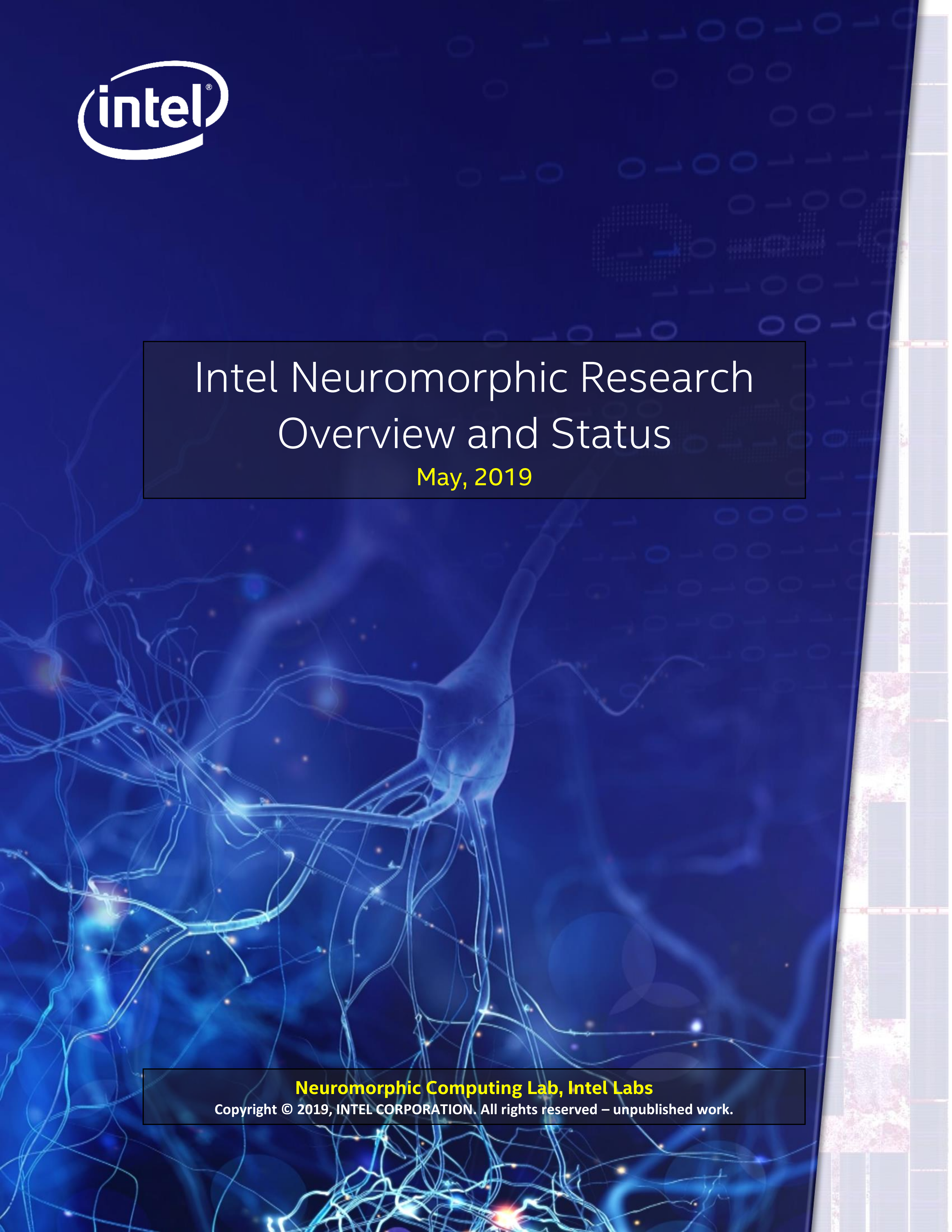


Intel Neuromorphic Research Overview and Status

May, 2019

Neuromorphic Computing Lab, Intel Labs

Background

Computer architects and machine learning researchers have long looked to the brain for inspiration. Some of the earliest computer architectures and neural network models were developed with crude models of neural computation in mind. These include John von Neumann's processor architecture that has thrived over many decades, as well as the Artificial Neural Network model that has more recently begun to deliver great practical value.

