Technology Brief

Intel Labs' Loihi 2 Neuromorphic Research Chip and the Lava Software Framework

Taking Neuromorphic Computing to the Next Level with Loihi 2

Intel Labs' new Loihi 2 research chip outperforms its predecessor by up to 10x and comes with an open-source, community-driven neuromorphic computing framework

"For the first time, we are seeing a quantitative picture emerge that validates this promise. Together, with our research partners, we plan to build on these insights to enable wide-ranging disruptive commercial applications for this nascent technology."

—Mike Davies
Director of Intel's Neuromorphic Computing Lab

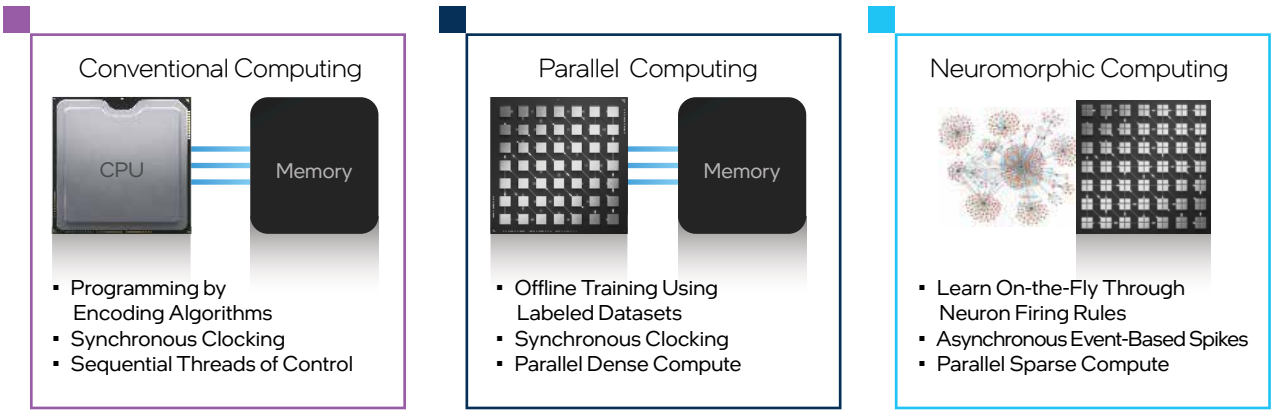
Introduction

Recent breakthroughs in AI have swelled our appetite for intelligence in computing devices at all scales and form factors. This new intelligence ranges from recommendation systems, automated call centers, and gaming systems in the data center to autonomous vehicles and robots to more intuitive and predictive interfacing with our personal computing devices to smart city and road infrastructure that immediately responds to emergencies. Meanwhile, as today's AI technology matures, a clear view of its limitations is emerging. While deep neural networks (DNNs) demonstrate a near limitless capacity to scale to solve large problems, these gains come at a very high price in computational power and pre-collected data. Many emerging AI applications—especially those that must operate in unpredictable real-world environments with power, latency, and data constraints—require fundamentally new approaches.

Neuromorphic computing represents a fundamental rethinking of computer architecture at the transistor level, inspired by the form and function of the brain's biological neural networks. Despite many decades of progress in computing, biological neural circuits remain unrivaled in their ability to intelligently process, respond to, and learn from real-world data at microwatt power levels and millisecond response times.

Guided by the principles of biological neural computation, neuromorphic computing intentionally departs from the familiar algorithms and programming abstractions of conventional computing so it can unlock orders of magnitude gains in efficiency and performance compared to conventional architectures. The goal is to discover a computer architecture that is inherently suited for the full breadth of intelligent information processing that living brains effortlessly support.

Today's Computing Architectures



Three Years of Loihi Research

Intel Labs is pioneering research that drives the evolution of compute and algorithms toward next-generation AI. In 2018, Intel Labs launched the Intel Neuromorphic Research Community (Intel NRC) and released the Loihi research processor for external use. The Loihi chip represented a milestone in the neuromorphic research field. It incorporated self-learning capabilities, novel neuron models, asynchronous spike-based communication, and many other properties inspired from neuroscience modeling, with leading silicon integration scale and circuit speeds.

