES
SEQUENTIAL THREADS OF CONTROL	PARALLEL DENSE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

Exploiting dynamics at the neuron level

Maximize computation by minimizing data movement

Artificial Neuron (Stateless)

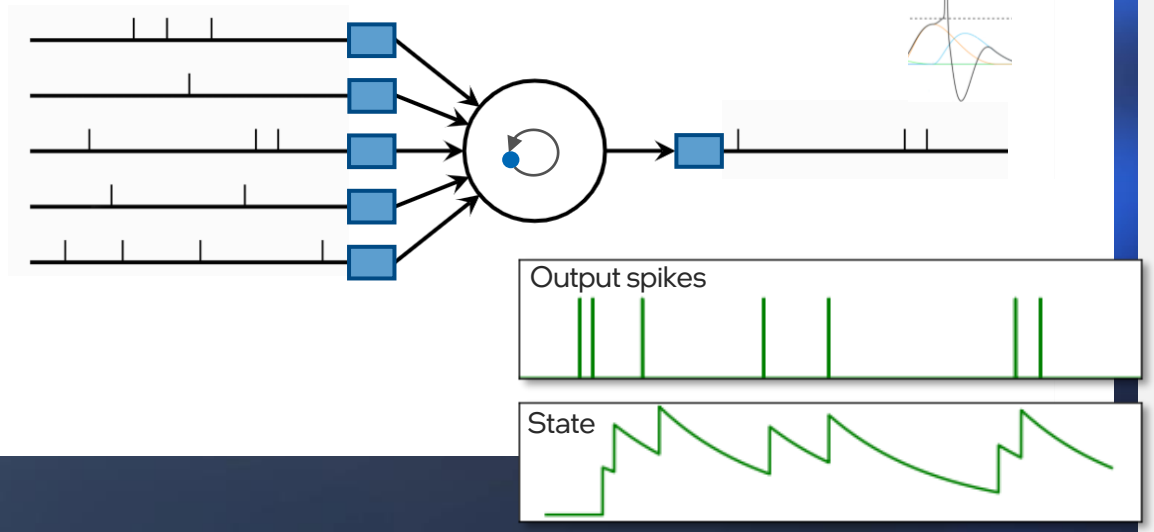
$$u_i = \sum_j w_{ij} f(u_j) + b_i$$



Spiking Neuron (Nonlinear Filter)

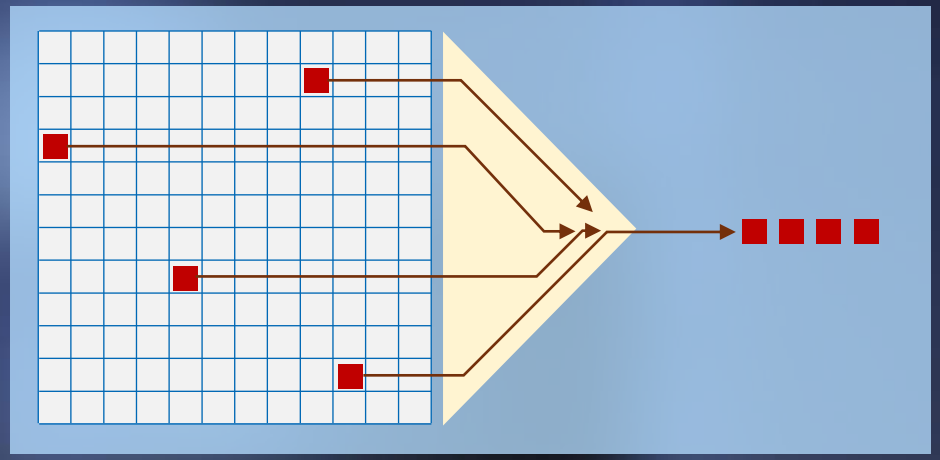
$$u_i(t) = \sum_j w_{ij} \left(\delta_j(t) * \alpha_u(t)\right) + b_i$$