Those in the neuromorphic research field believe that now, more than ever, breakthroughs in computing and artificial intelligence may come from the study of neuroscience. The recent success of deep learning has confirmed that connectionist models of intelligence can yield impressive and dramatic gains compared to alternative methods. Deep learning has become a widely deployed tool that now impacts nearly every aspect of our lives. Yet anyone who has observed the rapid learning behavior of a toddler, or anyone who wishes to deploy advanced AI capabilities in edge devices, appreciates that our AI technology still has a long way to go before it approaches the capabilities and efficiency of natural intelligence found in the brains of organisms.

All researchers in computing, including AI researchers, are at the mercy of the computer architectures available to them. In the 1980s, advances in neural networks were limited by the poor floating point support of that era's computers. At that time, simple linear threshold units were most commonly studied, the discontinuities of which precluded the use of backpropagation and scaling to large networks. Over time, improved floating point performance encouraged researchers like LeCun, Rumelhart, and Hinton to adopt sigmoidal and rectified linear activation functions. Neural networks thereby became differentiable and those researchers were able to apply powerful optimization methods, specifically the backpropagation algorithm, which could then be scaled to large multi-layered networks. These techniques came to fruition in the 2000s with the availability of plentiful training data that the GPUs of that era could rapidly process in batched form with highly provisioned floating point resources.

Today, those interested in more advanced neuro-inspired connectionist algorithms face an even more severe computational roadblock. The neural networks we find in nature are radically different from the deep learning networks in common use today. Natural neural networks are sparsely connected and pervasively recurrent. They continuously adapt on a wide range of timescales in response to individual, non-batched, and mostly unlabeled data samples. They utilize discontinuous spike events to achieve sparse activation in time. These characteristics, among others, lead to dramatic efficiencies in nature but make them extremely ill-suited for executing on Von Neumann processors and matrix arithmetic accelerators optimized for today's deep networks.

This perspective informs our neuromorphic research program in Intel Labs. We intend to move beyond Von Neumann and matrix arithmetic architectures to arrive at a general neural processor architecture that is much better suited for the types of neural connectionist algorithms under study at the forefront of neuroscience. We seek to apply the principles of neural computation as understood today to the form and function of silicon integrated circuits. We hope that these neuromorphic architectures may better realize the brain's energy efficiency compared to conventional architectures and may match the brain's ability to learn and adapt, leading to breakthroughs in neuro-inspired computation.

The Loihi Neuromorphic Research Processor

In January of 2018, Intel released its Loihi neuromorphic research chip [1]. This first-of-its-kind processor implements a programmable learning architecture supporting a wide range of neuroplasticity mechanisms under study in the modern field of computational neuroscience (for example, see [2] [3] [4] [5] [6]). These features provide Loihi with the ability to integrate and respond to real-world information received continuously over time and to intelligently adapt behavior in order to satisfy functional objectives, similar to the operation of the brains of organisms found in nature.

By replicating many of the fundamental architectural properties of biological neurons, Loihi offers highly efficient and scalable learning performance compared to conventional machine learning methods. In particular, Loihi's advanced *spiking neural network* feature set with programmable local learning rules enable a broad class of neuro-inspired algorithms supporting supervised, unsupervised, reinforcement-based, and one-shot learning paradigms. Loihi has a fully *asynchronous* design implementation that allows it to rapidly process information in an event-driven fashion, mapping those events to fine-grain parallel units that compute with temporally sparse spike messages.

Although Loihi is a research chip that will not, in its current form, be sold commercially, Loihi was designed to near-commercial standards and fabbed with Intel's 14nm FinFET CMOS process technology. Sufficiently large quantities of Loihi chips have been manufactured to support a vibrant research ecosystem. Over the past year, Intel has developed a number of Loihi hardware systems that are now being used by collaborators to evaluate the potential of Loihi's neuromorphic learning. These include a range of devices from single-chip USB form-factor systems to 32-chip boards that users access remotely to a large 768-chip scalable system under development.

In March of 2018, Intel demonstrated an example object recognition and learning application on Loihi that operates in real-time consuming 74 mW [7]. At the *2018 NICE workshop* in Hillsboro, Oregon, Intel's collaborator, Applied Brain Research (ABR), demonstrated a compliant 6-DOF robotic arm control application in which Loihi's learning features allow the arm to adapt in real time to unpredictable environment perturbations [8].