Over the past three years, Intel NRC members have evaluated Loihi in a wide range of application demonstrations. Some examples include:

- Adaptive robot arm control
- Visual-tactile sensory perception
- Learning and recognizing new odors and gestures
- Drone motor control with state-of-the-art latency in response to visual input
- Fast database similarity search
- Modeling diffusion processes for scientific computing applications
- Solving hard optimization problems such as railway scheduling

In most of these demonstrations, Loihi consumes far less than 1 watt of power, compared to the tens to hundreds of watts that standard CPU and GPU solutions consume. With relative gains often reaching several orders of magnitude, these Loihi demonstrations represent breakthroughs in energy efficiency.[1]

Furthermore, for the best applications, Loihi simultaneously demonstrates state-of-the-art response times to arriving data samples, while also adapting and learning from incoming data streams. This combination of low power and low latency, with continuous adaptation, has the potential to bring new intelligent functionality to power- and latency-constrained systems at a scale and versatility beyond what any other programmable architecture supports today.

Loihi has also exposed limitations and weaknesses found in today's neuromorphic computing approaches. While Loihi has one of the most flexible feature sets of any neuromorphic chip, many of the more promising applications stretch the range of its capabilities, such as its supported neuron models and learning rules. Interfacing with conventional sensors, processors, and data formats proved to be a challenge and often a bottleneck for performance. While Loihi applications show good scalability in large-scale systems such as the 768-chip Pohoiki Springs system, with gains often increasing relative to conventional solutions at larger scales, congestion in inter-chip links limited application performance.

Loihi's integrated compute-and-memory architecture foregoes off-chip DRAM memory, so scaling up workloads requires increasing the number of Loihi chips in an application. This means the economic viability of the technology depends on achieving significant improvements in the *resource density* of neuromorphic chips to minimize the number of required chips in commercial deployments.

One of the biggest challenges holding back the commercialization of neuromorphic technology is the lack of software maturity and convergence. Since neuromorphic architecture is fundamentally incompatible with standard programming models, including today's machine-learning and AI frameworks in wide use, neuromorphic software and application development is often fragmented across research teams, with different groups taking different approaches and often reinventing common functionality. Yet to emerge is a single, common software framework for neuromorphic computing that supports the full range of approaches pursued by the research community that presents compelling and productive abstractions to application developers.

The Nx SDK software developed by Intel Labs for programming Loihi focused on low-level programming abstractions and did not attempt to address the larger community's need for a more comprehensive and open neuromorphic software framework that runs on a wide range of platforms and allows contributions from throughout the community. This changes with the release of Lava.

Intel Labs is pioneering research that drives the evolution of compute and algorithms toward next-generation AI.

Loihi 2: A New Generation of Neuromorphic Computing Architecture

Building on the insights gained from the research performed on the Loihi chip, Intel Labs introduces Loihi 2. A complete tour of the new features, optimizations, and innovations of this chip is provided in the final section. Here are some highlights:

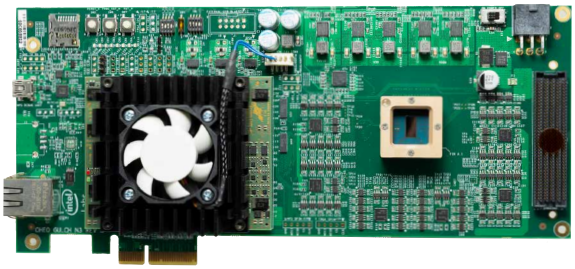
- **Generalized event-based messaging.** Loihi originally supported only binary-valued spike messages. Loihi 2 permits spikes to carry integer-valued payloads with little extra cost in either performance or energy. These generalized spike messages support event-based messaging, preserving the desirable sparse and time-coded communication properties of spiking neural networks (SNNs), while also providing greater numerical precision.
- **Greater neuron model programmability.** Loihi was specialized for a specific SNN model. Loihi 2 now implements its neuron models with a programmable pipeline in each neuromorphic core to support common arithmetic, comparison, and program control flow instructions. Loihi 2's programmability greatly expands its range of neuron models without compromising performance or efficiency compared to Loihi, thereby enabling a richer space of use cases and applications.