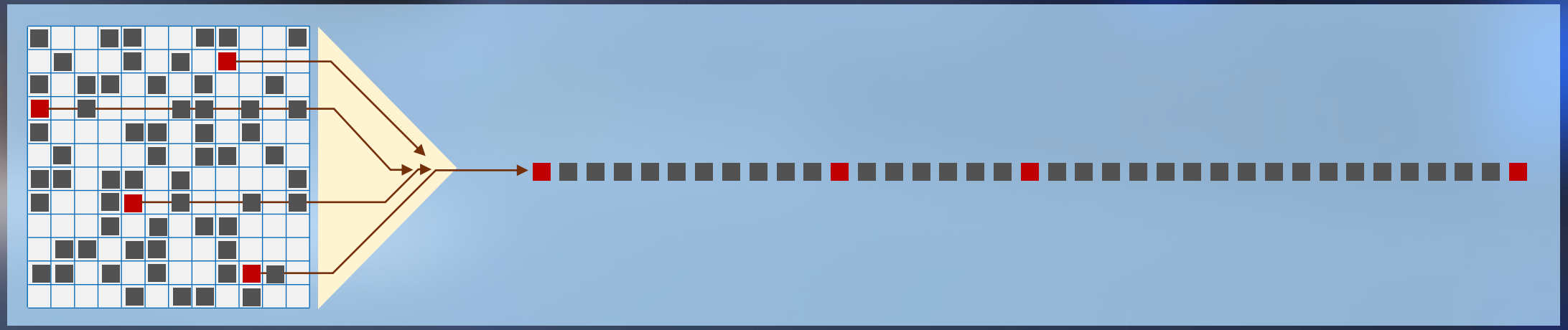
$$\tau \dot{v}_i(t) = (-v_i(t) + u_i(t)) - V_{thr}\delta_i(t)$$



Exploiting sparse, asynchronous communication

Fast and efficient, whether in brains or in computers



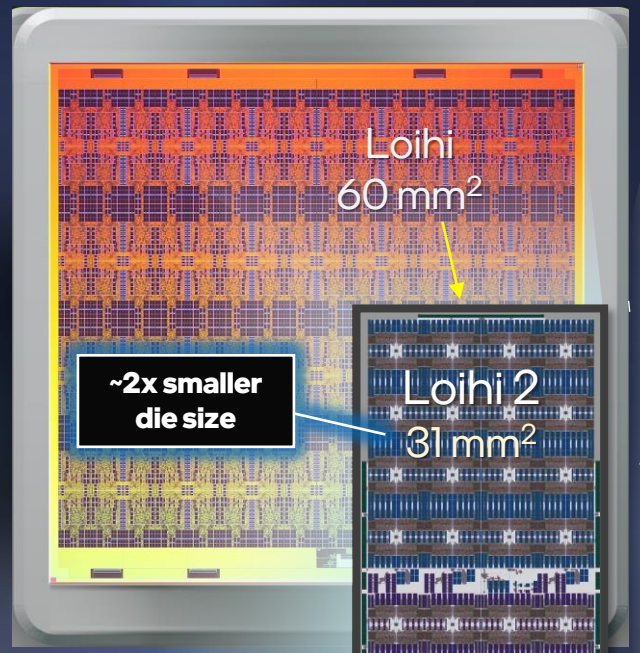


The Latest Loihi chip: Loihi 2

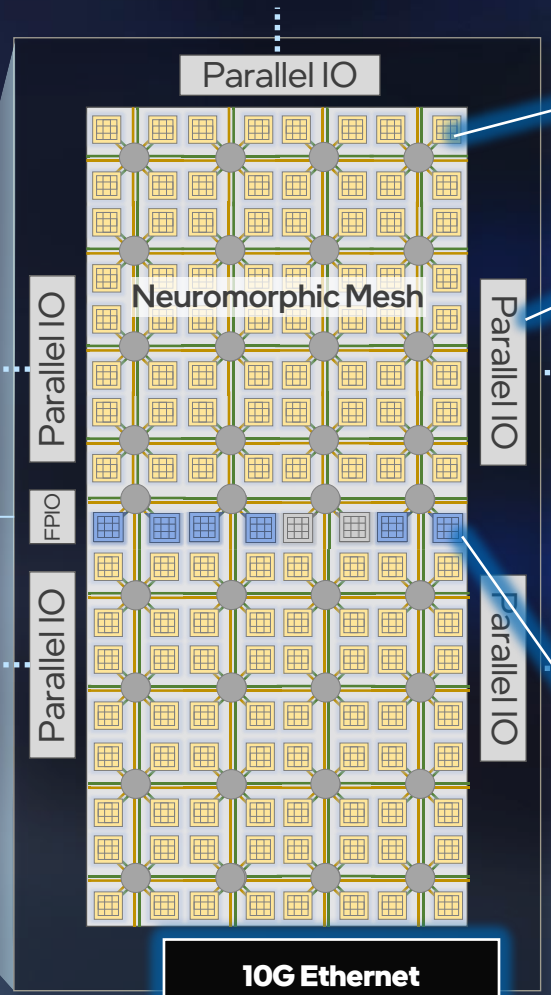
Generalized Spikes
Spikes carry int8 magnitudes for greater workload precision