Since those initial examples, the breadth of algorithms running on Loihi has steadily expanded, and researchers are now quantitatively evaluating Loihi's performance and efficiency compared to conventional architectures. The results are compelling. Applied Brain Research has shown that Loihi provides the most efficient solution among all commercially available computing architectures for audio deep network inference by factors ranging from 5x to over 100x [9]. Loihi supports a one-dimensional Simultaneous Localization and Mapping (SLAM) algorithm that operates at 100x lower power compared to standard CPU-based methods [10], and Intel's own evaluations of the Spiking Locally Competitive Algorithm and a graph search (Dijkstra's) algorithm show performance improvement factors from 100 to 10,000 times compared to CPUs[1]. In general, we are finding that whenever an algorithm can be formulated to run on Loihi, leveraging the architecture's fine-grain parallelism and sparse activity, it tends to outperform alternative methods on conventional architectures by orders of magnitude. Moreover, the relative gains on Loihi *increase* with increasing problem scale.

In some cases, as hoped, Loihi's bottom-up neuro-inspired architecture is enabling algorithms that have no direct analogy to conventionally coded algorithms. Working with Thomas Cleland, a neuroscientist from Cornell who studies biophysical models of the mammalian olfactory system, we have developed olfaction-inspired networks at a level of abstraction that run on Loihi and exhibit many of the same remarkable properties of the biological system: efficient spike-based oscillatory dynamics, one-shot learning, and classification performance that exceeds conventional artificial olfaction approaches [11]. Other researchers are exploring astrocyte modeling on Loihi in order to stabilize emergent neural oscillatory dynamics [12]. This new class of oscillatory spiking neural networks, what we are calling *phasor neural networks*, are uniquely suited for Loihi's architecture. We believe they will lead to a wide space of novel and highly efficient neuro-inspired algorithms with direct correspondence to the mysterious but pervasive rhythms observed in brains (for example, [13]).

To support productive algorithms and applications research, Intel is developing a novel software stack for Loihi systems, named *Nx SDK*. The Nx SDK framework provides a general API, compiler, and runtime for Loihi development, exposing the architecture's full range of learning capabilities. It may be extended by a variety of third party frameworks. For example, with Intel's support, Applied Brain Research has ported its *Nengo* toolchain [14] to Loihi platforms. Nengo on Loihi now supports a number of baseline capabilities, including the construction of working memories, nonlinear dynamical systems, multi-layer perceptron networks trained with TensorFlow, and learning modules using ABR's PES rule [15]. Over the coming months, we expect collaborators to implement Nx SDK integration with the *EONS* evolutionary optimization framework from University of

[1] LASSO comparison: FISTA algorithm from SPAMS (http://spams-devel.gforge.inria.fr/) running on an Intel Core i7-4790 3.6GHz with 32GB RAM was compared to the LCA module of NxSDK v0.8 on Loihi. For graph search: the NetworkX implementation of Dijkstra's Algorithm (https://networkx.github.io/) on an Intel Xeon 6136 3.00 GHz with 32GB RAM was compared to the performance of a 32-chip Nahuku system running NxSDK v0.8.

Performance results are based on testing as of December 2018 and may not reflect all publicly available security updates. No product can be absolutely secure.

Tennessee and Oak Ridge National Labs as well as with the Human Brain Project's *Neurorobotics Platform* [16].

More information about Loihi's architecture and programming model is available in recent Intel Labs publications [1] [17] [18], as well as on Intel's INRC website [19] (private access, invitation available upon request to inrc_interest@intel.com.)