- **Enhanced learning capabilities.** Loihi primarily supported two-factor learning rules on its synapses, with a third modulatory term available from non-localized "reward" broadcasts. Loihi 2 allows networks to map localized "third factors" to specific synapses. This provides support for many of the latest neuro-inspired learning algorithms under study, including approximations of the *error backpropagation* algorithm, the workhorse of deep learning. While Loihi was able to prototype some of these algorithms in proof-of-concept demonstrations, Loihi 2 will be able to scale these examples up, for example, so new gestures can be learned faster with a greater range of presented hand motions.
- **Numerous capacity optimizations to improve resource density.** Loihi 2 has been fabricated with a preproduction version of the Intel 4 process to address the need to achieve greater application scales within a single neuromorphic chip. Loihi 2 also incorporates numerous architectural optimizations to compress and maximize the efficiency of each chip's neural memory resources. Together, these innovations improve the overall resource density of Intel's neuromorphic silicon architecture from 2x to over 160x, depending on properties of the programmed networks.
- **Faster circuit speeds**. Loihi 2's asynchronous circuits have been fully redesigned and optimized, improving on Loihi down to the lowest levels of pipeline sequencing. This has provided gains in processing speeds from 2x for simple neuron state updates to 5x for synaptic operations to 10x for spike generation.[2] Loihi 2 supports minimum chip-wide time steps under 200ns; it can now process neuromorphic networks up to 5000x faster than biological neurons.
- **Interface improvements.** Loihi 2 offers more standard chip interfaces than Loihi. These interfaces are both faster and higher-radix**.** Loihi 2 chips support 4x faster asynchronous chip-to-chip signaling bandwidths,[3] a destination spike broadcast feature that reduces inter-chip bandwidth utilization by 10x or more in common networks,[4] and three-dimensional mesh network topologies with six scalability ports per chip. Loihi 2 supports glueless integration with a wider range of both standard chips, over its new Ethernet interface, as well as emerging event-based vision (and other) sensor devices.

Using these enhancements, Loihi 2 now supports a new deep neural network (DNN) implementation known as the Sigma-Delta Neural Network (SDNN) that provides great gains in speed and efficiency compared to the rate-coded spiking neural network approach commonly used on Loihi. SDNNs compute graded activation values in the same way that conventional DNNs do, but they only communicate significant changes as they happen in a sparse, event-driven manner. Simulation characterizations show that SDNNs on Loihi 2 can improve on Loihi's rate-coded SNNs for DNN inference workloads by over 10x in both inference speeds and energy efficiency.[5]

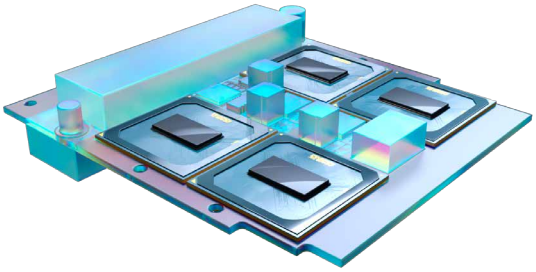
Available Loihi 2 Hardware

Intel currently offers two Loihi 2-based neuromorphic systems to researchers. Primary access to Loihi 2 is through the Neuromorphic Research Cloud, where teams engaged in the Intel NRC have access to shared systems.