Programmable Neurons
Neuron models described by microcode instructions

Programmable Neurons
Neuron models described by microcode instructions



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



Neuromorphic cores (128)
Fully redesigned with up to 8192 neurons

Parallel off-chip interfaces (6)
Async wave pipelined at 10 Gb/s with multicast compression

Microprocessor cores (6)
Asynchronous x86 and RISC-V

10G Ethernet
Accelerated host message + spike I/O

Realized in Loihi, improved in Loihi 2

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.

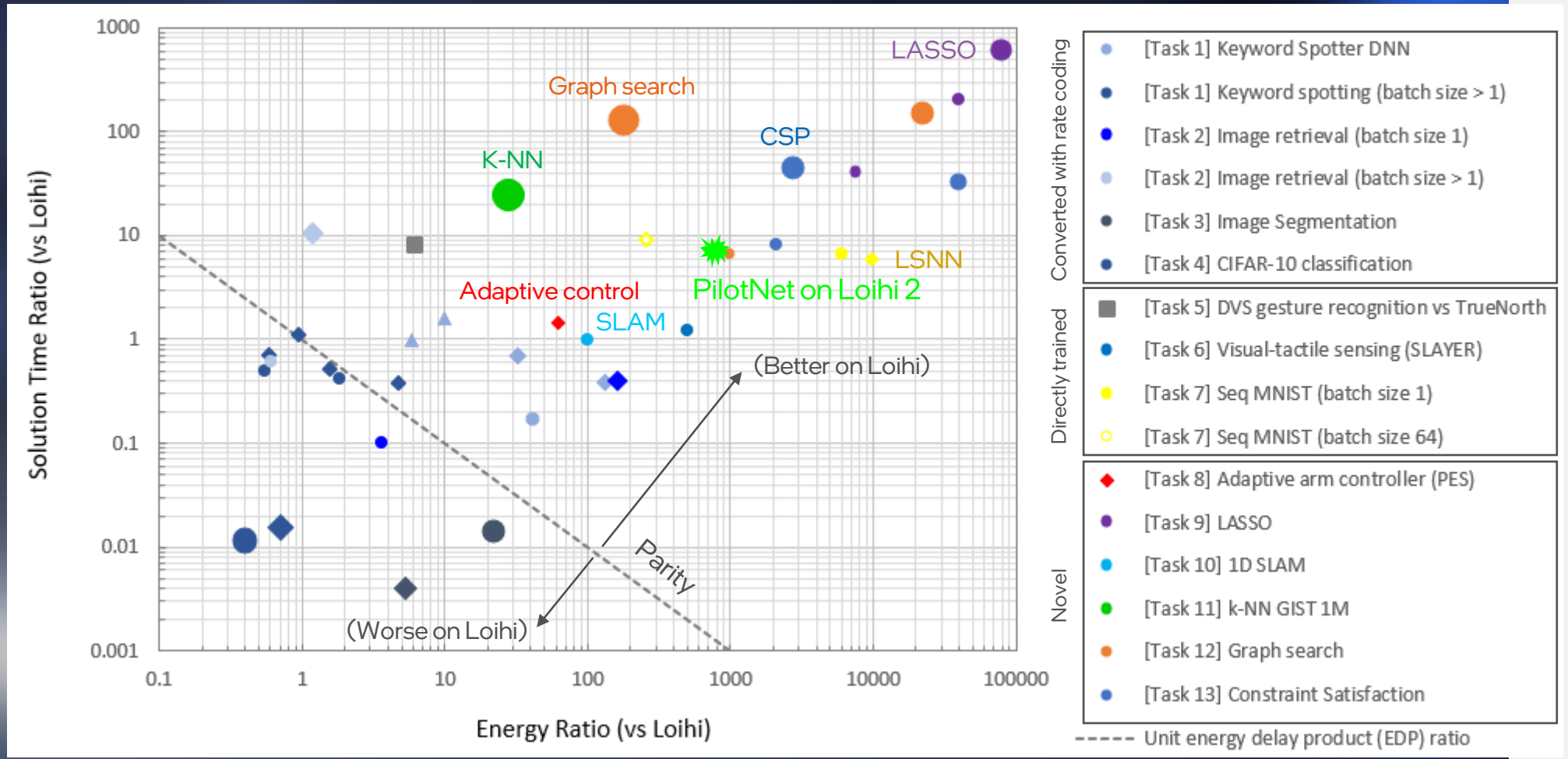
Where Loihi Shines…

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
- Jetson Orin

> 10^3 EDP gain of Loihi 2 vs. latest Nvidia Jetson Orin for real-time DNN signal processing