Intel Neuromorphic Research Community

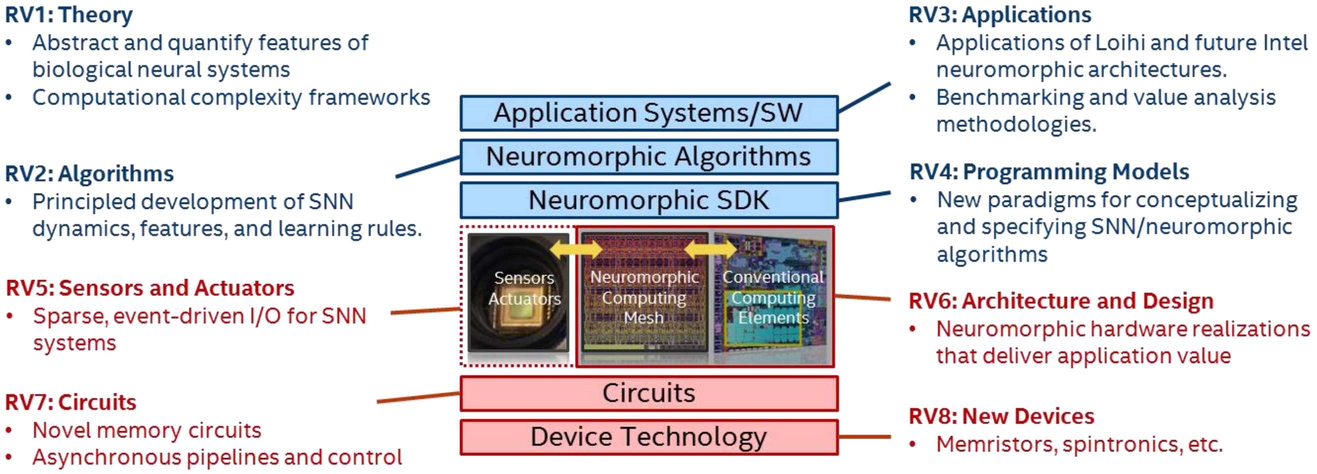
1. *Objectives and Structure*

In March of 2018, Intel Labs launched the Intel Neuromorphic Research Community (INRC). This collaborative research program is open to worldwide academic, government, and industry research groups interested in tackling the hurdles facing the adoption of neuromorphic architectures for mainstream computing applications. INRC members are using Loihi, as the industry's only fully functional and high performance neuro-inspired processor architecture, to accelerate their research. Intel hopes the findings of this community will drive future improvement of neuromorphic architectures, software, and systems, eventually leading to the commercialization of this nascent technology.

For the majority of INRC activities, Intel's primary role is to provide access to Loihi systems and the Loihi Software Development Kit (Nx SDK) across a range of engagement models. Additionally, a limited amount of funding is available from Intel's Corporate University Research Office (CUR) in support of particularly compelling academic project proposals. In April of 2018 a call for proposals was issued resulting in the submission of over sixty project proposals. Given a budget of $2.25M over three years ($750k per year), Intel was only able to fund 12 of these proposals, but the majority of PIs who submitted proposals have since managed to engage with the INRC and pursue their proposed research on Loihi in a self-funded manner, albeit with reduced project scope.

Since neuromorphic computing entails nothing less than a bottom-up rethinking of computer architecture, beginning potentially at the device technology level, unsolved and important research problems can be found at all levels of the computing stack, from the process technology level to circuits to microarchitecture and silicon design to algorithms to software programming models to end applications. Figure 1 illustrates the full range of neuromorphic computing research vectors Intel and others are engaged in. Red vectors indicate hardware-oriented areas; blue vectors indicate theory and usage-oriented areas.

It is our view that, with the availability of Loihi, the state of neuromorphic hardware capabilities now significantly leads the state of algorithmic, application, and programming understanding. As such, the focus of INRC is on research vectors RV1 through RV5. For the foreseeable future, we intend to use Loihi and future silicon iterations as our primary vehicle for collaboration with a broadening network of researchers. Although we support and internally pursue research vectors RV6-8, these areas are not the focus of the INRC.



RV1: Theory
- Abstract and quantify features of biological neural systems
- Computational complexity frameworks

RV2: Algorithms
- Principled development of SNN dynamics, features, and learning rules.

RV5: Sensors and Actuators
- Sparse, event-driven I/O for SNN systems

RV7: Circuits
- Novel memory circuits
- Asynchronous pipelines and control

RV3: Applications
- Applications of Loihi and future Intel neuromorphic architectures.
- Benchmarking and value analysis methodologies.

RV4: Programming Models
- New paradigms for conceptualizing and specifying SNN/neuromorphic algorithms

RV6: Architecture and Design
- Neuromorphic hardware realizations that deliver application value

RV8: New Devices
- Memristors, spintronics, etc.

Figure 1. Neuromorphic Research Vectors

Given the state of neuromorphic research and our ranking of priorities, we expect INRC activity and Intel's funding investments to focus disproportionately on RV2 initially with, we hope, a shift to RV3 over time as a broadening portfolio of mature algorithms enable a wide range of applications. We see RV1, RV4, and RV5 vectors as important but more limited areas of INRC focus.

The ultimate goals for INRC and Intel's neuromorphic research are the following:

1) Identify, develop, and characterize a space of algorithms that exploit the novel properties of spiking neural network hardware architectures (e.g. fine-grain parallelism, sparse activity, high degrees of connectivity) to deliver orders of magnitude gains in efficiency compared to the leading conventional algorithms.

2) Guide the iterative development of neuromorphic, non-Von Neumann architectures with the algorithmic and architectural insights that arise in the pursuit of goal (1).

3) Prototype real-world applications of Intel's neuromorphic silicon to assess the practical value that this architecture may provide if commercialized.