Oheo Gulch:
Single-chip system for early evaluation

Designed primarily for lab testing, each Oheo Gulch board contains a single-socketed Loihi 2 chip instrumented for characterization and debug. An Intel® Arria® 10 FPGA interfaces to Loihi 2 and provides remote access over Ethernet. These are the first systems made available to Intel NRC partners through Intel's Neuromorphic Research Cloud, while larger-scale systems remain in development.



Kapoho Point:
Compact, stackable 8-chip system (coming soon)

Kapoho Point improves upon Loihi's Kapoho Bay, offering eight Loihi 2 chips in an approximately 4x4-inch form factor with an Ethernet interface. This system is ideal for portable projects and exposes general-purpose input/output (GPIO) pins and standard synchronous and asynchronous interfaces for integration with sensors and actuators for embedded edge and robotics applications. Kapoho Point boards can be stacked to create larger systems in multiples of eight chips. Kapoho Point will be available for remote access in the Neuromorphic Research Cloud and on loan to Intel NRC research teams.

Lava: A Community-Driven, Open-Source Neuromorphic Computing Framework[6]

One of the more fundamental challenges facing the field of neuromorphic computing has been the lack of clear, productive programming models for the hardware. Loihi 2 comes with Lava—a new, open-source software framework for developing neuro-inspired applications and mapping them to neuromorphic platforms. The Lava architecture is platform-agnostic; we have intentionally structured the code so that it is not tied to our own neuromorphic chips.

It's also modular and composable, so people can integrate the best algorithmic ideas from different groups and contribute back to a common code base. Lava is extensible and hierarchical, so people can build up levels of abstraction to make neuromorphic programming accessible to a broader developer community. Our overriding goal with Lava is to encourage convergence in the field with a common, open, professionally developed software foundation.

Lava includes Magma, a low-level interface for mapping and executing neural network models and sequential processes to neuromorphic hardware. This layer now includes cross-platform execution support, so applications can be developed in simulation on CPUs/GPUs before being deployed to Loihi 2 (or other) neuromorphic platforms. This layer also includes a profiler that can measure or estimate performance and energy consumption across the targeted back-end platforms.

Lava also supports channel-based asynchronous message passing. Lava specifies, compiles, and executes a collection of processes mapped to a heterogenous execution platform including both conventional and neuromorphic components. Communication between all processes occurs over an event-based message-passing backbone and API available to all processes. This presents developers with an overall programming paradigm of communicating sequential processes or the actor model, which supports extreme levels of parallelism. Messages in Lava vary in granularity from single-bit spikes to buffered packets with arbitrary payloads.

Other aspects of Lava include:

- **Offline training.** Lava supports tools such as SLAYER, enabling a range of different event-driven neural networks to be trained offline with backpropagation and integrated with other modules specified in Lava.
- **Integration with third-party frameworks.** Lava is fully extensible, supporting eventual interfaces to frameworks like Robotic Operating System (ROS), YARP, TensorFlow, PyTorch, Nengo, and more. These interfaces enable people to construct applications spanning heterogeneous systems and real-world applications.
- **Python interfaces.** For ease of adoption, all libraries and features in Lava are exposed through Python, with optimized libraries and underlying C/C++/CUDA/OpenCL code where necessary to provide excellent performance.
- **Open-source framework with permissive licensing.** Lava is freely available on GitHub to encourage community growth and convergence, and runs on CPU/GPU platforms without requiring any legal agreement with Intel. The software is available for free use under BSD-3 and LGPL-2.1 licensing. The lowest-level components necessary for deploying applications to Loihi 2 hardware systems remain accessible only to engaged Intel NRC members, at no cost.

Collaborating to Advance Neuromorphic Computing

The Intel Neuromorphic Research Community (Intel NRC) is a collaborative research effort that brings together teams from academic, government, and industry organizations around the world to overcome the wide-ranging challenges facing the field of neuromorphic computing. Members of the NRC receive access to Intel's Loihi research chips in support of their neuromorphic projects. Intel offers several forms of support to engaged members, including Loihi 2 hardware, academic grants, early access to results, and invitations to community workshops. Membership is free and open to all qualified groups.