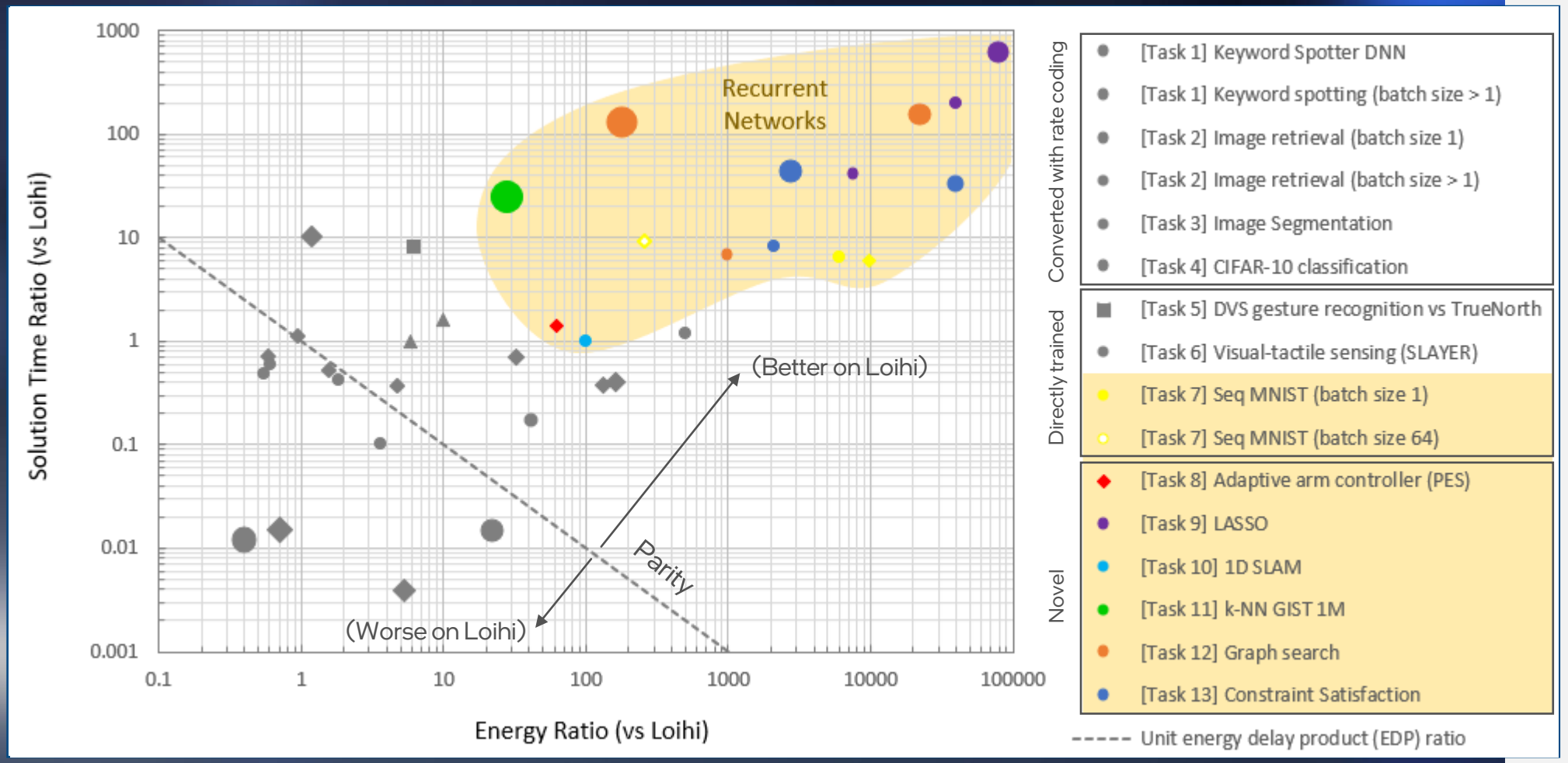


M. Davies et al, "Advancing Neuromorphic Computing With Loihi: A Survey of Results and Outlook," Proc. IEEE, 2021. Results may vary.