4) Develop an ecosystem of researchers and developers who can routinely and successfully apply neuromorphic silicon to solve new problems, paving the way to a broad commercial ecosystem that can support the proliferation of this technology into the world.

Intel hopes its network of INRC members will advance the state-of-the-art understanding of neuromorphic learning algorithms, demonstrating the value of this emerging technology for a wide range of application domains. Most of the enabling software and results from these efforts will be contributed to the public domain in the form of publications and open source software.

As a commercial enterprise, it's difficult for Intel to prioritize the advancement of neuroscience understanding among its goals for the INRC, but we do hope and expect this to be an important side benefit of the community's work. To the degree that such pursuits are synergistic with our goals above, Intel may offer support in the form of funding, letters of support, and of course access to our Loihi systems. To date, we've engaged with a number of groups whose primary motivation is to model biological systems for neuroscience research purposes.

2. INRC Engagement

As of Q1 of 2019, Intel has formally engaged with over 50 active research projects across 47 groups. The teams span 31 universities, five U.S. national labs, four nonprofit research institutes, and five industry partners. INRC academic partners include 18 in the U.S., one in Canada, 10 in the E.U., and two in Asia. To date, we've allocated approximately two thirds of INRC's $750k/year funding over three years to 12 of these academics groups.

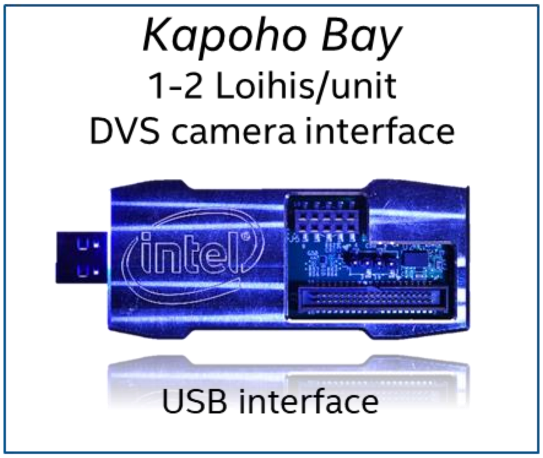
Intel has hosted two week-long INRC workshops, one in the Fall of 2018 located in Reykjavik, Iceland; the second in January of 2019 in Portland, Oregon. A number of travel grants were provided to students to support their attendance at these two workshops, which drew 65 and over 90 researchers respectively. Additionally, Intel has hosted or supported extended tutorials at numerous conferences over the past year (e.g. at the NICE workshops in 2018 and 2019, ICONS, Riken, Telluride, and others). Our third INRC workshop is planned for October, 2019 in Munich, Germany.

We continue to accept new proposals, especially for unfunded access, on an ongoing basis. Interest continues to rise, with seven new project proposals received in Q1 of 2019. Our Loihi remote cloud systems now have over 160 user accounts.

Loihi Neuromorphic Systems and Software

1. *Kapoho Bay*

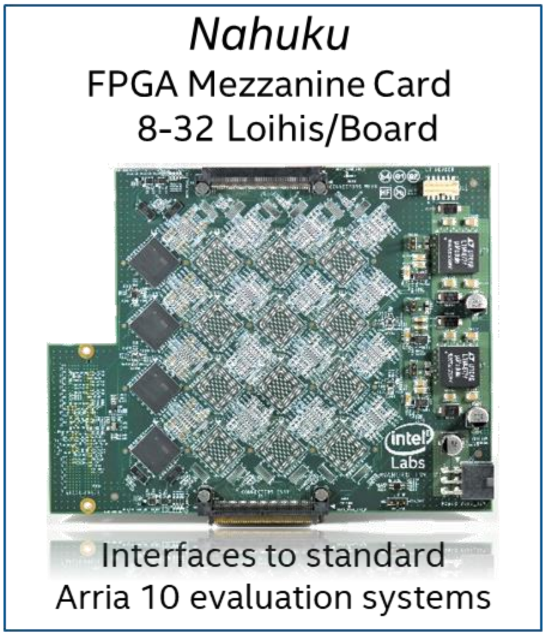
Intel has developed a modular small-scale system, code named *Kapoho Bay*, for deploying Loihi in portable or embedded environments. The system is integrated in a USB stick form factor, shown right, providing a host computer with a Loihi-accelerated subsystem comprising up to 256,000 spiking neurons and 256 million synapses (*i.e.* two Loihi chips). Kapoho Bay supports a variety of peripheral interfacing options: GPIO pins, I2C, and a DVS AER interface supporting the IniVation DAVIS 240C DVS camera.