Intel created the NRC because we believe no single organization can effectively unlock the full potential of neuromorphic computing. By collaborating with some of the leading researchers in this field spanning academia, industry and government, Intel is working to overcome the challenges in the development of neuromorphic computing and to progress it from research prototypes to industry-leading products over the coming years. The group has grown rapidly since its inception in 2018 and now includes more than 140 members. The community's body of research and results paint a picture of neuromorphic computing being well-suited for an emerging class of bio-inspired intelligent workloads that also have commercial relevance.

"In just a few years, we've formed a vibrant community comprising hundreds of researchers around the world inspired by the promise of neuromorphic computing to deliver orders of magnitude gains in computing efficiency, speed, and intelligent functionality. For the first time, we are seeing a quantitative picture emerge that validates this promise. Together, with our research partners, we plan to build on these insights to enable wide-ranging disruptive commercial applications for this nascent technology."

–Mike Davies,
Director of Intel's Neuromorphic Computing Lab

As the NRC grows, Intel will continue investing in this unique ecosystem and working with members to provide technology support and explore where neuromorphic computing can add real-world value for problems, big and small. Additionally, Intel continues to apply learnings from the NRC to future generations of Loihi chips.

A First Tour of Loihi 2

Loihi 2 has the same base architecture as its predecessor Loihi, but comes with several improvements to extend its functionality, improve its flexibility, increase its capacity, accelerate its performance, and make it easier to both scale and integrate into a larger system (see Figure 1).

Base Architecture

Building on the strengths of its predecessor, each Loihi 2 chip consists of microprocessor cores and up to 128 fully asynchronous neuron cores connected by a network-on-chip (NoC). The neuron cores are optimized for neuromorphic workloads, each implementing a group of spiking neurons, including all synapses connecting to the neurons. All communication between neuron cores is in the form of spike messages. The number of embedded microprocessor cores has doubled from three in Loihi to six in Loihi 2. Microprocessor cores are optimized for spike-based communication and execute standard C code to assist with data I/O as well as network configuration, management, and monitoring. Parallel I/O interfaces extend the on-chip mesh across multiple chips—up to 16,384—with direct pin-to-pin wiring between neighbors.

New Functionality

Loihi 2 supports fully programmable neuron models with graded spikes. Each neuron model takes the form of a program, which is a short sequence of microcode instructions describing the behavior of a single neuron. The microcode instruction set supports bitwise and basic math operations in addition to conditional branching, memory access, and specialized instructions for spike generation and probing. Table 1 summarizes the different types of instructions.

Table 1. Highlights of the Loihi 2 Instruction Set

OP CODES	DESCRIPTION
RMW, RDC *read-modify-write, read-and-clear*	Access neural state variables in the neuron's local memory space
MOV, SEL *move, move if 'c' flag*	Copy neuron variables and parameters between registers and the neuron's local memory space
AND, OR, SHL *and, or, shift left*	Bitwise operations
ADD, NEG, MIN *add, negate, minimum*	Basic arithmetic operations
MUL_SHR *multiply shift right*	Fixed precision multiplication
LT, GE, EQ *less than, not equal, equals*	Compare and write result to 'c' flag
SKP_C, JMP_C *skip ops, jump to program address based on 'c' flag*	Branching to navigate program
SPIKE, PROBE *spike, send probe data*	Generate spike or send probe data to processor

Neurons can optionally generate and transmit *graded spikes*—a generalization of Loihi's binary spike messages—that carry a 32-bit spike payload, specified by microcode. A graded spike's integer-valued payload multiplies the weights of downstream synapses. Typically, only eight bits of precision are used, representing a negligible extra energy cost compared to binary spike processing.