Novel recurrent networks give the best gains

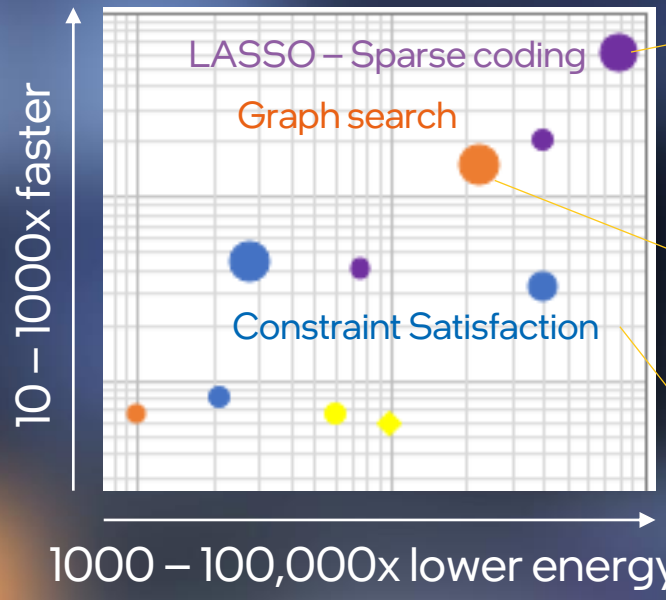
Reference architecture

- 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.

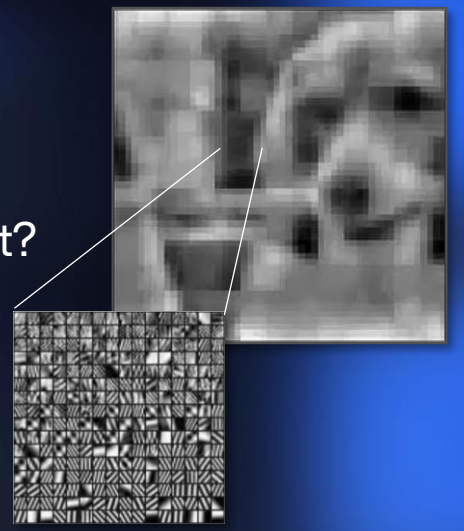
Zooming in on the best examples: Optimization problems



What features best explain the sensory input?

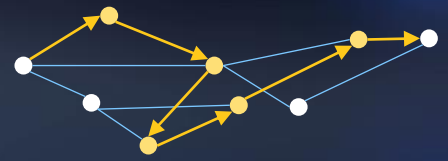
$$\underset{z}{\operatorname{argmin}} \|x - Dz\|_2^2 + \lambda \|z\|_1$$

Input Reconstruction Sparse regularization



What is the shortest path to my goal?

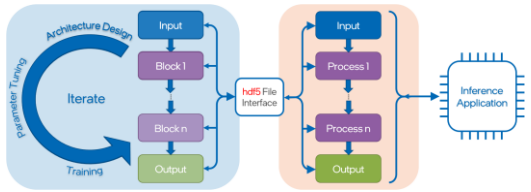
What is the shortest path while visiting each waypoint exactly once?



Current algorithmic focus areas: Lava algorithm libraries

lava-dl

- Direct & HW-aware training of event-based DNNs
- Rich neuron model library (feed-forward & recurrent)

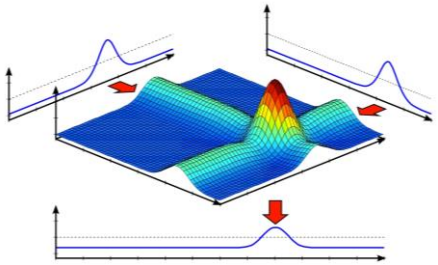


lava-optim

- Family of constraint optimization solvers
- Today: QP, QUBO
- Future: MPC, LCA, ILP, ...
- Standalone use or as part of AI applications