A *peripheral toolkit* is under development for Kapoho Bay, expected to be released later in 2019, that will allow researchers to enhance an I/O FPGA on one of the Kapoho bay boards with hardware support for new sensors and actuators. This capability will allow Kapoho Bay boards to be embedded in larger systems in order to support diverse mobile deployments, e.g. on a robot or drone.

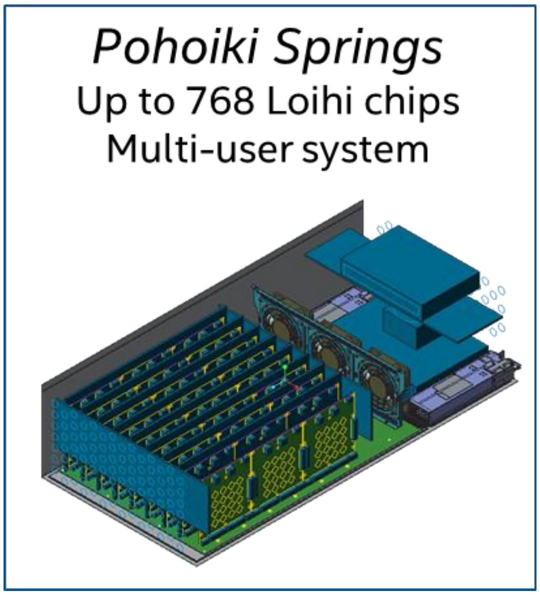
2. *Nahuku FPGA Mezzanine Card*

For networks requiring higher neuron counts than what Kapoho Bay can support, Intel has developed an 8- to 32-chip Loihi carrier card, code named *Nahuku*, intended to interface with the Intel® Arria® 10 SoC Development Kit via the kit's FPGA Mezzanine Card (FMC) connector. Depending on the number of populated Loihi chips, a Nahuku FMC system implements up to 1-4 million spiking neurons and 1-4 billion synapses.



The Arria 10 SoC includes a dual-core ARM subsystem which runs an Ubuntu distribution of Linux. Intel's Nx SDK software runs on the Arria 10's internal ARM CPU and connects to the Nahuku mesh of Loihi chips via the ARM AXI Bus. The user interacts with the system via the Arria 10's Ethernet connection, calling the Nx SDK on the ARM to set up and execute a specified spiking neural network.

The Arria 10 Dev Kit supports a wide range of interfaces that allow users to connect sensors and other peripherals to the Loihi chips, either by developing software on the ARM CPU or by mapping custom hardware logic into the Arria 10 FPGA. Intel offers a DVS Adapter FPGA Mezzanine Card that may be inserted into a second FMC connector on the Arria 10 Dev Kit that



provides connectivity to the IniVation DAVIS 240C event-based camera.

Currently, Intel hosts a pool of Nahuku-enabled Arria 10 systems in its datacenter available for remote use by INRC members. Users log on to Xeon servers and then launch their Loihi jobs via the SLURM job scheduling system into the attached pool of Nahuku neuromorphic systems. A similar cloud-based Loihi service has been set up at Argonne National Labs for DOE researchers.

3. *Pohoiki Springs*

Pohoiki Springs is our largest Loihi system under development, expected to be available for external use in Q3 of 2019. It is intended for deployment in a standard 19-inch datacenter rack. The rack-mounted chassis includes a Xeon host processor, dual power supplies, and a backplane with connectivity for up to 24 Nahuku boards and three FPGA mesh interface cards. In its maximum configuration, one Pohoiki Springs system will support up to 768 Loihi chips, providing up to 100 million neurons and 100 billion synapses. Although it will be possible to map very large spiking neural networks into Pohoiki Springs, its parallelism is equally intended for two other use cases: (1) simultaneous use by a large number of remote users, and (2) acceleration of evolutionary optimization algorithms that require spawning large numbers of networks whose fitness will each be evaluated concurrently.

4. *Neuromorphic Software Development Kit (Nx SDK)*

Intel has developed a software development kit, named Nx SDK, that is needed in order to compile and run spiking neural networks (SNNs) on Loihi systems. Nx SDK presents a Python API that will be familiar to those with experience developing for other spiking neural network simulators and frameworks. The Nx SDK spans up to three software layers running on CPUs at different proximities to the Loihi mesh of neuromorphic resources:

1. **Embedded layer**: SNN-interfacing processes that run on the embedded x86 processors in each Loihi chip.