Support for three-factor learning rules. Loihi 2 programmable neuron models may now manipulate synaptic input received from its dendritic compartments with arbitrary microcode and assign the results to third factor modulatory terms available to a neuron's synaptic learning rules. This

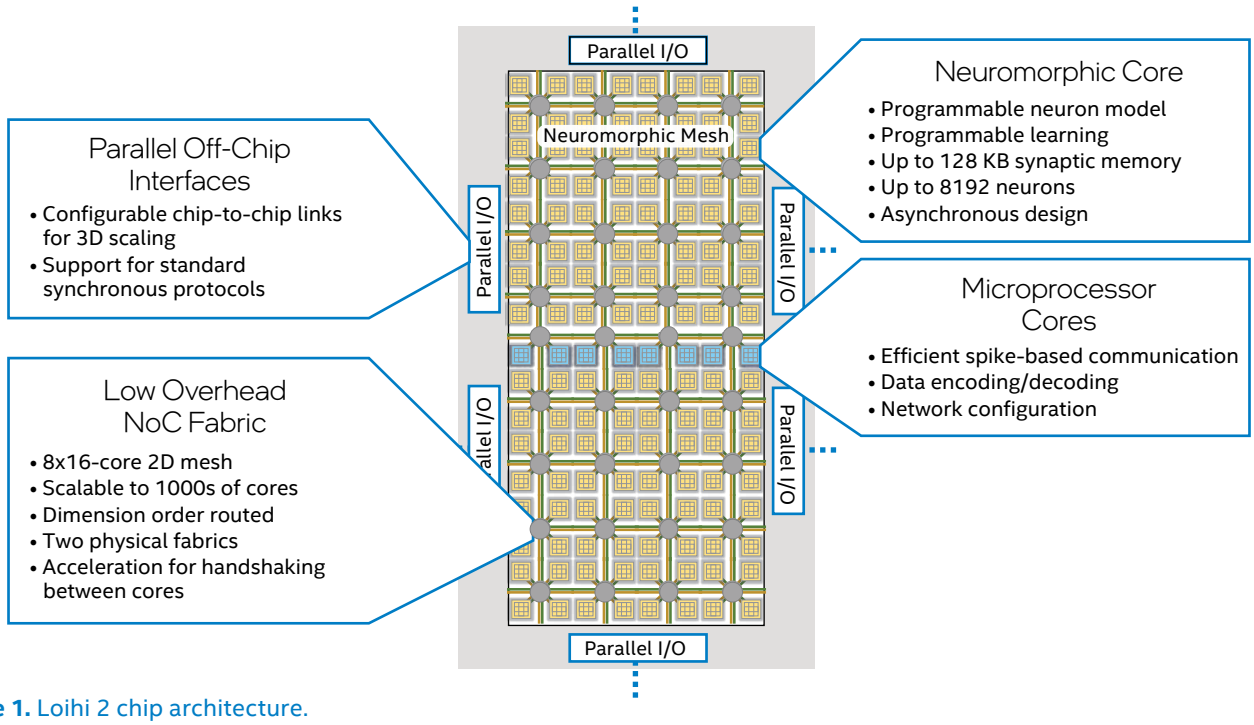


Figure 1. Loihi 2 chip architecture.

greatly generalizes the support Loihi offered for a reward-based broadcast mechanism that was intended for the same purpose of modulating learning rules. Loihi 2's modulatory factors may be mapped uniquely per post-synaptic neuron, such as to represent errors that apply to individual neurons in support of supervised learning algorithms.

Capacity Improvements

Loihi 2 achieves a 2x higher synaptic density than Loihi. Compared to Loihi, the Loihi 2 neuron cores are approximately half the size for a similar synaptic memory capacity. Although each Loihi 2 core has slightly less aggregate memory, the effective core capacity is significantly higher because the memory architecture in Loihi 2 is more efficient.