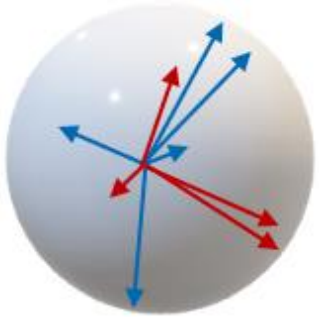
lava-dnf

- Design models with attractor dynamics
- Stabilize temporal data
- Selective data processing
- Dynamic working memories



lava-vsa (WIP)

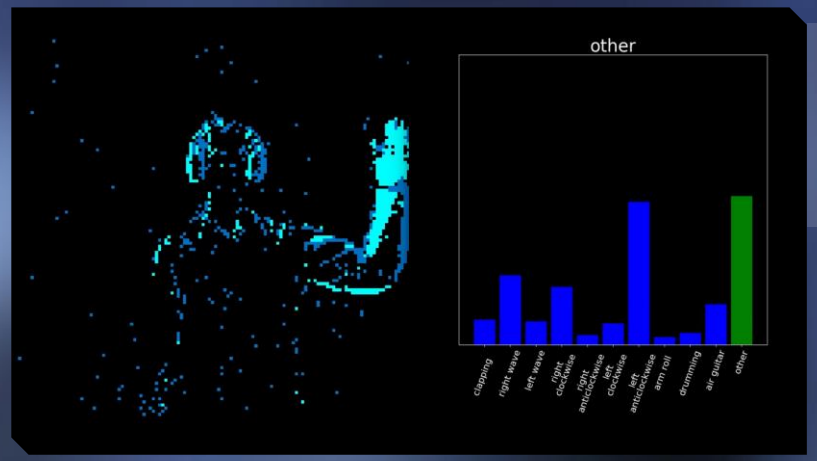
- API for algebraic model description for VSAs
- Library of data types and operations (composition, binding, factorization, ...)



Future directions

- lava-io (sensor/actuator interfaces)
- lava-robotics (control, planning, physical simulator interfaces)
- lava-evolve (evolutionary training methods)
- lava-ui (graphical network creation, visualization, debugging)
- Signal processing
- Off-the-shelf apps (segmentation, tracking, keyword detection, ...)
- Neural simulators (Brian2Lava, ...)

Loihi Has Confirmed the Value in these areas in the past

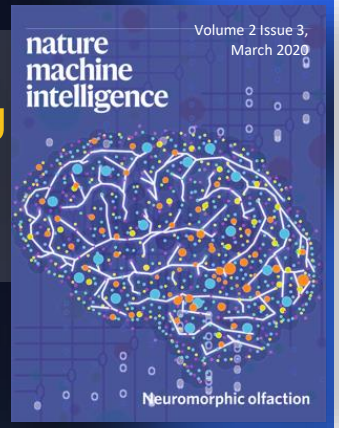


Gesture recognition + learning

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

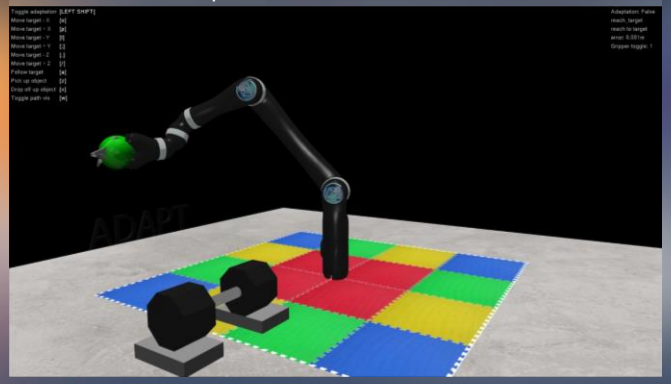
Olfaction-inspired odor recognition and learning