2. **Host layer**: Processes responsible for communicating data in and out of the Loihi chip mesh.

3. **Super-host layer**: User-interface code that compiles user-specified networks into the Loihi chips and provides execution feedback to the user (e.g. in the form of graphical waveforms).

In some cases, such as for Kapoho Bay, the Host and Super-host layers are combined and both run on the system's USB host CPU. Nx SDK requires Python 3.5.5 and Ubuntu (16.04 LTS), which may be run inside a super-host virtual machine.

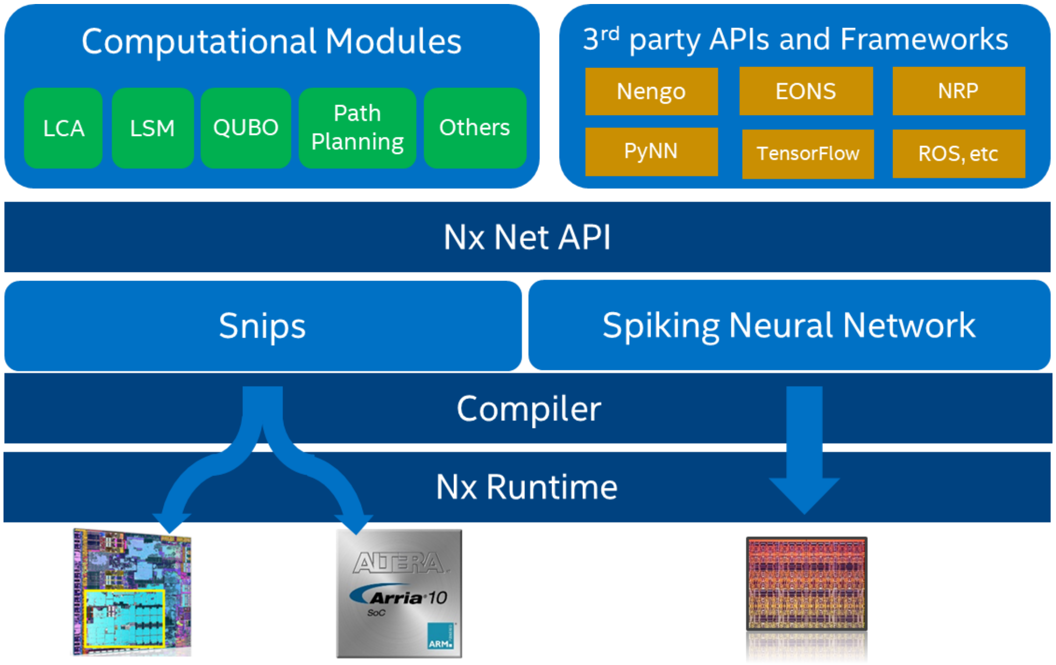


Figure 2. Nx SDK Architecture

Figure 2 shows the high-level architecture of Nx SDK. Today, most users develop for Loihi systems using the Nx Net API, which provides abstractions for constructing spiking neural networks similar to those of the Brian, Nengo, or PyNN frameworks. Users may also define conventionally coded sequential processes, called Snips (for "Sequential neural interfacing processes"). Snips are intended to interact closely with the defined spiking neural network, often performing encoding and decoding functions necessary to convert real-world data to and from spatiotemporal spike patterns. Snips may also implement more complex functions such as background structural plasticity algorithms.

The NxSDK compiler maps the user-specified snips and spiking neural network into the underlying heterogeneous computing substrate. Spiking neurons are allocated to the optimal neuromorphic cores across the multi-chip Loihi mesh in order to maximize resource utilization. Each snip is mapped to the CPU nearest to its associated SNN resources with sufficient resources to satisfy the snip's memory and performance requirements. All communication between snips and spiking neurons occurs either over spike messages or generalized message-passing channels.

NxSDK is intended to be extended upward to higher-level computational modules that abstract and hide the complexity of snips and spiking neural network dynamics. To date, a number of modules have been released, e.g. the Locally Competitive Algorithm (LCA) for solving LASSO optimization problems [20] and a constraint satisfaction solver [21]. Many other will follow, including those independently developed and released by INRC members.

As shown in Figure 2, we expect to integrate Nx SDK with a variety of third party APIs and machine learning frameworks. To date, integration with Nengo [14] has been implemented, which also offers a facility to train deep spiking neural networks using TensorFlow. Preliminary support for the Human Brain Project's Neurorobotics Framework [16] has also been demonstrated. Others will follow over the coming months.