Loihi 2 cores support flexible memory partitioning to increase the effective core capacity. Where Loihi used many discrete individual memories, each with a fixed allocation, Loihi 2 asynchronously aggregates its memories to allow different functions in the core to access a variable number of memory banks (see Figure 2). This allows the overall memory resources to be soft-partitioned in order to achieve the optimal balance for a particular application— for instance, between neurons and synapses—on a per-core basis. This results in a higher effective core capacity. For example, the most common configuration of a Leaky-Integrate-and-Fire neuron model requires 4x fewer memory resources in Loihi 2 compared to Loihi. In this case, Loihi 2 can implement 4x the neurons in the same memory footprint. Even greater neuron density can be achieved by reducing neuron precision or by reclaiming memory from other features.

Loihi 2 offers advanced connectivity compression features to better utilize available memory. Loihi 2 still supports the various sparse and dense synapse encodings supported by Loihi, but longer synapse lists and more flexibility in synaptic precision results in more efficient synapse encoding. Bigger improvements come from Loihi 2's support for convolutional, factorized, and stochastic connections. Loihi 2's convolution feature supports strided, dilated, and sparse kernels. The kernel is only stored once on the core, with synaptic targets computed on-the-fly. The factorized connectivity feature can compress synaptic memory from $O(n^2)$ to $O(n)$ when the connectivity matrix can be expressed as the product of two vectors. Finally, the stochastic connectivity feature supports the procedural generation of synapses from a single seed. These powerful new features can increase effective synaptic capacities by many factors; for example, 17x for some convolutional networks to over 80x for stochastic connections.

Loihi 2 increases the number of embedded processors per chip to 6 from 3 in Loihi. These processors are programmed with conventional C or Python code and perform many essential tasks related to encoding and decoding data across the neuromorphic domain, as well as management and housekeeping operations. The increased processor count helps to prevent these conventional processing tasks from bottlenecking overall application performance, as occasionally occurred in Loihi.

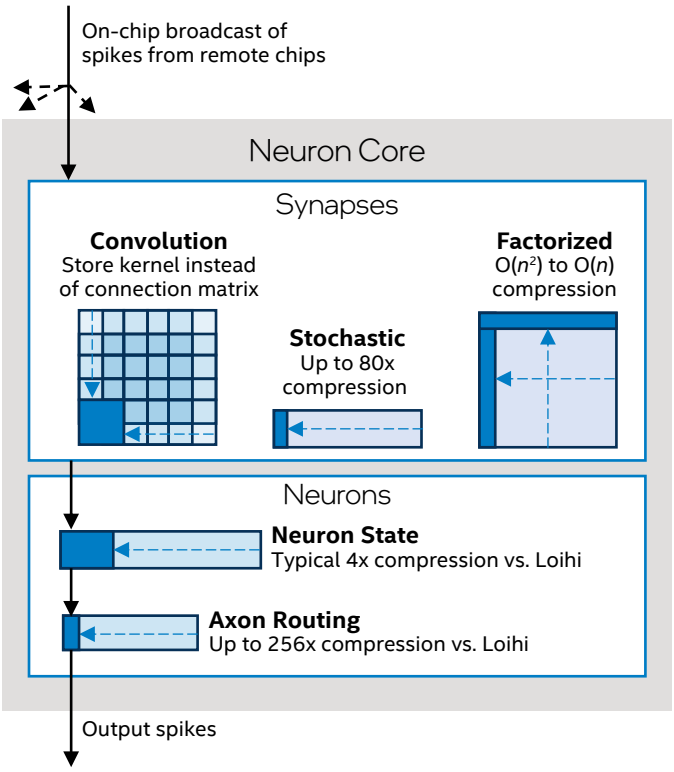


Figure 2. Loihi 2's memory partitioning is more flexible and efficient than Loihi's.

I/O and Scalability

Loihi 2 supports local broadcast of spikes at a destination chip to alleviate congestion on chip-to-chip channels. Inter-chip links are fewer in number and inherently slower than on-chip links, thus introducing a potential bottleneck through which all inter-chip traffic must flow. Loihi 2's new local broadcast feature significantly reduces inter-chip traffic, which can result in increases of over 10x in effective bandwidth for multi-chip workloads while freeing up routing table resources in the sending core.

Loihi 2 supports 3D multi-chip scaling, resulting in shorter routing distances between chips and further reducing the congestion of inter-chip links. A variety of asynchronous inter-chip protocols optimized for different distances and pin-counts is available to allow flexibility in building systems of chips with different densities and physical configurations.

Loihi 2 supports standard interfaces for easier system integration with non-Loihi devices. Supported interfaces include 1000BASE-KX, 2500BASE-KX and 10GBase-KR Ethernet, GPIO, and both synchronous (SPI) and asynchronous (AER) handshaking protocols. A spike I/O module at the edge of the chip provides configurable hardware accelerated expansion and encoding of input data into spike messages, reducing the bandwidth required from the external interface and improving performance while reducing load on the embedded processors.

Loihi 2 at a Glance

Table 2 provides a comprehensive comparison of Loihi 2 features versus Loihi features.

Table 2. Comparison of Loihi to Loihi 2

Resources/Features	Loihi	Loihi 2
Process	Intel 14nm	Intel 4
Die Area	60 mm^2	31 mm^2
Core Area	0.41 mm^2	0.21 mm^2
Transistors	2.1 billion	2.3 billion
Max # Neuron Cores/Chip	128	128
Max # Processors/Chip	3	6
Max # Neurons/Chip	128,000	1 million
Max # Synapses/Chip	128 million	120 million
Memory/Neuron Core	208 KB, fixed allocation	192 KB, flexible allocation
Neuron Models	Generalized LIF	Fully programmable
Neuron State Allocation	Fixed at 24 bytes per neuron	Variable from 0 to 4096 per neuron depending on neuron model requirements
Connectivity Features	Basic compression features: • Variety of sparse and dense synaptic compression formats • Weight sharing of source neuron fanout lists	In addition to the Loihi 1 features: • Shared synapses for convolution • Synapses generated from seed • Presynaptic weight-scaling factors • Core fan-out list compression and sharing • Broadcast of spikes at destination chip
Information Coding	Binary spike events	Graded spike events (up to 32-bit payload)
Neuron State Monitoring (for development/debug)	Requires remote pause and query of neuron memory	Neurons can transmit their state on-the-fly
Learning Architecture	Programmable rules applied to pre-, post-, and reward traces	Programmable rules applied to pre-, post-, and generalized "third-factor" traces
Spike Input	Handled by embedded processors	Hardware acceleration for spike encoding and synchronization of Loihi with external data stream
Spike Output	1,000 hardware-accelerated spike receivers per embedded processor	In addition to the Loihi 1 feature, hardware accelerated spike output per chip for reporting graded payload, timing, and source neuron
External Loihi Interfaces	Proprietary asynchronous interface	Support for standard synchronous (SPI) and asynchronous (AER) protocols, GPIO, and 1000BASE-KX, 2500BASE-KX, and 10GBase-KR Ethernet
Multi-Chip Scaling	2D tile-able chip array Single inter-chip asynchronous protocol with fixed pin-count	3D tile-able chip array Range of inter-chip asynchronous protocols with variable pipelining and pin-counts optimized for different system configurations
Timestep Synchronization	Handled by cores	Accelerated by NoC routers

Help accelerate research and adoption of breakthrough AI systems.
For more information, visit **intel.com/content/www/us/en/research/neuromorphic-community.html** or email **inrc_interest@intel.com**.

[1] A survey of these results was recently published in M. Davies et al, "Advancing Neuromorphic Computing with Loihi: A Survey of Results and Outlook," Proc IEEE, 2021. Results may vary.

[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; Loihi 2 parameters measured from both silicon characterization with the N3B1 revision and pre-silicon circuit simulations. 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, 32 GB 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.

[3] 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.

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

[5] 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.

[6] Lava replaces the first-generation Loihi chip's Nx SDK.