3000x more data efficient learning than a deep autoencoder



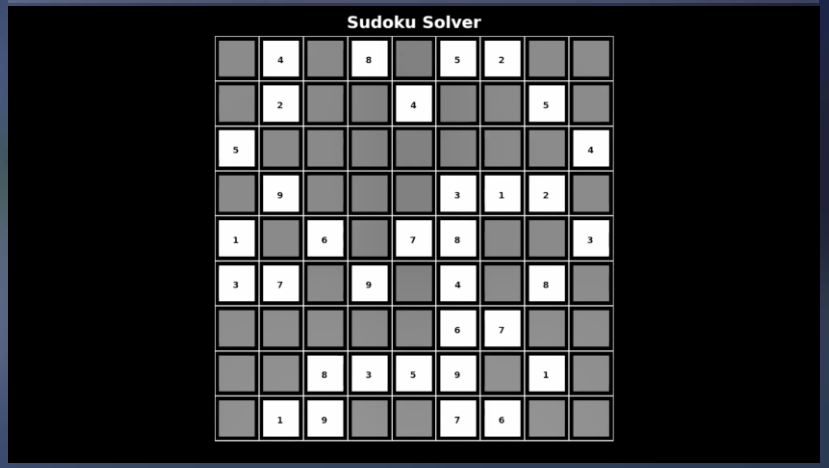
Adaptive robotic arm control

40x lower power, 50% faster vs GPU



Combinatorial optimization (CSP, SAT, ILP, QUBO)

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



Scene understanding

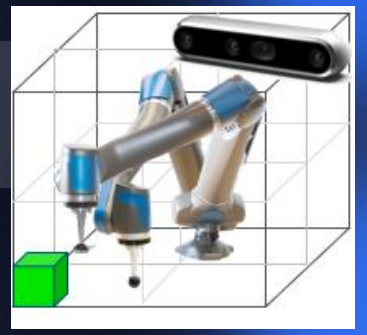
Integrated behaviors: Object recognition, tracking, learning
100x lower power vs CPU

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

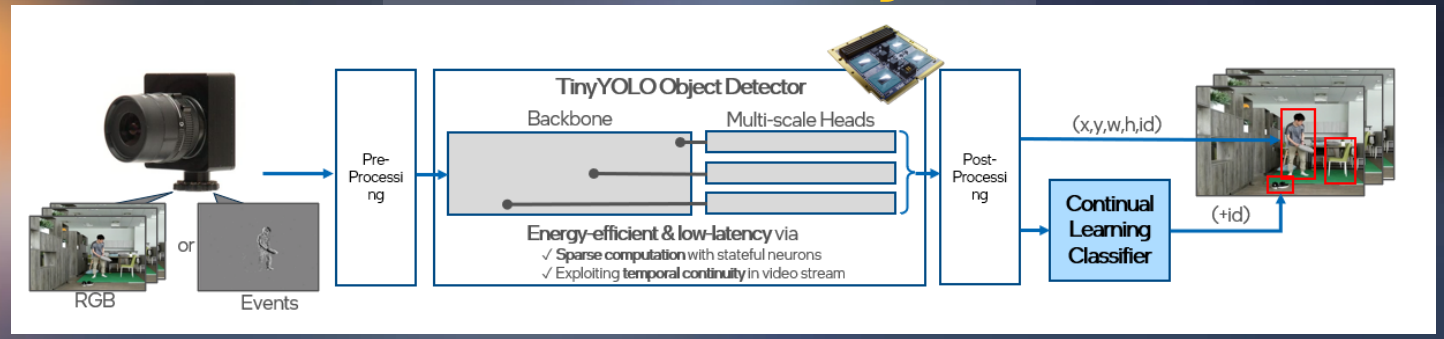
Loihi will confirm value soon in …

Model predictive control for robotic control

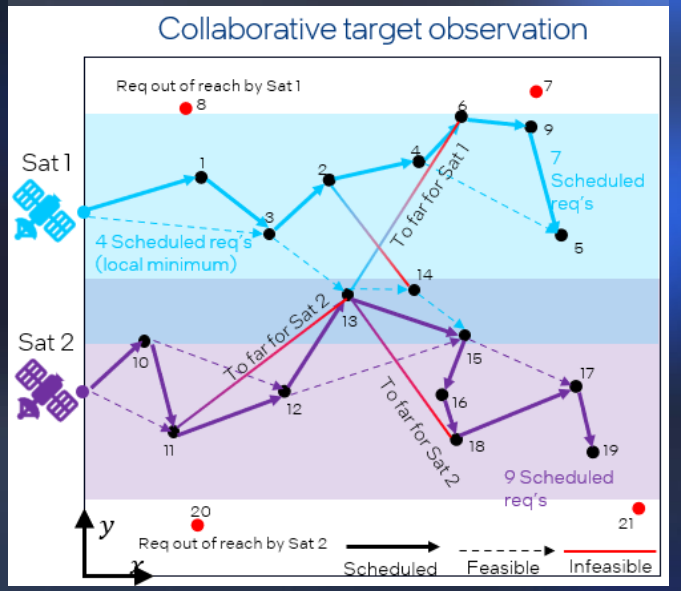
Graph search & motion planning

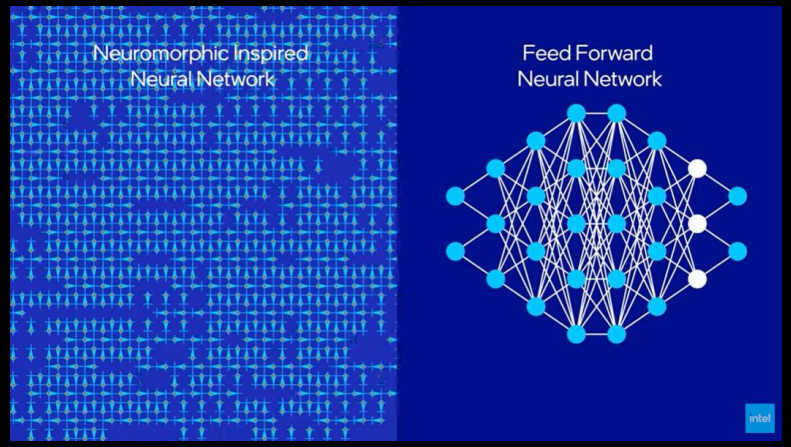


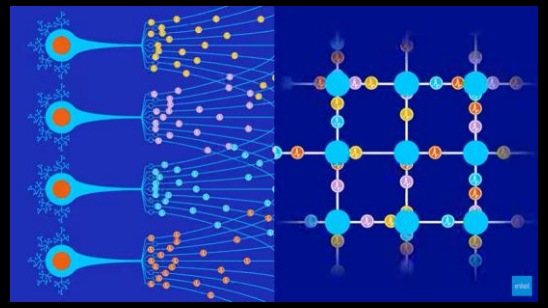
Continual Learning

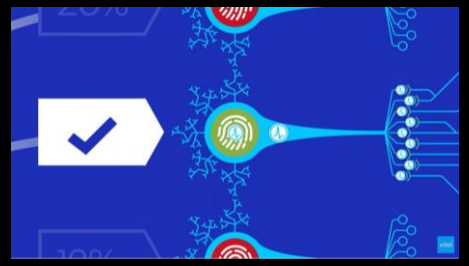


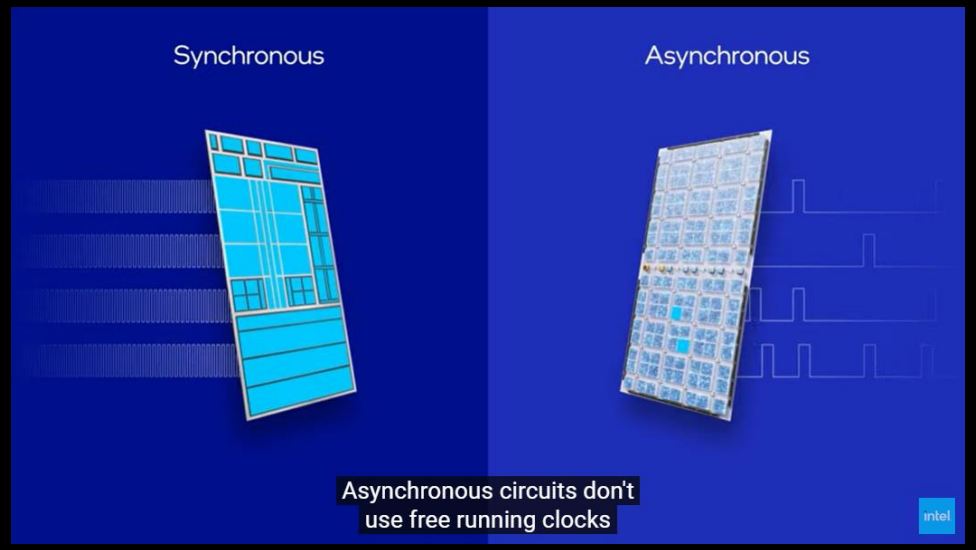
Satellite Scheduling











Watch them on YouTube: [Part 1]() [Part 2]()

For more info contact:

Ashish Rao Mangalore, ashish.rao.mangalore@intel.com
Gabriel Fonseca Guerra, gabriel.fonseca.guerra@intel.com
Andreas Wild, andreas.wild@intel.com