References

[1] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, Y. Liao, C.-K. Lin, A. Lines, R. Liu, D. Mathaikutty, S. McCoy, A. Paul, J. Tse, G. Venkataramanan, Y.-H. Weng, A. Wild, Y. Yang and H. Wang, "Loihi: a Neuromorphic Manycore Processor with On-Chip Learning," *IEEE Micro,* 2018.

[2] E. M. Izhikevich, "Solving the distal reward problem through linkage of STDP and dopamine signaling," *Cerebral Cortex,* 2007.

[3] F. Zenke, W. Gerstner and S. Ganguli, "The temporal paradox of Hebbian learning and homeostatic plasticity," *Current Opinion in Neurobiology,* no. 43, pp. 166-176, 2017.

[4] J. Sacramento, R. P. Costa, Y. Bengio and W. Senn, "Dendritic error backpropagation in deep cortical microcircuits," *arXiv:1801.00062,* Dec 2017.

[5] C. Tetzlaff, S. Dasgupta, T. Kulvicius and F. Wörgötter, "The Use of Hebbian Cell Assemblies for Nonlinear Computation," *Scientific Reports,* p. 12866, 2015.

[6] G. Bellec, D. Salaj, A. Subramoney, R. Legenstein and W. Maass, "Long short-term memory and learning-to-learn in networks of spiking neurons," *arXiv,* May 2018.

[7] M. Mayberry, "Intel Creates Neuromorphic Research Community to Advance 'Loihi' Test Chip," 1 March 2018. [Online]. Available: https://newsroom.intel.com/editorials/intel-creates-neuromorphic-research-community/.

[8] C. Eliasmith, "Intel 2018 Demo Video," Vimeo, 10 April 2018. [Online]. Available: https://vimeo.com/264198863.

[9] P. Blouw, X. Choo, E. Hunsberger and C. Eliasmith, "Benchmarking Keyword Spotting Efficiency on Neuromorphic Hardware," *arXiv.1812.01739,* 2018.

[10] G. Tang, A. Shah and K. P. Michmizos, "Spiking Neural Network on Neuromorphic Hardware for Energy-Efficient Unidimensional SLAM," *arXiv:1903.02504,* 2019.

[11] N. Imam and T. A. Cleland, "Rapid learning and robust recall in a neuromorphic olfactory circuit," *(in review).*

[12] K. P. Michmizos, "Towards Critical SNNs: Astrocytes Detect Chaos in Neuronal Dynamics," in *Neuro-Inspired Computating Elements Workshop*, Albany, NY, 2019.

[13] E. P. Frady and F. T. Sommer, "Robust computation with rhythmic spike patterns," *arXiv:1901.07718,* 2019.

[14] Applied Brain Research, "Nengo Ecosystem," 2016. [Online]. Available: https://www.nengo.ai/projects.html.

[15] D. MacNeil and C. Eliasmith, "Fine-Tuning and the Stability of Recurrent Neural Networks," *PLoS ONE,* vol. 6, no. 9, 2011.

[16] Human Brain Project, "HBP Neurorobotics Platform," 2018. [Online]. Available: https://neurorobotics.net/.

[17] C.-K. Lin, A. Wild, G. N. Chinya, Y. Cao, M. Davies, N. Srinivasa, D. M. Lavery and H. Wang, "Programming Spiking Neural Networks on Intel's Loihi," *Computer,* vol. 51, no. 3, pp. 52-61, 2018.

[18] M. Davies, "Neuro-Inspired Computational Elements Workshop," 27 March 2019. [Online]. Available: https://niceworkshop.org/wp-content/uploads/2019/04/NICE-2019-DAY-2a-Mike-Davies.pdf.

[19] Intel Corporation, "Intel Neuromorphic Research Community," 2018. [Online]. Available: http://neuromorphic.intel.com/.

[20] P. T. P. Tang, T.-H. Lin and M. Davies, "Sparse Coding by Spiking Neural Networks: Convergence Theory and Computational Results," *arXiv,* May 2017.

[21] G. A. F. Guerra and S. B. Furber, "Using Stochastic Spiking Neural Networks on SpiNNaker to Solve Constraint Satisfaction Problems," *Frontiers in Neuroscience,* vol. 11, p. 714, 2017.

Notices and Disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more.

Intel, the Intel logo, Xeon, Arria, and Stratix are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries. Other names and brands may be claimed as the property of others. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors.