## DISCLAIMER

**This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Reference herein to any social initiative (including but not limited to Diversity, Equity, and Inclusion (DEI); Community Benefits Plans (CBP); Justice 40; etc.) is made by the Author independent of any current requirement by the United States Government and does not constitute or imply endorsement, recommendation, or support by the United States Government or any agency thereof.**

**2024 Neuromorphic Computing for Science Workshop**

# Position Papers

**September 12-13, 2024**
**Bethesda, MD**


**Co-Chairs**
Gina Adam, George Washington University
Garrett Kenyon, Los Alamos National Laboratory
Thomas Potok, Oak Ridge National Laboratory

**Organizing Committee**
Giorgio Ascoli, George Mason University
Frances Chance, Sandia National Laboratories
Yiran Chen, Duke University
Joseph Friedman, University of Texas at Dallas
Cory Merkel, Rochester Institute of Technology
Maryam Parsa, George Mason University
Midya Parto, University of Central Florida
Catherine Schuman, University of Tennessee Knoxville
Shinjae Yoo, Brookhaven National Laboratory
Yuping Zeng, University of Delaware



U.S. DEPARTMENT *of* ENERGY | Office of Science

## Disclaimer



U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research
Points of Contact:

Robinson Pino, robinson.pino@science.doe.gov
Hal Finkel, hal.finkel@science.doe.gov
Marco Fornari, marco.fornari@science.doe.gov
Margaret Lentz, margaret.lentz@science.doe.gov
Kalyan Perumalla, kalyan.perumalla@science.doe.gov
David Rabson, david.rabson@science.doe.gov
Bill Spotz, william.spotz@science.doe.gov

https://doi.org/10.2172/2506703

# Contents

**Jha, Flow-based Crossbar Computing and Neuronal Stochastic Dynamics**

**Johnson, Robust Autonomy via Reward-Modulated Insect Circuitry**

**Kharel et al., Beyond Energy Efficiency: Neuromorphic Primitives for More General Metrics**

**Kudithipudi, What are the Building Blocks for Neuro-Inspired Continual Learning?**

**Kulkarni et al., Transformers-enabled Discovery of Neuromorphic Circuit Primitives from Large Scale Network Simulations**

**Kumar, The Need for Beyond-Backpropagation AI Training Algorithms**

**Li et al., A Brain-Inspired-Attention-Based Spiking-Driven Neuromorphic System with Compute-in-Memory Design**

**Li et al., Hypothesis-Driven Applications, Neuromorphic Circuits, and Technology Co-Design**

**Mansingh, Bio-inspired Emergent Intelligence for Scientific Computing**

**Marquez et al., Scalable Ultrafast Superconducting Neuromorphic Circuits (SNC)**

**Mukim et al., Optimal Data Encoding Methods for Neuromorphic Computing**

**Noy et al., Neuromorphic ionic computing for next-generation information processing and AI**

**Patton, Brain Inspired Large Scale Simula6on of Spiking Neural Networks**

**Porter et al., A plausible neuromorphic implementation of a novel coding scheme for memory of daily experience**

**Purohit, A multi-mode simulation framework for hardware-software co-development of neuromorphic systems**

**Qiu, Leveraging Circuit Dynamics for Temporal Applications and Event Driven Computing**

**Robinson, Neuromorphic Learning with Over-Parameterized Generalized Feedback Networks**

**Roos, Recurrent quasi-Boolean circuits for neuromorphic primitives and micro-brain modeling**

**Rothganger, What is the basic element of neural computing?**

**Saxena, Electronic Photonic Integrated Circuits (EPICs) For Neuromorphic Computing**

**Shainline, Neuromorphic Supercomputing with Superconducting Optoelectronic Networks**

**Shrivastava et al., Tissue Vs Silicon – Holistic ML Systems Perspective To Unravel AGI, Scaling Laws, and Energy Efficiency**

**Sornborger, Spike Coincidence as a General Control Mechanism for Neuromorphic Processing**

**Stan, The Case for Waferscale Neuromorphic Architecture using Asynchronous Stream Computing**

**Thakkar, Scaling Photonic Interconnects for High Neuromorphic Connectivity**

**Torbunov, Bio-inspired Adaptive and Self-Organized Learning**

**Tsang et al., Neuromorphic compressed temporal representation using spiking autoencoder**

**Varshnay, Co-Design Methodologies for Integrating Small Organism-Inspired Chiplets**

**Wang, Enabling scalable neuromorphic systems with error-aware simulation frameworks**

**Wang, Towards a Scalable Neuromorphic Domain Specific Language**

**Yakopcic et al., Exploiting Brain-Scale Computing Through a Memristor Based Bio-Inspired Analog Architecture**

**Yanguas-Gil et al., The forgotten chemical connectome: how to implement a neuromodulatory system and why should we care about it**

**Yoo, Brain-Derived Neuromorphic Computing with 3D Photonic-Electronic-Ionic Circuits**

**Young, Deep Co-design, with Cross-discipline Teams, Leveraging State of the Art Technology to Push Neuromorphic Computing Forward**

**Zand, Livewired Neuromorphic Systems: Where Evolving SNNs Meet Evolvable Hardware**

# Part 3: Pre-Workshop Report

# Part 1: Call for Position Papers

## Call for Position Papers

Engineering novel neuromorphic computing systems with functionalities, capabilities, and energy efficiency similar to biological brains is one of the most exciting and challenging scientific endeavors of our time. This workshop aims to identify key research needs, challenges, and next steps necessary to develop biologically-realistic neuromorphic circuits primitives that capture the functionality of neural systems found in nature. Moreover, simulating neuromorphic computing primitives integrated into networks will be key to understanding their behavior at scale, particularly for those computing architectures where full-scale commercial fabrication is not yet readily accessible. Appropriate neuroscience datasets and metrics will have to be established to vet proposed neuromorphic circuits.

In the development of new circuits and methodologies for neuromorphic computing, it is critical that there is close collaboration among circuit designers, computer engineers, computational neuroscientists, and algorithms and simulation researchers. This workshop aims to bring together a diverse range of experts across three complementary technical areas.

Submit your position paper to the technical areas below:

1. **Neuroscience algorithms and translation to neuromorphic analog circuits**

   This technical area is driven by the fundamental question *"What are the key neuromorphic circuit primitives that are needed to capture the full functionality of critical biological computing mechanisms?"*. The goal of the activities in this space is to understand what principles and circuit structures of brain organization and dynamics underpin its functionality and robustness capabilities and how these principles can be translated into functionally-equivalent neuromorphic circuits and systems that could be practically implemented (with available technology?). Ideas related to neuromorphic computing principles inspired from brain regions/functions (cortical, hippocampus, thalamus, sensing, motor control, etc.) are sought after. Topics related to neuromorphic approaches and emulations of small invertebrate brains are also of interest.

2. **Technologies and prototyping of neuromorphic analog primitives**

   This technical effort is driven by the fundamental question *"What are the technologies needed to demonstrate and prototype key neuromorphic circuit primitives?"* Ideas related to novel neuromorphic circuits based on new devices and designs, and new principles guided by neuroscience-inspired functionality are of interest. Ideas related to emerging analog technologies that provide orders of magnitude in performance, parallelism, energy efficiency, tunability range, temporal delays, etc., and that mimic the biological behavior and robustness of key primitives are welcomed. Also of interest are topics related to high neuromorphic connectivity capabilities, e.g. optoelectronic technologies and photonic interconnects.

3. **Scalable integration for neuromorphic computing modeling**

   The fundamental question driving this technical area is *"What are the critical characteristics for effective large-scale simulation of neuromorphic circuits and systems?"* New approaches are needed to create simulations of large-scale biologically-realistic neural networks, diverse synapse connectivity, and sophisticated network activity. Of interest are ideas related to novel methods to integrate and to scale up the simulation of the neuromorphic circuit primitives using high-performance computing in order to understand their interactions in the context of hundreds of millions of neurons and synapses. Also welcomed are novel methodologies for the efficient exploration of the large co-design space between neuromorphic algorithms and circuit technologies.

When discussing the technical idea and how it fits in the technical area(s) and the overall vision of the workshop, include a discussion on the benchmarks, metrics, and/or datasets requirements for neuromorphic computing for your proposed implementation.

# Part 2: Position Papers

Stochastic Thermodynamics and Quantum Technologies: Neuromorphic Circuit Optimization

Yohannes Abate
The University of Georgia

Are there fundamental reasons for the significant disparity in energetic efficiency between artificial computers and those observed in nature? If so, can these disparities be at least partially mitigated? Furthermore, why do biological computers consume considerably more energy than the minimum dictated by Landauer's bound, despite energy expenditure being a critical fitness cost in biological evolution? We posit that addressing these questions requires two parallel and mutually inclusive strategies. Firstly, it is imperative to re-evaluate the characteristics of computational units at the synaptic level. Secondly, we must investigate the interplay between the energy consumption of physical computational systems, their other performance metrics (such as "space" and "time" complexity as explored in computer science theory), and the constraints imposed on permissible physical processes in networks of neuromorphic systems. Such investigations have the potential to yield significant advancements in the design of energy-efficient brain-like systems and provide profound insights into the relationship between artificial and biological computation systems.

The primary focus of this proposal is twofold: i) to stimulate discussions that explore ways to optimize energy budgets and develop understanding and analysis of neuromorphic hardware at the network level using stochastic thermodynamics and fluctuation theorems and ii) to stimulate discussions on how we can implement quantum technologies at a synaptic level that are based on unitary dynamics and are inherently dissipationless, offering huge potential to further reduce energy costs, in addition to the reductions offered by neuromorphic computing.

**(i) Stochastic Thermodynamics for Optimizing Network Physical Architecture**: We envision to stimulate discussions that explore ways to optimize energy budgets and develop understanding and analysis of neuromorphic hardware using stochastic thermodynamics and fluctuation theorems. We have lots of evidence that thermodynamic costs have played a major role in determining the physical architecture of computing networks. Yet to date, there has been almost no application of stochastic thermodynamics to investigate these systems in order to deepen our understanding of the relationship among their energetic behavior, computational behavior, robustness, etc.

The relationship between the thermodynamic and computational properties of physical systems has gained central practical importance due to the escalating energetic costs of digital devices. Real-world computers adhere to multiple physical constraints, influencing their thermodynamic properties. These constraints are applicable to both natural computers, such as brains and eukaryotic cells, and artificial neuromorphic systems. Notably, all these systems must complete computations rapidly, utilizing minimal degrees of freedom, thus operating far from thermal equilibrium. 20th-century analyses of computational thermodynamics did not account for the constraints of nano-equilibrium dynamics. However, the emerging field of stochastic thermodynamics offers formal tools for analyzing computational systems under nano-equilibrium conditions. These tools can provide deeper insights into the fundamental thermodynamic properties of neuromorphic systems and their relationship to computational performance.

**(ii) Quantum Technologies**: Various materials and physical phenomena based on classical physics have been explored to utilize intrinsic materials physics to enable energy-efficient computation by replicating the functionalities of biological neurons and synapses. Merging insights from brain-inspired hardware and software with quantum phenomena and quantum technologies at the synaptic level, could unlock a new era of information processing. Quantum technologies, grounded in unitary dynamics, exhibit a fundamental property of being inherently dissipationless. Unitary dynamics means that the evolution of quantum states are described by unitary operators. Physically this means that there is no dissipation in the system. In such systems, energy is conserved, and no energy is lost to heat or other forms of dissipation during computation. This intrinsic characteristic implies that such technologies operate without energy loss due to dissipation, making them fundamentally more efficient. This contrasts sharply with classical computing systems, where energy dissipation and heat generation are significant concerns, leading to inefficiencies and increased energy costs. Consequently, quantum technologies hold significant potential to further diminish energy consumption. This reduction is poised to complement the energy savings already achievable through neuromorphic computing paradigms. Together, these advanced technologies promise a future where computational processes are not only faster and more powerful but also substantially more energy-efficient.

**Multiscale Encoding in Neuromorphic In-Memory Computing with Spintronics (Themes 1, 4)**

***Tosiron Adegbija (University of Arizona)***

**Introduction:** The human brain's remarkable computational efficiency is attributed to its complex neural architecture and diverse signaling mechanisms, and in-memory computing paradigm. One key principle underlying this efficiency is multiscale encoding, where information is represented across multiple temporal and spatial scales. This position paper advocates for understanding and implementing multiscale encoding as a cornerstone for developing next-generation neuromorphic computing systems. Furthermore, we propose leveraging multiscale encoding in neuromorphic in-memory computing architectures using non-volatile memory technologies, such as spintronics, to unlock unprecedented computational capabilities, thereby revolutionizing fields like artificial intelligence and robotics.

**Multiscale Encoding:** In vivo research has revealed that neural systems employ multiscale encoding to efficiently process and represent information[1,2]. For instance, in the auditory system, different timescales encode various aspects of sound[3], while in the visual system, different spatial scales capture details ranging from textures to global object shapes. This hierarchical organization enables efficient processing of complex sensory scenes and facilitates robust object recognition.

**Multiscale Neuromorphic In-Memory Computing:** We propose implementing multiscale encoding in neuromorphic in-memory computing architectures using spintronics[4], a nonvolatile memory technology that leverages the spin of electrons to store and process information. Spintronics offers several advantages over traditional CMOS technology, including low power consumption, high density, adaptable volatility, and neuromorphic analogs[5]. These advantages make it an ideal candidate for implementing neuromorphic in-memory computing architectures that mimic the brain's energy-efficient and persistent memory capabilities, pushing the boundaries of current computational paradigms.

**Implementation Strategy:** To achieve this, we propose a cross-disciplined implementation strategy: 1) *In vivo research*: Investigate neural mechanisms underlying multiscale encoding in various brain regions and sensory modalities. This research will provide invaluable insights into the specific circuit motifs, neuronal properties, and synaptic plasticity rules that enable multiscale encoding. 2) *Neuromorphic spintronic primitives*: Develop spintronic circuit primitives that can emulate the observed multiscale encoding mechanisms. These primitives could include spintronic neurons, synapses, and adaptive filters. These circuits would enable the representation and processing of information at various temporal and spatial scales within the neuromorphic system. 3) *Neuromorphic in-memory architectures*: Design in-memory computing architectures that leverage these spintronic primitives to implement multiscale encoding in machine learning applications like spiking neural networks. These architectures could be inspired by specific brain regions, such as the auditory or visual cortex, or by more general principles of neural organization.

**Benchmarks, Metrics, and Datasets:** To evaluate multiscale encoding in neuromorphic in-memory computing systems, we propose developing benchmarks that assess performance on tasks like object recognition and speech recognition under varying conditions (e.g., image resolutions, noise levels, sampling rates). Metrics will quantify energy efficiency, robustness, and adaptability to multiscale input, focusing on performance changes across scales. Evaluation will utilize existing and novel datasets comprising multiscale sensory stimuli, such as images with varying resolutions and audio recordings with different sampling rates.

**Potential Impact:** By mimicking the brain's energy-efficient and robust information processing capabilities, multiscale encoding in neuromorphic in-memory computing using spintronics can revolutionize fields such as artificial intelligence, robotics, and high-performance computing. This approach promises breakthroughs in areas like natural language processing, computer vision, and autonomous systems. To unlock this potential, we call upon the scientific and engineering community to collaboratively integrate in vivo research, develop neuromorphic spintronic primitives, and design innovative in-memory architectures, thereby advancing this exciting research in energy-efficient neuromorphic computing.

**References**


1. B. Harland, M. Contreras, and J. Fellous, “A role for the longitudinal axis of the hippocampus in multiscale representations of large and complex spatial environments and mnemonic hierarchies,” *The Hippocampus: Plasticity and Functions,* p. 67, 2018.
2. B. Harland, M. Contreras, M. Souder, and J. Fellous, “Dorsal ca1 hippocampal place cells form a multi-scale representation of megaspace,” *Current Biology,* vol. 31, no. 10, pp. 2178-2190, 2021.
3. S. Norman-Haignere, et al., “Multiscale integration organizes hierarchical computation in human auditory cortex,” *Nature Human Behavior,* vol. 6, no. 3, pp. 455-469, 2022.
4. D. Gajaria, T. Adegbija, and K. Gomez, “CHIME: Energy-efficient STT-RAM-based concurrent hierarchical in-memory processing,” *IEEE International Conference on Application-Specific Systems, Architectures and Processors (ASAP)*, 2024.
5. S. Kulkarni, D. Kadetotad, S. Yin, J. Seo, and B. Rajendran, “Neuromorphic hardware accelerator for SNN inference based on STT-RAM crossbar arrays,” *IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, 2019.

**Neural Computing: Is it time to take a step back and, dare I suggest, perhaps start over?**

**Brad Aimone, Distinguished Member of Technical Staff; Sandia National Laboratories**
**Themes: *Neuroscience-inspired computing principles; Performance metrics***

*Warning: this will be blunt*. The neuromorphic computing field is in crisis. Despite exciting successes in novel approaches, scalable systems, and algorithm design, we still struggle to define ourselves and our potential impact to the world. Meanwhile, the AI revolution is entering its 2nd decade, with tremendous successes far beyond any expectation coupled with failures to incorporate many brain-like capabilities and a shocking disregard for energy efficiency[1]. That neuromorphic is not broadly seen as a solution to these shortcomings should be taken as an indictment on us for failing to offer a compelling path forward.

Here, I contend that is the lack of a cohesive theoretical framework that is holding the neural computing field back. This lack of a framework has led to an increasingly divided community[2] that confuses outsiders and ultimately limits the field's broader adoption. Furthermore, it is hindering progress: any proposed neural algorithm is diminished in impact if it does not leverage everyone's proposed hardware, while any novel hardware is seen as limited if it cannot handle all known algorithms. This is a recipe for long-term failure. The field needs independent goal posts to guide its development and inform the world. To achieve this, the field needs a clearly-defined theoretical computational framework that is both hardware- and algorithm-agnostic.

Do we not already have a theoretical framework? Yes and no. There are many *algorithm* frameworks in use today, ranging from reservoir computing to spiking neural networks to probabilistic neural circuits, among others (**Fig 1A**). How these relate to different hardware strategies is messy[3] and it is awkward to expect end users to learn multiple programming models[4]. Further, "every approach is its own framework" does not scale, thus novelty in algorithms or technology is viewed with concern rather than celebrated.

Neuromorphic requires a framework that is similarly abstracted from today's technical work. Until such a framework exists, our community will remain trapped in a cycle of speaking past one another as opposed to communicating broadly. Further, it is mistaken to argue that we do not understand enough of the brain for this formalization[1,5-7]: just as quantum computers do not aim to capture every aspect of quantum physics, neural computing does not need to capture everything about the brain. However, because the brain offers a much broader and much less well-defined source of inspiration[2], the need for a grounded theoretical framework is perhaps even *more* important for neuromorphic.

What should this framework look like? I propose that quantum computing can be a useful guide (**Fig 1B**). When people speak of quantum computing, they do not speak in generalities of computing with quantum mechanics; rather they refer to a specific theoretical model of quantum circuit computation. This model utilizes ***well-defined operations*** that leverage ***well-defined quantum mechanical concepts*** to perform computation. Independent of likelihood of success or prospective impact, this theoretical model provides a concrete reference point for hardware and algorithm development – something the neuromorphic field sorely lacks. Importantly, the quantum circuit model is not based on what quantum hardware can do today (clearly), nor does it make any restrictions on what quantum systems should do in terms of applications. As such, it is liberating for scientific exploration across the field – neither hardware nor algorithms researchers need to justify the basic assumptions of their field; they simply move forward.

We are closer today than ever…if we have courage. A framework that makes use of established hardware and algorithm communities (e.g., analog circuits, synaptic devices, spike-based communication, probabilistic computing, *etc*.) would be more likely to be accepted (**Fig 1C**); but we should not limit the field's scope to today's capabilities. For instance, the brain has more to offer as an inspiration for circuits than mechanisms alone[5]. The framework must also meet stringent theoretical criteria to be useful[2]. The goal should be to develop a framework to inspire what we should do going forward, *not looking backwards*. Defining this framework can be the first step of a revitalized, truly brain-inspired community.

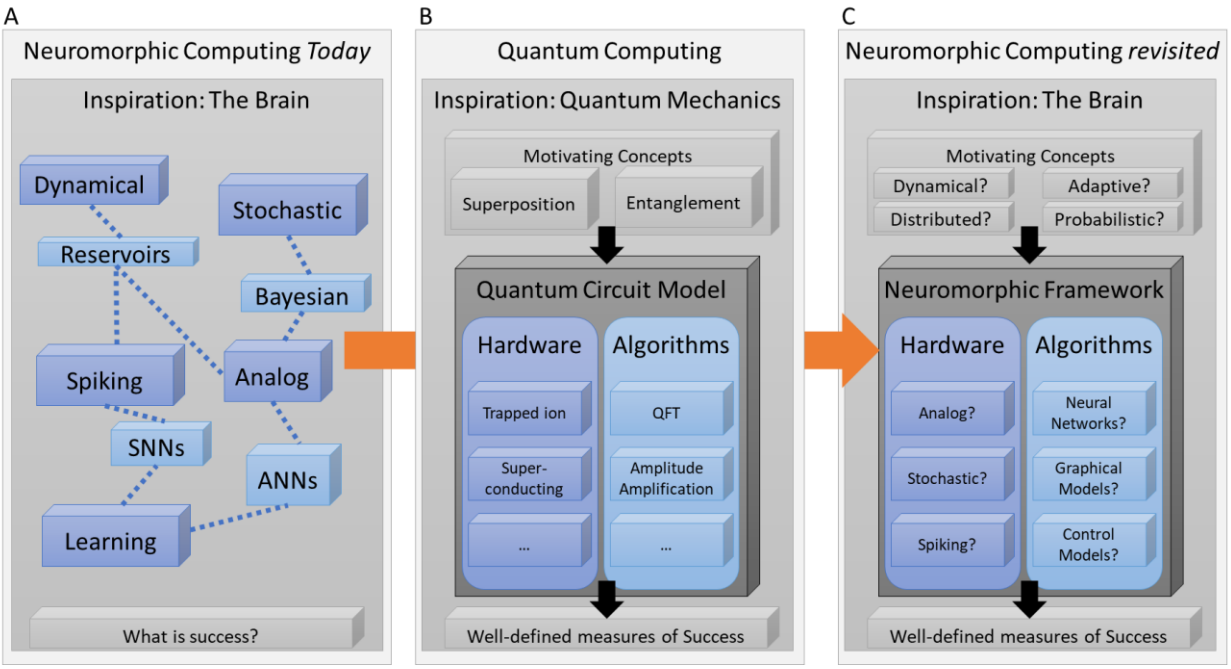



**Figure 1: (A)** Neuromorphic computing as a field has developed in an ad hoc, unorganized manner. **(B)** Quantum computing leverages a well-defined computational framework that constrains both hardware and algorithm research; yet each can advance independently of progress in the other. **(C)** This paper proposes stepping back and revisiting what such a framework would be for neuromorphic computing which would enable hardware and algorithm research to be complementary and provide concrete measures of progress and success.

**Bibliography of relevant papers**


1. Aimone, James B. "A roadmap for reaching the potential of brain-derived computing." *Advanced Intelligent Systems* 3, no. 1 (2021): 2000191.
2. Aimone, James B., and Ojas Parekh. "The brain's unique take on algorithms." *nature communications* 14, no. 1 (2023): 4910.
3. Aimone, James B., and Shashank Misra. "Will stochastic devices play nice with others in neuromorphic hardware?: There's more to a probabilistic system than noisy devices." *IEEE Electron Devices Magazine* 1, no. 2 (2023): 50-56.
4. Aimone, James B., William Severa, and Craig M. Vineyard. "Composing neural algorithms with Fugu." In *Proceedings of the International Conference on Neuromorphic Systems*, pp. 1-8. 2019.
5. Aimone, James B. "Neural algorithms and computing beyond Moore's law." *Communications of the ACM* 62, no. 4 (2019): 110-110.
6. Chance, Frances S., James B. Aimone, Srideep S. Musuvathy, Michael R. Smith, Craig M. Vineyard, and Felix Wang. "Crossing the cleft: communication challenges between neuroscience and artificial intelligence." *Frontiers in computational neuroscience* 14 (2020): 39.
7. Aimone, James B., Prasanna Date, Gabriel A. Fonseca-Guerra, Kathleen E. Hamilton, Kyle Henke, Bill Kay, Garrett T. Kenyon et al. "A review of non-cognitive applications for neuromorphic computing." *Neuromorphic Computing and Engineering* 2, no. 3 (2022): 032003.

**A dynamic neural intelligence primitive for neuromorphic systems**

Alireza Alemi, Center for Neuroscience, University of California at Davis, Davis, CA

The energy cost of training large AI models is projected to grow exponentially, raising environmental concerns [1]. In addition to energy inefficiency, gradient-based trained models are often brittle and subject to overfitting with limited data. These properties contrast with learning in biological agents, which is energy- and data-efficient with remarkable robustness to various perturbations. Neural dynamics underlie sensorimotor (e.g., walking) and cognitive (e.g., decision-making) computations. Therefore, we consider the adaptive dynamics of a neural population to act as a dynamic computational primitive of neural intelligence such that given a small number of training trajectories, it learns a generative dynamic model of the training data and is able to generalize to new circumstances.

We utilize four fundamental principles of neuroscience in designing the proposed neuromorphic computing primitive: (1) A balance of excitatory and inhibitory inputs ("E-I balance") in neocortical networks, linked to the observed irregular, Poisson-like firing, (2) Dendritic nonlinearity, (3) Localized plasticity rules, and (4) Online, causal learning through a feedback loop with the environment. The first principle can be realized with a recurrent spiking neural network framework whose architecture is derived from enforcing coding efficiency to perform dynamical computations and modeling [2,3]. Unlike the popular approach of encoding information in the firing rate of spikes, i.e., *rate coding*, the information in this framework is encoded in both spike counts and timing as efficiently as possible. The recurrent network requires a small number of spikes (tens of neurons) to perform most tasks, orders of magnitudes fewer neurons than rate coding frameworks such as Neural Engineering Framework [4], as the scaling of coding error with the number of neurons $N$ is $1/N$, compared to the scaling for rate coding which is $1/\sqrt{N}$. Incorporating the second principle as dendritic nonlinearities in this framework leads to learning *nonlinear* dynamical computation while preserving the desired efficiency and robustness properties [5]. All the weights can be trained using the third principle, i.e., local plasticity rules. We use the fourth principle to learn a stable dynamic model from training trajectories in a self-supervised manner where prediction errors are used as online teaching signals [6].

The resulting network operates on multiple timescales of neural and learning dynamics and is guaranteed to achieve global stability of the concurrent learning and computation dynamics using Lyapunov function theory. Our recent results in learning a nonlinear continuous attractor, a fundamental computational motif in cognitive and sensorimotor function, demonstrate that the learning rules provide an inductive bias for generalization from a few training trajectories, outperforming the backpropagation-through-time (BPTT) algorithm in a recurrent network which needs an order of magnitude more data. This Lyapunov-based learning can capture low-dimensional latent dynamics underlying training trajectories in high-dimensional embedding. While the backprop algorithm struggles with extrapolation, we show that our model can extrapolate the walking speed when learned from two speeds of human walking data. Importantly, thanks to the extreme sparsity of spiking, this acts as a representational regularizer, forcing the model to not only learn the most parsimonious mode overfitting to a high level of noise in the data. In addition, the learned spiking model carries out robust dynamical computations despite various forms of common perturbation including the failure of a large fraction of neurons. Behavioral dynamic datasets such as human motion capture datasets and large electrophysiological recordings underlying brain functions such as decision-making and motor control can be used to benchmark these spiking models.

We are currently exploring how to incorporate more architectural constraints, such as hierarchy and sparse connectivity, into this framework. We are currently developing an FPGA prototype of the proposed neuromorphic primitive to understand and explore the design space trade-offs and hardware complexity.

Potential applications: The proposed framework can not only be used to reverse engineer cortical dynamics, but it can also be used as an accelerator in high-performance computing to efficiently and robustly capture dynamic rules underlying sequential data, adaptively update it, and simulate the dynamic model with new conditions.

[1] Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M. and Villalobos, P., 2022, July. Compute trends across three eras of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

[2] Boerlin, M., Machens, C.K. and Denève, S., 2013. Predictive coding of dynamical variables in balanced spiking networks. *PLoS computational biology*, *9*(11), p.e1003258.

[3] Denève, S. and Machens, C.K., 2016. Efficient codes and balanced networks. *Nature neuroscience*, *19*(3), pp.375-382.

[4] Eliasmith, C. and Anderson, C.H., 2003. *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. MIT press.

[5] Alemi, A., Machens, C., Deneve, S. and Slotine, J.J., 2018, April. Learning nonlinear dynamics in efficient, balanced spiking networks using local plasticity rules. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

[6] Denève, S., Alemi, A. and Bourdoukan, R., 2017. The brain as an efficient and robust adaptive learner. *Neuron*, *94*(5), pp.969-977.

# Neuromorphic Computing for Neuromophic Cameras

**Harbir Antil, Professor of Mathematics and Director of Center for Mathematics and Artificial Intelligence, George Mason University, Fairfax, Virginia 22030**

**Themes covered.** (1) Neuroscience-inspired computing principles; (3) Modeling and simulation approaches; (4) Data requirements and energy efficiency.

**Introduction.** Event (Neuromorphic) cameras are novel biologically inspired sensors that record data based on the change in light intensity at each pixel asynchronously. If the change in light intensity at a given pixel is larger than a preset threshold then an event (spike, $\pm 1$) is recorded at that pixel. Due to the sparsity of data, events can be recorded on the order of micro-seconds. The events are a non-redundant stream of spikes through the time dimension of each pixel. Due to fine scale and relevant sampling, we can see objects in high contrast environment, fast moving objects subject to motion blur (e.g., hypersonics) or objects in scenarios with limited power or memory.

**Algorithms.** Recently in [2, 3, 1], we have developed innovative optimization algorithms and applications of event based cameras in image segmentation, motion estimation, and image deblurring. The article [2] develops a mathematically rigorous *bilevel optimization framework* for neuromorphic sensors addressing the challenge of reconstructing high-quality images from event-based camera data. The algorithm can simultaneously handle the actual images and neuromorphic data. A result is shown in Figure 1.

This work has been extended in [3] to address image segmentation and motion estimation under a *Generalized Nash Equilibrium*-based optimization framework for the event camera data. This framework capitalizes on the temporal and spatial information derived from the event stream, enabling accurate segmentation and motion estimation. Theoretical foundations are established through the derivation of existence criteria and the proposal of a multi-level optimization method to calculate equilibrium.

A completely new optimization framework to dynamically track objects has been recently introduced in [1]. Each pixel is modeled temporally and we can carry out reconstruction only using the event data without any additional knowledge of events as in [2]. See Figure 2 for two examples.

**Advantages and Outlook.**

- **State-of-the-art.** The articles [2, 3, 1], are the first and only mathematically rigorous models and algorithms currently on neuromorphic imaging.
- **Generic algorithms.** The optimization algorithms introduced are general and they can be applied to other neuromorphic applications such as neuromorphic audio sensors (work is in progress on this) and neuromorphic chips.
- **Scalability.** The above algorithms use the least amount of data at a very fine scale. They can operate either pixel-wise or on a collection of pixels. This makes them scalable. In fact, the reconstruction approach introduced in [1], which only needs event data, can be done in real time.
- **Interdisciplinary collaboration.** We are closely working with remote sensing division at the US Naval Research Lab in Washington DC, BlackSky (satellite company), and United State Air Force Academy (USAFA) to truly understand the needs for development of our mathematically rigorous models and algorithms. See the second example in Figure 2.
- **Hardware acquisition.** Air Force Office of Scientific Research (AFOSR) has awarded us a DURIP project titled: "Neuromorphic Imaging and Digital Twins" to purchase hardware. Eleven iniVation and Prophesee sensors have been procured in June 2024, in addition to two drones. New neuromorphic data is now leading to completely new algorithmic developments.
- **Test case.** Various benchmarks, including data from International Space Station (provided by USAFA), stereo configuration, sensor mounting on drones are currently under construction.

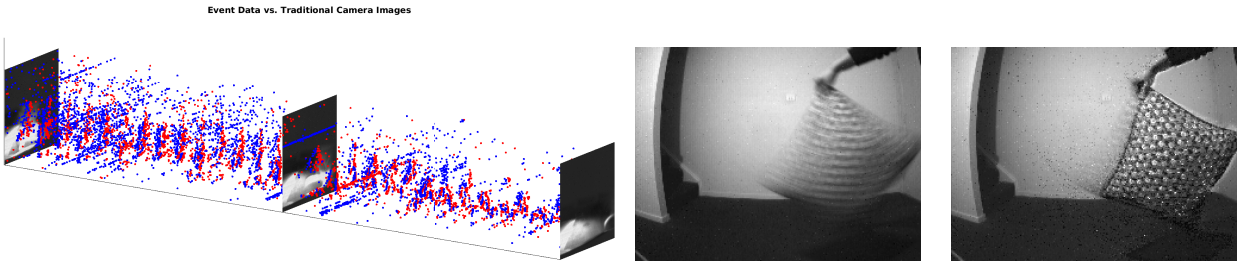



FIGURE 1. **Left:** Comparison of the output of a standard frame-based camera (3 frames) and event camera (dots indicate the events). **Middle:** reconstruction using standard camera. **Right:** reconstruction using our algorithm applied to standard camera data combined with neuromorphic data. Notice the thumb of the person.

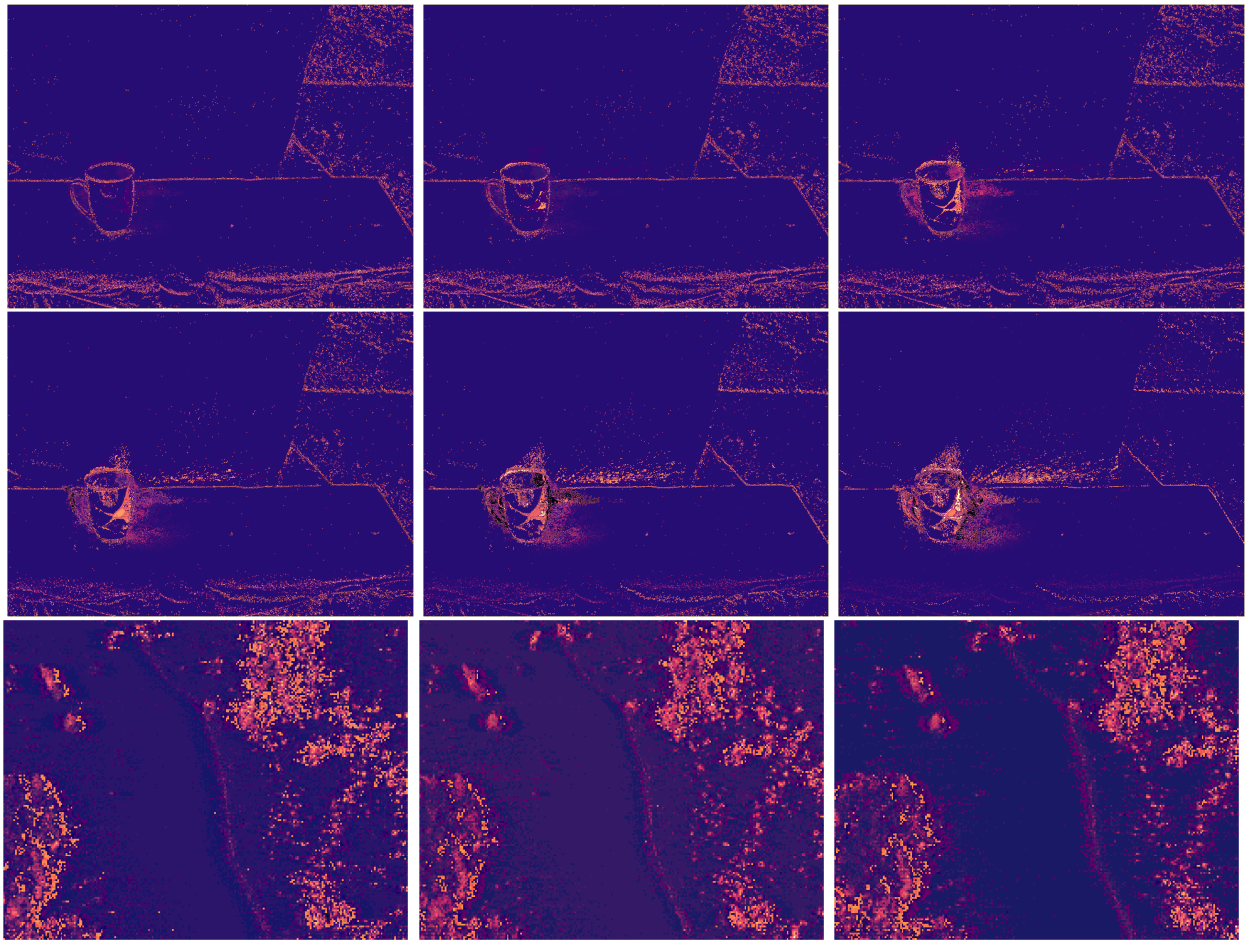


FIGURE 2. **Top two rows:** Dynamic ballistic sequence of cup generated using our algorithm from [1] (with a false color map). **Bottom row:** sequence of earth shots taken by the event camera on International Space Station (data from USAFA).

## REFERENCES

[1] H. Antil, D. Blauvelt, and D. Sayre. Dynamics reconstruction from neuromorphic data. *Submitted*, 2024.

[2] H. Antil and D. Sayre. Bilevel inverse problems in neuromorphic imaging. *Inverse Problems*, 39(9):094003, aug 2023.

[3] H. Antil and D. Sayre. Gnep based dynamic segmentation and motion estimation for neuromorphic imaging. *Foundations of Data Science*, 2024.

# Simulation of Neuromorphic Architectures with Emerging Memory Technologies

Kazi Asifuzzaman
Oak Ridge National Laboratory, Oak Ridge, TN, USA
asifuzzamank@ornl.gov

## 1 Challenge

Advanced software functionalities, specially the ones that build on Artificial Intelligence (AI) and Machine learning (ML) algorithms, require unprecedented compute and memory performance. With *Moore's Law* and *Dennard Scaling* approaching their end, conventional *Von-Neumann* architectures struggle to keep up with the exceeding demands of modern applications. Leveraging recent advancements in material and device-level innovations, researchers are in pursuit to push boundaries with novel technologies (e.g. ReRAM, ECRAM, STT-MRAM etc.) and unconventional computing approaches. One promising avenue of such efforts explore the potentials of Neuromorphic Computing, a special computing paradigm that closely mimics human brain to perform computation, adopting a non-*Von-Neumann* approach, that significantly reduces memory transfer overhead with collocated processing and memory. Neuromorphic architectures are comprised of neurons and synapses, where the programs are defined by neural networks and associated parameters [1]. A Spiking Neural Network (SNN), depicted in Figure 1(b), most closely resembles the behavior of a biological neuron, where it communicates with other neurons through discrete and binary "*spikes*". When a *spike* is communicated to a neuron, the weight is accumulated in its *membrane potential* and it is compared to the threshold at every time step. When the *membrane potential* exceeds the threshold, the neuron fires and resets its value [2]. Figure 1(a) shows an *Integrate-and-Fire (IF)* neuron model. This event-driven characteristics of SNNs allow it to operate in low power.

## 2 Opportunity

Vector Matrix Multiplication (VMM) is the most critical and dominant operation in Artificial Neural Networks (ANNs) including SNNs. Therefore, efficient VMM operations directly contribute towards performance improvement and energy efficiency of neuromorphic systems [3]. The structure of neurons and synapses in SNNs, provides a unique opportunity to naturally map them on crossbar architectures prevalent in emerging memory technologies such as ReRAM, as shown in Figure 1(c). Moreover, these crossbars, by their structures, are inherently capable of performing VMM operations efficiently without any additional compute units. When the input vector for the VMM is fed into the crossbar array as wordlines, and the weight matrix is programmed into the crossbar devices, the current summed at each bitlines represent the resultants of the VMM according to the Kirchhoff's law [4], providing a huge opportunity to harness some of the unique features novel memory technologies offer. However, some of these technologies are susceptible to *non-idealities* such as programming errors, noise error, drift error and array parasitic resistances [5]. For an accurate modeling of neuromorphic architectures realized with emerging memory technologies, it is essential to have robust and realistic simulation infrastructures that faithfully model these memory technologies with their non-idealities in detail. So far, only a handful of studies investigated this promising avenue [6]–[8], and as the development of these technologies are continuously evolving, it is imperative to develop and maintain such simulation infrastructures that take into account all key features, non-idealities and latest advancements in their development.

## 3 Timeliness

Emerging memory technologies have matured in recent years through extensive research and development. It is now crucial to investigate their feasibility and potential for neuromorphic computing through reliable and accurate simulation.

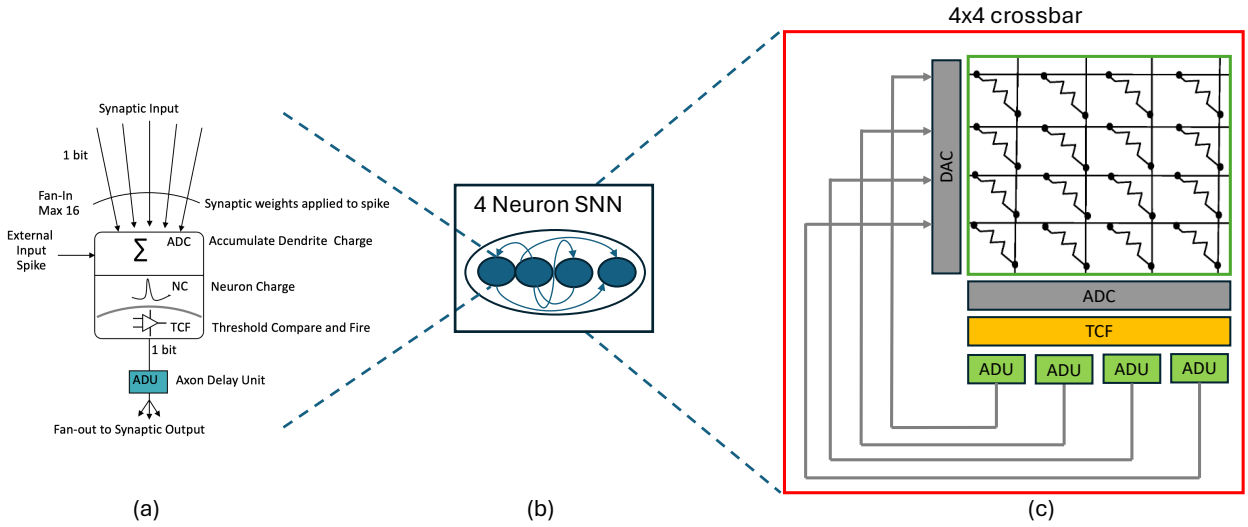



Figure 1: A simplified overview of the concepts discussed in the paper. (a) A generic representation of an *Integrate-and-Fire* neuron model; (b) A conceptual diagram of a Spiking Neural Network consisting four neurons.; and (c) A 4X4 Crossbar structure for a 4-neuron SNN — showing memory cell arrays with Threshold Compare and Fire (TCF), Axonal Delay Unit (ADU), Analog to Digital Converters (ADC) and Digital to Analog Converters (DAC).

## References


[1] Catherine D. Schuman, Shruti R. Kulkarni, Maryam Parsa, et al. "Opportunities for neuromorphic computing algorithms and applications". In: *Nature Computational Science*. 2022.

[2] Sangyeob Kim, Sangjin Kim, Soyeon Um, et al. "Neuro-CIM: ADC-Less Neuromorphic Computing-in-Memory Processor With Operation Gating/Stopping and Digital–Analog Networks". In: *IEEE Journal of Solid-State Circuits* 58.10 (2023), pp. 2931–2945. DOI: 10.1109/JSSC.2023.3273238.

[3] Fernando Aguirre et al. "Hardware implementation of memristor-based artificial neural networks". In: *Nature Communications*. 2024.

[4] Yun Long, Taesik Na, and Saibal Mukhopadhyay. "ReRAM-Based Processing-in-Memory Architecture for Recurrent Neural Network Acceleration". In: *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 26.12 (2018), pp. 2781–2794. DOI: 10.1109/TVLSI.2018.2819190.

[5] T. Patrick Xiao, Ben Feinberg, Christopher H. Bennett, et al. "On the Accuracy of Analog Neural Network Inference Accelerators". In: *IEEE Circuits and Systems Magazine* 22.4 (2022), pp. 26–48. DOI: 10.1109/MCAS.2022.3214409.

[6] Hazan H, Saunders DJ, Khan H, et al. "BindsNET: A Machine Learning-Oriented Spiking Neural Networks Library in Python". In: *Frontiers in neuroinformatics*. 2018. DOI: https://doi.org/10.3389/fninf.2018.00089.

[7] Marcel Stimberg, Romain Brette, and Dan FM Goodman. "Brian 2, an intuitive and efficient neural simulator". In: *eLife* 8 (Aug. 2019). Ed. by Frances K Skinner, Ronald L Calabrese, Frances K Skinner, et al., e47314. ISSN: 2050-084X. DOI: 10.7554/eLife.47314. URL: https://doi.org/10.7554/eLife.47314.

[8] T. Patrick Xiao, Christopher H. Bennett, Ben Feinberg, et al. "CrossSim: accuracy simulation of analog in-memory computing". In: (). URL: https://github.com/sandialabs/cross-sim.

# Cryogenic Neuromorphic Systems: Pioneering a New Frontier

Ahmedullah Aziz, University of Tennessee, Knoxville, TN, USA. E-mail: *aziz@utk.edu*

## 1 Challenge

Current neuromorphic computing systems, inspired by the structure and function of the human brain, hold great promise for revolutionizing artificial intelligence(AI)-based computational tasks such as pattern recognition, sensory processing, and adaptive learning. Conventional CMOS-based neuromorphic hardware faces critical limitations in power efficiency and parallelism [1]. With the rapid increase in demand for highly parallel and energy-efficient computing, CMOS-based platforms are facing even more scrutiny [2]. Superconducting neuromorphic hardware [3], operating at extremely low temperatures, offers a promising alternative by harnessing the zero-resistive characteristics and inherent quantum phenomena to achieve unprecedented levels of energy-efficiency [4]–[6] and speed [7]. Despite the potential benefits, the major challenges include the development of cryogenic-compatible components, managing heat dissipation at such low temperatures, and integrating these systems with existing computing infrastructures [8]. A comprehensive multi-level (materials, device, circuits, and systems) co-design approach and benchmarking framework is necessary to optimize the performance of these emerging technologies for high-performance AI hardware. Additionally, the development of suitable learning algorithms is crucial for further advancement of the field.

## 2 Opportunity

Recent advancements in cryogenic technology have enabled the development of neuromorphic hardware that operates at extremely low temperatures [9]–[11]. These systems leverage superconducting materials to achieve ultra-low power consumption and high-speed processing, making them ideal for highly parallel neuromorphic architectures [12], [13] where efficiency and performance are critical [14]. Notable technologies include superconducting nanowires [15]–[17], quantum phase slip junctions [18], and Josephson junction-based devices [19], which exhibit significantly lower switching energy and ultra-fast response times compared to conventional room temperature devices. Their exceptional efficiency allows for the design of neuromorphic architectures with massive scale and parallelism. Furthermore, these technologies can overcome the limitations of traditional CMOS-based neuromorphic hardware, providing higher integration density, improved scalability, and enhanced energy efficiency. The next steps involve developing algorithms and software specifically tailored for superconducting neuromorphic systems, as well as creating a design automation framework for simulation and benchmarking. ***The biological brain does not need cryogenic temperature to operate. The objective is not to replicate the brain itself but to develop a system that addresses computationally intensive tasks through methods inspired by neural processes.***

## 3 Urgency and Relevance

The soaring energy consumption of data centers and supercomputing facilities is projected to hit critical levels by 2030, driven by the burgeoning demands of artificial intelligence (AI) systems. Recent reports indicate that by 2025, data centers alone could account for up to 8% of the global electricity supply, a substantial increase from the current 3%. Additionally, AI workloads are expected to consume an estimated 300 terawatt hours (TWh) annually [20], rivaling the total energy consumption of major industrial nations. This dramatic escalation underscores the pressing need to explore alternative computing technologies that can deliver high performance while significantly reducing energy consumption. Cryogenic neuromorphic hardware presents a timely and essential solution to this challenge. By harnessing the unique properties of superconducting materials and quantum phenomena, these systems can achieve the high efficiency and performance required to sustain the future of AI hardware in an era of rapidly escalating energy demands.

## References


[1] Min-Kyu Kim, Youngjun Park, Ik-Jyae Kim, et al. "Emerging materials for neuromorphic devices and systems". In: *Iscience* 23.12 (2020).

[2] Felipe Torres, Ali C Basaran, and Ivan K Schuller. "Thermal management in neuromorphic materials, devices, and networks". In: *Advanced Materials* 35.37 (2023), p. 2205098.

[3] Md Mazharul Islam, Shamiul Alam, Md Shafayat Hossain, et al. "A review of cryogenic neuromorphic hardware". In: *Journal of Applied Physics* 133.7 (2023).

[4] Christopher L Ayala, Tomoyuki Tanaka, Ro Saito, et al. "Mana: A monolithic adiabatic integration architecture microprocessor using 1.4-zj/op unshunted superconductor josephson junction devices". In: *IEEE Journal of Solid-State Circuits* 56.4 (2020), pp. 1152–1165.

[5] Shamiul Alam, Md Shafayat Hossain, Srivatsa Rangachar Srinivasa, et al. "Cryogenic memory technologies". In: *Nature Electronics* 6.3 (2023), pp. 185–198.

[6] Shamiul Alam, Md Mazharul Islam, Md Shafayat Hossain, et al. "CryoCiM: Cryogenic compute-in-memory based on the quantum anomalous Hall effect". In: *Applied Physics Letters* 120.14 (2022).

[7] Wei Chen, AV Rylyakov, Vijay Patel, et al. "Rapid single flux quantum T-flip flop operating up to 770 GHz". In: *IEEE Transactions on Applied Superconductivity* 9.2 (1999), pp. 3212–3215.

[8] Michael Schneider, Emily Toomey, Graham Rowlands, et al. "SuperMind: a survey of the potential of superconducting electronics for neuromorphic computing". In: *Superconductor Science and Technology* 35.5 (2022), p. 053001.

[9] Michael L Schneider, Christine A Donnelly, Stephen E Russek, et al. "Energy-efficient single-flux-quantum based neuromorphic computing". In: *2017 IEEE International Conference on Rebooting Computing (ICRC)*. IEEE. 2017, pp. 1–4.

[10] Michael L Schneider, Christine A Donnelly, Ian W Haygood, et al. "Synaptic weighting in single flux quantum neuromorphic computing". In: *Scientific Reports* 10.1 (2020), p. 934.

[11] Jeffrey M Shainline, Sonia M Buckley, Richard P Mirin, et al. "Neuromorphic computing with integrated photonics and superconductors". In: *2016 IEEE International Conference on Rebooting Computing (ICRC)*. IEEE. 2016, pp. 1–8.

[12] Michael L Schneider and K Segall. "Fan-out and fan-in properties of superconducting neuromorphic circuits". In: *Journal of Applied Physics* 128.21 (2020).

[13] Mikail Yayla, Simon Thomann, Md Mazharul Islam, et al. "Reliable Brain-inspired AI Accelerators using Classical and Emerging Memories". In: *2023 IEEE 41st VLSI Test Symposium (VTS)*. IEEE. 2023, pp. 1–10.

[14] Lillian Sharpe, Julia Steed, Md Mazharul Islam, et al. "Impact of Neuron Firing Rate on Application and Algorithm Performance". In: *Proceedings of the 2023 International Conference on Neuromorphic Systems*. ICONS '23. New York, NY, USA: Association for Computing Machinery, Aug. 2023, pp. 1–4. ISBN: 9798400701757. DOI: 10.1145/3589737.3605996. URL: https://dl.acm.org/doi/10.1145/3589737.3605996 (visited on 10/19/2023).

[15] Md Mazharul Islam, Shamiul Alam, Md Shafayat Hossain, et al. "Dynamically reconfigurable cryogenic spiking neuron based on superconducting memristor". In: *2022 IEEE 22nd International Conference on Nanotechnology (NANO)*. IEEE. 2022, pp. 307–310.

[16] Md Mazharul Islam, Shamiul Alam, Nikhil Shukla, et al. "Design Space Analysis of Superconducting Nanowire-based Cryogenic Oscillators". In: *2022 Device Research Conference (DRC)*. IEEE. 2022, pp. 1–2.

[17] Md Mazharul Islam, Shamiul Alam, Catherine D Schuman, et al. "A deep dive into the design space of a dynamically reconfigurable cryogenic spiking neuron". In: *IEEE Transactions on Nanotechnology* (2023).

[18] Ran Cheng, Uday S Goteti, and Michael C Hamilton. "High-speed and low-power superconducting neuromorphic circuits based on quantum phase-slip junctions". In: *IEEE Transactions on Applied Superconductivity* 31.5 (2021), pp. 1–8.

[19] Md Mazharul Islam, Shamiul Alam, Md Rahatul Islam Udoy, et al. "A cryogenic artificial synapse based on superconducting memristor". In: *Proceedings of the Great Lakes Symposium on VLSI 2023*. 2023, pp. 143–148.

[20] Christopher Metz. "Towards sustainable artificial intelligence systems: enhanced system design with machine learning based design techniques". PhD thesis. Universität Bremen, 2024.

# Energy efficient transformer architecture for LLMs based on Spiking Neural Networks

Adarsha Balaji[1] (abalaji@anl.gov, corresponding author)
Sandeep Madireddy[1]
[1]*Argonne National Laboratory*

## Challenge

Recent advances in Artificial Intelligence (AI) have transformed our approach to solving scientific problems using AI foundational models for use in the fields of material science, cancer research and climate change. However, in order to efficiently train and deploy then models, there exists a trade-off between the cost, throughput and computational complexity that often determines the scalability, hardware choice, parallelism strategy, latency, and throughput of these models. These foundational models are traditionally designed using the transformer architecture, a sequence-to-sequence model based on the multi-headed self attention mechanism (SA). However, the transformer is a memory and computation intensive block, leading to the need for expensive and slow AI hardware (GPUs or custom accelerators) to train and infer these models. For instance, the 175-billion parameter GPT-3 model, designed by OpenAI, is estimated to need $\sim$ 175 years to train using a single Tesla V100 GPU. To address this, we explore spiking neural network (SNNs) based transformers, a computing paradigm inspired by the biological concepts of the mammalian brain, which provides a number of advantages for data-, energy-, and resource-efficient execution of such foundational models.

Spiking Neural Networks (SNNs) [1] are an energy efficient alternative to traditional neural networks when executed on neuromorphic hardware. The mainstream approaches to design large-scale SNN are ANN-SNN conversion and direct training SNN. However, training large-scale SNN, like Transformer architectures, using existing surrogate learning methods is inefficient and time-consuming. ANN-SNN conversion techniques are not scalable and are only able to achieve optimal performance at the cost of a large number of time-steps, i.e. increased latency. To address this, a methodology to design large-scale transformer-based SNN using the principles of transfer learning and knowledge distillation with existing ANN-SNN conversion methods. The methodology works in three steps: (1) conversion of trained ANN into SNN, and (2) replacing ANN-based self-attention (ASA) mechanism with a SNN-based self attention (SSA), and (3) fine-tuning the SSA block using SNN-based surrogate learning algorithms.

## Opportunity

In the transformer architecture, the vanilla self-attention (VSA) mechanism transforms an input sequence into an attention map. The VSA takes the Query (Q), Key (K) and Value (V) vectors as input and perform three operations: matrix multiplication (dot product), scale and softmax activation. The dot product and softmax activation operations cannot be readily implemented in SNNs due to the discrete (spike) nature of SNN activations. Therefore, existing transformer architectures are not accurately realizable in SNNs. To address this, we propose spike-based transformer architecture (STA), based on the model architecture that replaces the complex matrix and activation operations of the VSA with modified Hadamard and sparse addition operations. The hadamard operator performs a low-power binary multiplication operation (bit-level AND gate) and replaces the computationally complex and expensive dot-product operation used in the VSA mechanism.

Training large-scale transformer models based on the STA is still a challenge, due to the limitations of existing spike-based learning algorithms. This limitation is attributed to the non-differentiable nature of SNN activations (spikes) and the limited ability of surrogate methods to

compute gradients in SNNs. With the limitations of SNN learning algorithms to train large-scale SNNs and the extreme compute resources needed to train transformer-based foundational models, we propose a methodology to design SNN-based foundational models by exploiting ANN-SNN conversion methods to convert trained transformer-based models to SNNs followed by a knowledge distillation-based fine-tuning of the attention layers of the network to minimize the conversion loss of the network.

**ANN-SNN conversion** The basic principle of converting ANNs into SNNs is that the firing rates of spiking neurons match the graded activations of analog neurons. To achieve this, a relation is established between the ANNs using ReLU activation and the SNN integrate-and-fire (IF) neuron. The ReLU can be considered a firing rate approximation of an IF neuron with no refractory period, whereby the output of the ReLU is proportional to the number of spikes produced by an IF neuron within a given time window. The first step is to replace the ReLU-based ANN neurons in the feed-forward block with the Integrate-and-Fire (IF) neuron. To ensure high conversion accuracy, the Poisson spike rate of the IF neurons is expected to be proportional to the activation of its respective ANN neurons. This can be achieved by finding the right balance of IF neurons thresholds, input weights and input firing rates.

**Fine-Tuning Spiking Self-Attention Block** Step 1 and 2 of the proposed methodology involve the ANN-SNN conversion of the fully connected layers of the network and replacing the traditional analog self attention (ASA) mechanism with the proposed spiking equivalent, respectively. However, the outcome of the proposed spiking self-attention (SSA), as shown in equation 2, does not equate to the expected outcome of the ASA, as shown in equation 3. Therefore, we initialize the weights of the SSA with the trained weights from the ASA to ensure a well-initialized SNN that can be fine-tuned. We explore a surrogate gradient function ($spike_{grad}$) [2], to fine-tune the SSA block of the network, while freezing the weights of the embedding and the feed forward block of the network.

$$AttentionMap(AM) = LIF((Q \otimes K^T)_{Columnwise}) \tag{1}$$

$$SSA(Q, K, V) = (AM \otimes V) \tag{2}$$

$$ASA(q_n, K, V) = \sum_{m=1}^{M} \frac{exp(q_n^T.k_m)}{\sum_{m=1}^{M} exp(q_n^T.k_m)} \cdot v_m^T \tag{3}$$

**Implementation of Spiking Self Attention** The input to the self-attention block are the spike-encoded sequence (S) generated from the input encodings $I \in [0, 1]^{T \times N}$, where T is the time duration of the input spike train and N is the sequence length of the input. The query (Q), key (K), and value (V) matrices $\in \mathbb{R}^{T \times N \times D}$, where D ($d_{q,k,v}$) is the model dimension hyperparameter, are computed by performing a linear transformation using three learnable matrices ($W_q$, $W_k$ and $W_v$), respectively, followed by a layer of spiking neurons (LIF) to generate the output spike response. The spiking self-attention is computed by performing a column-wise ($d_k \times 1$) hadamard operation between the $Q$ and $K^T$ vectors, as shown in equation 1, followed by a spiking IF activation. The output of the SSA block is computed as the hadamard product of the attention map generated using equation 1 and the value (V) vector, as shown in equation 2.

# References


[1] Yi, Zexiang, et.al. Learning rules in spiking neural networks: A survey, Neurocomputing 2019.

[2] Neftci, Emre O, et al. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks, IEEE Signal Processing Magazine, 2019

# Hardware-aware continual learning on neuromorhpic hardware

Adarsha Balaji[1] (abalaji@anl.gov, corresponding author)
Sandeep Madireddy[1], Prasanna Balaprakash[2]
[1] *Argonne National Laboratory,* [2] *Oak Ridge National Laboratory*

## Challenge

Scientific experiments at the DOE experimental facilities such as Argonne's Advanced Photon Source (APS) and Oak Ridge's Spallation Neutron Source (SNS) require the transfer of large datasets from the detectors to supercomputers in order to perform data analysis. While the data acquisition step in these workflows is relatively fast, the lack of access to on-demand computing required to process greater than GB/s streaming data from x-ray detectors, at the edge, presents significant *time, energy, and throughput challenges* to scientific discovery. Neuromorphic computing (NmC) based on spiking neural networks (SNNs), a computing paradigm inspired by the biological concepts of the mammalian brain, provides a number of advantages for data-, energy-, and resource-efficient machine learning at the edge. Hardware implementations of NmC, executing SNNs, enable large-scale data analysis beyond what is feasible with the emerging high-performance-computing-to-edge computing paradigm. This key advantage can be attributed to the low-power design of the underlying computing circuits, the distributed implementation of its compute and storage, and novel technology integration in the form of non-volatile memory-based neuro-synaptic cores. In order to deploy the NmC capabilities at the facilities such as the APS, however, SNNs need to continually adapt to variations in the data and the NmC hardware. Existing state-of-the-art SNN-based continual learning (SCL) algorithms are not tolerant to process, operational, and reliability variations in NmC hardware. To overcome these challenges, there is a need for an SCL framework to train data-, energy-efficient, and hardware-fault-tolerant SNN-based models on NmC hardware.

## Opportunity

To deploy NmC hardware capabilities at the edge, SNN learning algorithms need to continually adapt to variations in experimental data and the NmC hardware. However, existing NmC systems suffer from two key issues making them inapplicable for scientific applications. First, SNNs lack efficient and scalable learning algorithms [1] and thus are unsuitable for efficient continual learning at the edge. Second, energy-efficient implementations of NmC hardware are extremely sensitive to process, operating (voltage and temperature), and fault-induced variations [2], hindering the performance of the SNN in terms of learning and inference accuracy. An SCL framework will address these limitations and allow for the deployment of SCL on NmCs at the edge. The proposed framework will support the execution of the SCL on (1) a custom, mixed-signal implementation of NmC, (2) existing digital implementations of NmCs, such as Intel's Loihi [3], and (3) an analog implementation of NmC [4]. This position paper describes the need for SCL on energy-efficient NmC hardware at the edge, allowing for accelerated discoveries at the experimental facilities.

Continual learning (CL) enables a model to learn tasks in a sequential fashion and adapt to a shift in data, without the loss of existing knowledge. However, because of the phenomenon of catastrophic forgetting (CF), models trained for an initial task $T_A$ tend to "forget" information about $T_A$ while learning a new task $T_B$. Since learning in the mammalian brain does not suffer from CF, researchers have taken inspiration from bio-inspired learning on spiking neural networks to improve memory retention.

**Neuromorphic Continual Learning** Develop regularization-based and network architecture-based approaches to prevent CF in SNNs. Regularization will involve a surrogate gradient descent

learning rule, while architecture-based methods will leverage sparse synaptic connections to minimize CF.

*Regularization Approach* - a surrogate gradient descent learning rule can be adopted to train the network, such as event-driven backpropagation. To address CF while learning a new task $T_B$, changes to synaptic connections (weights) that significantly influence the performance of the network for task $T_A$ should be minimized. First, the influence of a synapse on task $T_A$ is determined by measuring its contribution to task $T_A$, during the training process. Next, each synapse is now assigned a new *plasticity parameter* p - {0, 1}. The value of p controls the plasticity of the synapse, its degree of freedom to learn/change the higher the value of $p$, and its ability to learn/adapt to the new task. The value of p is modulated by using a reward-modulated spike-time-dependent plasticity (R-STDP) learning rule.

*Network Architecture Approach* - Prior work [5] studying the impact of the network architecture on CF hypothesizes that when overlapping regions/synapses in the network are used to represent multiple tasks, the process contributes heavily to CF. Exploiting the sparse nature of synaptic connections in SNNs by periodically pruning redundant synapses to a specific task (T) and exploring and encouraging task-based sub-networks, to preserve the overall performance of the network.

**Fault-Tolerant Learning** Design a fault-tolerant learning algorithm using hardware fault injections during training. This involves creating a fault modeling framework (FMF) for accurate hardware-level fault simulations and a fault injection framework (FIF) to integrate these faults into the learning process. A hardware-software co-optimization framework is needed to model, simulate, extract and inject the hardware faults, during application training time. The framework can study the impact of hardware variations, such as resistance drift, on the application accuracy and allow for the variations to be injected during the training phase of the model.

**Evaluation on State-of-the-Art Hardware** Test the proposed SCL algorithms on existing NmC hardware, such as Intel's Loihi and Rain Neuromorphics' analog implementations. Additionally, design a custom mixed-signal NmC hardware for comprehensive evaluation of the SCL framework.

**Hardware Implementation** Develop a mixed-signal neuromorphic integrated circuit design that combines digital and analog elements for efficient SNN execution. The hardware platform will have to combine a digitally implemented integrate-and-fire neurons with an analog ReRAM crossbar-based synaptic element. The neurosyaptic core includes necessary peripheral circuits, such as analog-to-digital and digital-to-analog spike converters and memory units and supports backpropagation-based global learning (GL) and bioinspired local learning (LL) rules. For the complex computations of GL, we will integrate a fully digital von Neumann CPU to perform global weight updates.

# References


[1] Li et al. Spiking neural network learning, benchmarking, programming and executing, 2020.

[2] Hu et al. In *2018 ITC-Asia*, 2018.

[3] Davies et al. Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro*, 38(1):82–99, 2018.

[4] Kendall et al. Training end-to-end analog neural networks with equilibrium propagation. arxiv, 2020.

[5] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999.

Position Paper

# 2024 Neuromorphic Computing for Science Workshop

## *HPC and Accelerator-driven Scaling up of Neuromorphic Models and Simulations*

by Srutarshi Banerjee, Data Science and Learning Division, Argonne National Laboratory, IL, USA

Modeling and simulations have been an integral part of neuroscience research for several decades. To mimic the performance of human brain, neuromorphic simulations involving complex neurons, synapse models, density of neurons, and other realistic biological parameters, neural behaviors at scale of human brain is required. This is fueled by the large-scale data collection from structural and functional imaging tools [1]. Existing models and simulators [2] (such as NEST [3] and others) lack these capabilities. To harness the full potential of neuromorphic computing, the need to scale the existing Spiking Neural Network (SNN) models to hundreds of billions of parameters is necessary. High performance computing and accelerator driven scale-up of the models and simulations using existing hardware (for example GPUs) is one of important research directions aimed in this position paper.

The most dominant method for training neural networks using backpropagation [4] and other evolution-based algorithms [5] have created several SNN algorithms which are able to perform vision [6], tactile sensing tasks [7], and others. Despite these progresses, there has not been a significant progress in the scale up of these models – not anywhere close to the scale of human brain. The challenges in this scaling up process is how to effectively use existing hardware such as GPUs, CPUs, Accelerators, FPGAs and how to simulate the synapse interactions between billions of neurons.

The focus of the position paper is to explore the feasibility of scaling up the models using reconfigurable computing paradigm with on-demand compute resources needed by the algorithm. Basic progress has been made mostly in the material and device level [8], however, without the successful development of reconfigurable algorithms, the scaling up is not feasible. To pave way for performing massive parallel computation as in our human brain, the model building must consider the available memory and associated compute available – which can be optimized using Neural Architecture Search (NAS) framework. Although some works in literature has optimized the SNN model based on performance and memory available [9], [10], we argue that the NAS framework has a vital role to play not only in the scaling up of the models, but also in determining the optimal compute, memory and performance of the scaled-up model. Additional scaling up of compute using multiple processes and multiple threads in one or multiple nodes is envisaged.

While the scaling up of the neuromorphic computing models not only involves an efficient SNN architecture, but there is also an associated training effort which is of utmost importance. In that regard, the use of novel learning rules and strategies is of importance. The need of huge memory and compute resources in the traditional backpropagation-based learning methodology is inefficient in scaling up of the SNN models. While other training strategies – Surrogate Gradient Descent, Real-time Recurrent Learning [11], Meta-learning [12], and others, have been tried with the SNN models, there is no clear vision of training strategy when it comes to the effective scaling up. The prospects of Meta-learning or ‘learning-to-learn’ framework for training during the scaling up of the SNN models from fewer trainable weights to larger trainable weights is one of the interesting research directions to explore. Additional meta-learning optimizations can be performed on the considerations of available memory, compute resources and process communications footprint during the training process.

## References


[1] Wang, Felix, Shruti Kulkarni, Bradley Theilman, Fredrick Rothganger, Catherine Schuman, Seung-Hwan Lim, and James B. Aimone. "Scaling neural simulations in STACS." *Neuromorphic Computing and Engineering* 4, no. 2 (2024): 024002.

[2] Kulkarni, Shruti R., Maryam Parsa, J. Parker Mitchell, and Catherine D. Schuman. "Benchmarking the performance of neuromorphic and spiking neural network simulators." *Neurocomputing* 447 (2021): 145-160.

[3] Gewaltig, Marc-Oliver, and Markus Diesmann. "Nest (neural simulation tool)." *Scholarpedia* 2, no. 4 (2007): 1430.

[4] Eshraghian, Jason K., Max Ward, Emre O. Neftci, Xinxin Wang, Gregor Lenz, Girish Dwivedi, Mohammed Bennamoun, Doo Seok Jeong, and Wei D. Lu. "Training spiking neural networks using lessons from deep learning." *Proceedings of the IEEE* (2023).

[5] Stanley, Kenneth O., Jeff Clune, Joel Lehman, and Risto Miikkulainen. "Designing neural networks through neuroevolution." *Nature Machine Intelligence* 1, no. 1 (2019): 24-35.

[6] Li, Yuhang, Ruokai Yin, Youngeun Kim, and Priyadarshini Panda. "Efficient human activity recognition with spatio-temporal spiking neural networks." *Frontiers in Neuroscience* 17 (2023): 1233037.

[7] Kang, Peng, Srutarshi Banerjee, Henry Chopp, Aggelos Katsaggelos, and Oliver Cossairt. "Boost event-driven tactile learning with location spiking neurons." *Frontiers in Neuroscience* 17 (2023): 1127537.

[8] Xu, Minyi, Xinrui Chen, Yehao Guo, Yang Wang, Dong Qiu, Xinchuan Du, Yi Cui, Xianfu Wang, and Jie Xiong. "Reconfigurable neuromorphic computing: Materials, devices, and integration." *Advanced Materials* 35, no. 51 (2023): 2301063.

[9] Kim, Youngeun, Yuhang Li, Hyoungseob Park, Yeshwanth Venkatesha, and Priyadarshini Panda. "Neural architecture search for spiking neural networks." In *European conference on computer vision*, pp. 36-56. Cham: Springer Nature Switzerland, 2022.

[10] Putra, Rachmad Vidya Wicaksana, and Muhammad Shafique. "Spikenas: A fast memory-aware neural architecture search framework for spiking neural network systems." *arXiv preprint arXiv:2402.11322* (2024).

[11] Neftci, Emre O., Hesham Mostafa, and Friedemann Zenke. "Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks." *IEEE Signal Processing Magazine* 36, no. 6 (2019): 51-63.

[12] Stewart, Kenneth M., and Emre O. Neftci. "Meta-learning spiking neural networks with surrogate gradient descent." *Neuromorphic Computing and Engineering* 2, no. 4 (2022): 044002.

# Neuronal Dynamics in Neuromorphic Systems with Oscillations

Frank Barrows
Los Alamos National Laboratory

Artificial neural networks struggle with computing time series tasks and predictions, these same tasks are where biological neural networks excel.[1] The brain processes time varying information using neuronal dynamics, networks of neurons that evolved to handle such input. Recapitulating this activity in analog neuromorphic materials can enable novel functionalities and computational capabilities.

Most existing neural networks do not leverage neuronal-like dynamics. At the level of a single neuron, phenomena including action potential propagation, refractory periods, and signal integration arise from the simple dynamics of ion channel gates.[2] At the network level, models of neural populations such as Wilson-Cowan model capture aspects of communication across the brain, e.g, dynamics of the cortical-thalamic axis, and oscillatory activity resulting from competing excitatory and inhibitory connections.[3, 4] Importantly, neuromorphic systems and materials have their own inherent dynamics that can be leveraged in neuromorphic computers. These dynamics include the evolution under external forcing, the material response to perturbations and relaxation dynamics.[5, 6] These can be used to devise training approaches and improve our understanding of how to embed time-varying information. Dynamical neuromorphic networks are an essential step in designing functionally equivalent neuromorphic systems.

We propose using oscillating functional units to implement local dynamical learning through two approaches:

**Implement Oscillations:** Oscillating neuromorphic units would capture an essential function of the brain. In the brain oscillations influence information transfer, mediate interactions between structures, e.g., gamma oscillations in the cortico-thalamic loops, and relate attention to signaling between the primary visual cortex and the lateral geniculate nucleus.[7] Oscillating neuromorphic units include memristor based LRC, LR, and RC circuit motifs,[8–10] Figure 1 (a) depicts such motifs. This goes beyond neuromorphic spiking networks, oscillating neurons can represent inhibitory signals, acting as a proxy for adversarial training. Developing tunable oscillating units requires co-designing training algorithms, control mechanisms, and motif couplings to ensure stability and functionality. A proposed coupling scheme is shown in Figure 1 (b).

**Leveraging Neuromorphic Dynamics for Computation:** To implement local dynamic learning we must exploit the inherent dynamics of active neuromorphic systems. This involves developing algorithms that utilize these dynamics for learning, integrating both phase and amplitude in training, while controlling the network such that it remains stable. The underlying assumption is that a unified mechanism in the brain can implement learning and network control. Preliminary studies suggest that continuous learning can effectively manage oscillations,[11] with local spike-timing-dependent plasticity (STDP) kernels naturally arising in models of coupled oscillating neurons with signal integration. Dynamical neuromorphic algorithms can leverage these local energy based rules for training. These training rules could be implemented in hardware by designing the coupling between circuit motifs, thus enabling analog learning in physical neural networks.

By focusing on dynamic networks and oscillating units, we can enhance our understanding of how time series information is embedded and computed in neural systems. Embracing these dynamics offers a pathway to more faithful and reliable representations in neuromorphic computing, ultimately bridging the gap between artificial and biological intelligence.

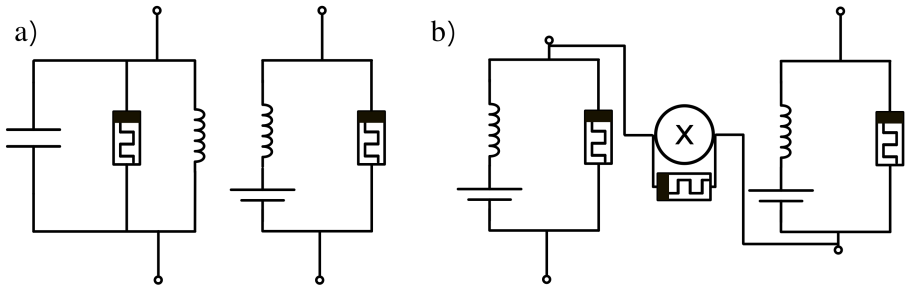



Figure 1: (a) Two proposed neuromorphic oscillating circuit motifs with memristors. (b) Schematic of coupled oscillating circuit motifs, coupling consists of a trainable weight implemented by a memristor linked to a functional element, e.g., a signal convolution element, filter, op-amp, or ReLU.

# 1 Bibliography

# References

[1]M. Schneider, N. Greifzu, L. Wang, C. Walther, A. Wenzel, and P. Li, "An end-to-end machine learning approach with explanation for time series with varying lengths", Neural Computing and Applications **36**, 7491–7508 (2024).

[2]A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve", The Journal of Physiology **117**, 500–544 (1952).

[3]P. Fries, "Rhythms for cognition: communication through coherence", Neuron **88**, 220–235 (2015).

[4]J. Zang, S. Liu, P. Helson, and A. Kumar, "Structural constraints on the emergence of oscillations in multi-population neural networks", eLife **12**, edited by T. Tchumatchenko and L. L. Colgin, RP88777 (2024).

[5]F. Caravelli, F. L. Traversa, and M. Di Ventra, "Complex dynamics of memristive circuits: analytical results and universal slow relaxation", Physical Review E **95**, 022140 (2017).

[6]S. Kumar, X. Wang, J. P. Strachan, Y. Yang, and W. D. Lu, "Dynamical memristors for higher-complexity neuromorphic computing", Nature Reviews Materials **7**, 575–591 (2022).

[7]V. L. Mock, K. L. Luke, J. R. Hembrook-Short, and F. Briggs, "Dynamic communication of attention signals between the lgn and v1", Journal of Neurophysiology **120**, PMID: 29975169, 1625–1639 (2018).

[8]X. Liao and N. Mu, "Self-sustained oscillation in a memristor circuit", Nonlinear Dynamics **96**, 1267–1281 (2019).

[9]V. Rajamani, C. Yang, H. Kim, and L. Chua, "Design of a low-frequency oscillator with ptc memristor and an inductor", International Journal of Bifurcation and Chaos **26**, 1630021 (2016).

[10]F. Barrows, F. C. Sheldon, and F. Caravelli, "Network analysis of memristive device circuits: dynamics, stability and correlations", (2024), arXiv:2402.16015.

[11]M. Ernoult, J. Grollier, D. Querlioz, Y. Bengio, and B. Scellier, "Equilibrium propagation with continual weight updates", (2020), arXiv:2005.04168.

**Control of brain dynamics & learning as a computational primitive for neuromorphic computing**

Kris Bouchard, Staff Scientist & Lead, Computational Bioscience Group, Scientific Data Div., LBNL;
Adj. Prof, Dept. of Neuroscience and Redwood Center for Theoretical Neuroscience, UC Berkeley.

**Specific Challenge**: Current AI models compute with non-neurobiological attention mechanisms and require vast volumes of data for training; together, these characteristics impede low-power, real-time, flexible control and adaptation. In contrast, computations in the brain emerge from dynamics across neural populations, which manifestly enables low-power, real-time, flexible control and adaptation and learning from few examples. Furthermore, neuroanatomical feedback loops are present both within and between brain areas, and the same neural populations can perform diverse functions on demand. **Together, these observations indicate that feedback control is a computational principle of neural population dynamics that enables efficient on-line learning**. While it is recognized that brain computations emerge from dynamics which must be controlled with feedback, efforts to instantiate this principle into neuromorphic hardware are nascent with many challenges.

**Outcome:** Addressing this challenge would result in novel, co-designed, hardware-accelerated, neuromorphic systems based on computations through neural population dynamics with feedback control based on brain-derived principles, redefining the core mechanisms by which computations are instantiated.

**Broader Impact:** The voracious energy consumption of current computing architectures presents a clear challenge for both next generation HPC systems and edge computing in energy deprived environments or robotic systems. Furthermore, this energy consumption contributes to the existential threat of climate change. Neuromorphic systems that compute with dynamics instantiated with next-generation materials provides a viable low-energy computing alternative. This research represents a critical first step away from the 'brain inspired' dogma of current neuromorphic computing and towards 'brain derived' neuromorphic systems deployed in robotic systems. The paradigm shift we propose will open new research directions at the intersection of computer science, material science, neuroscience, and AI/ML.

**Background:** Current AI models utilize the attention mechanism as a core computational principle. In contrast, diverse brain functions, ranging from perception (e.g., facial recognition) to cognition (e.g., navigating an environment) to action (e.g., reaching to targets) are produced by collective and emergent dynamics of populations of neurons distributed across the brain. We have advanced the theory that functionally relevant brain dynamics are feedback controllable and tested this theory in electrophysiological recordings. In diverse neural datasets from across the brain (hippocampus, primary somatosensory and motor cortex, high-level visual areas), we developed methods to show that feedback controllable (FBC) 'directions' of high-dimensional neural population activity robustly outperform feedforward controllable (FFC) directions in reaching/location/face decoding, indicating that FBC is a computational primitive of diverse neural circuits. FBC and FFC subspaces emerged from collective interactions of a population of neurons with distinct activity profiles that map well to distinct populations of L5 pyramidal neurons. Finally, we showed analytically that the divergence between FBC and FFC subspaces depends on the degree of non-normality in neural dynamics, which results from the neuroanatomical constraint imposed by the fact that biological neurons are either excitatory or inhibitory, but not both. ***Together with other results, this indicates that feedback control is a well-defined computational primitive that influences behavior, cognition, and learning, and is implemented by specific neurons and neurobiological circuit principles.***

**Potential Approaches, Data Sets, & Metrics**: Outstanding computational research includes: 1) identifying 'special' high-order network motifs that enhance circuit controllability from mouse/fly connectomes, 2) HPC simulations of FBC of neurobiologically constrained spiking RNN dynamics through neuronal response modulation (a neurobiologically understood phenomena) with Hebbian plasticity to develop a simple (in RNNs with $N$-neurons, this would require $\sim N$ directly controlled parameters vs. direct optimization of synaptic weights in AI RNNs $\sim N^2$ # synapses) online adaptation and learning algorithm, 3) ensuring that neuromorphic circuits exhibit similar dynamics (e.g., rotations) as observed in experimental data *in vivo,* 4) exploring neuromorphic hardware codesign leveraging dynamicity of emerging materials.

**Outlook**: For large-scale RNNs, FBC provides a clear computational primitive for developing/simulating circuits employing electro-optical components and phase-change materials in a codesign paradigm towards deployment in future HPC systems, at the edge, or in robotic systems.

# Dendritic Computation: Routing Neuroscience to Neuromorphic Circuits

Suma G. Cardwell[1] and Frances S. Chance
Center for Computing Research, Sandia National Laboratories

**Primary Theme: Neuroscience algorithms and translation to neuromorphic analog circuits**

## Neuroscience-inspired computing principles

Neuroscience insights have long informed and inspired the design of neuromorphic systems [1] . This interdisciplinary approach has enabled the development of novel computing technologies but also holds promise to advance our understanding of the brain. However, while the neuroscience field continues to grow with tremendous explosion of new data, the neuromorphic field is still lagging in its translation. For example, the fundamental computational unit in a brain is the neuron. A single neuron is quite complex in biology across species from invertebrates to mammals. The structure of dendrites within a neuron are often quite intricate, vary widely and depend on the neuron's specific sub-type and location within a brain, and can be specialized to support the function of the neuron [2,3,4]. However, modern neuromorphic and artificial neural networks adopt a much simpler and homogeneous model of the neuron leading to increased reliance on scaling. Our hypothesis is that thoughtful consideration of the complexity and heterogeneity of a single neuron (specifically with dendrites) can lead to the design of smaller systems with increased computational complexity.

## Translation to analog microelectronic circuits

Modeling dendrites in analog circuits is crucial to getting the computational density (operations/unit) as well as computational efficiency (energy/operations/second). Dendrites within a neuron can be thought of as a "neural network within a single neuron" [5] enabling compute-on-wire. Current architectures focus solely scaling number of neurons and synapses per neuron. Most emerging non-volatile devices can be used to construct dendrites offering a "dendritic toolkit" which include non-linear filtering, spatio-temporal processing, gain modulation, coincidence detection while offering dense connectivity [6,7,8]. In prior work, we have demonstrated the efficacy of analog dendrites using the following circuits:

1) Direction-Selective Dendrite Circuits (Pattern Recognition [9], Near-sensor processing [9]), Dragonfly TSDNs (Target Selective Descending Neurons)[10]), Fruit Fly Visual ON Circuit[11])
2) Dendrites for Gain modulation (Dragonfly Coordinate Transformation for Interception [9])
3) SNNs with Dendrites (Coincidence Detection[12], Non-linear functions[12], Pooling layers[11]).

These circuit motifs can be used to demonstrate large-scale dendrite-based neural networks. Emerging devices as well as novel approaches in fabrication like 3D architectures, wafer-scale technology, and in-memory computing could further alleviate current communication and connectivity bottlenecks.

## Modeling and simulation approaches

Existing neural simulators such as NEURON, BRIAN 2, and NEST can be used to model large-scale SNNs. To incorporate dendrites in SNNs, we can use Dendrify as well as our own dendrite library that models hardware-based parameters for snnTorch [12]. We have also developed a neuromorphic architecture simulator called SANA-FE [13] with plans to incorporate multiple emerging hardware-based dendrites and enable design space exploration of novel dendrite-based architectures.

## Performance metrics, data requirements, and energy efficiency

Dendrites can be evaluated based on performance (operations/second) as well as energy efficiency (energy/operation/second). Other metrics include accuracy, latency, area, and ease-of-integration. Ideally, analog datasets should be created (much like event camera datasets) for evaluation and benchmarking various hardware-based approaches.

[1] Corresponding Author

## Acknowledgments

*SNL is managed and operated by NTESS under DOE NNSA contract DE-NA0003525*

## References

[1] Mead, Carver. "How we created neuromorphic engineering." *Nature Electronics* 3, no. 7 (2020): 434-435.
[2] Stuart, Greg, Nelson Spruston, and Häusser Michael, eds. *Dendrites*. Oxford University Press, 2016.
[3] Segev, Idan. "Sound grounds for computing dendrites." *Nature* 393, no. 6682 (1998): 207-208.
[4] Poirazi, Panayiota, and Athanasia Papoutsi. "Illuminating dendritic function with computational models." *Nature reviews neuroscience* 21, no. 6 (2020): 303-321.
[5] Ramakrishnan, Shubha, Richard Wunderlich, Jennifer Hasler, and Suma George. "Neuron array with plastic synapses and programmable dendrites." *IEEE transactions on biomedical circuits and systems* 7, no. 5 (2013): 631-642.
[6] London, Michael, and Michael Häusser. "Dendritic computation." *Annu. Rev. Neurosci.* 28, no. 1 (2005): 503-532.
[7] Boahen, Kwabena. "Dendrocentric learning for synthetic intelligence." *Nature* 612, no. 7938 (2022): 43-50.
[8] D'Agostino, Simone, Filippo Moro, Tristan Torchet, Yiğit Demirağ, Laurent Grenouillet, Niccolò Castellani, Giacomo Indiveri, Elisa Vianello, and Melika Payvand. "DenRAM: neuromorphic dendritic architecture with RRAM for efficient temporal processing with delays." *Nature communications* 15, no. 1 (2024): 3446.
[9] Cardwell, Suma G., and Frances S. Chance. "Dendritic computation for neuromorphic applications." In *Proceedings of the 2023 International Conference on Neuromorphic Systems*, pp. 1-5. 2023.
[10] Cardwell, Suma G., Mark Plagge, Luke Parker, Claire Plunkett, David Munkvold, Paloma Gonzalez-Bellido, Scott Koziol, Conrad James, "Compute-On-Wire: Leveraging Dendritic Complexity for Neuromorphic Computing", *Neuromorphic IOP, In Review*
[11] Parker, Luke, Suma Cardwell, Frances S. Chance, and Scott Koziol, "Bio-Inspired Active Silicon Dendrite for Direction Selectivity." In *Proceedings of the 2024 International Conference on Neuromorphic Systems*, 2024.
[12] Plagge, Mark, Suma George Cardwell, and Frances S. Chance. "Expressive Dendrites in Spiking Networks." In *2024 Neuro Inspired Computational Elements Conference (NICE)*, pp. 1-8. IEEE, 2024.
[13] Boyle, James, Mark Plagge, Suma George Cardwell, Frances S. Chance, and Andreas Gerstlauer. "Performance and energy simulation of spiking neuromorphic architectures for fast exploration." In *Proceedings of the 2023 International Conference on Neuromorphic Systems*, pp. 1-4. 2023.

# More than Spikes: Neurons as Dynamical Systems for Intracellular Processing

William Chapman
Sandia National Laboratories

# More than Spikes: Neurons as Dynamical Systems for Intracellular Processing

**Introduction** Traditional computing systems utilize the computational primitive of Boolean logic, operations on a population of bits, and a synchronous operating clock, to enable logical and symbolic operations. Largely influenced by analogy to this approach, recent developments in neuromorphic computing have focused on the action potential, or spike, as the fundamental unit of neurally inspired computing. This has led to a focus on population-level encoding approaches, in which simplified units such as the leaky-integrate-and-fire (LIF) are utilized as universal function approximators, and largely implements 'spiking equivalents' of traditional computer algorithms [1], at the cost of requiring thousands of units and hand-engineered algorithms. In contrast, biological neurons are nonlinear units, with a plethora of intracellular dynamics which directly operate on inputs, rather than relying solely on population-level representations. *We argue that these intracellular dynamics, which perform specific computational operations rather than a generalized logic, are the computational primitives of neural systems.*

**Neurons Process by Intracellular Dynamics** The primary difference between artificial neural networks (ANNs) and biological systems is the presence of spatiotemporal dynamics, which range order of magnitude in both temporal and spatial scales, compared to ANNs which primarily utilize instantaneous activation functions. These dynamics allow intracellular processing by integrating inputs at different time-constants and allowing mixtures of these signals. Even the simplest dynamics, such as a continuously evolving internal state, allows computations such as integrals and correct assignment of error through time [2]. Utilizing even such simple dynamics in a recurrently connected system can lead to complex but predictable population-level dynamics that can act as function generators for a variety of tasks [3]. When further expanded from a first-order to second-order differential equation, internal dynamics are able to implement history-dependent effects such as rebound spiking, which has been shown as essential for maintenance of information over extended periods of time [4], and implicated as a mechanism for working memory [5]. Further allowing for spatial compartmentalization of membrane dynamics allows otherwise complex operations such as division and multiplication to occur within the arborization of a single neuron [6]. Collections of dendrites are then able to implement complex operations such as context-dependent processing [7], sensory fusion in neocortex [8], and continual learning on multiple tasks in hippocampus [9]. Local processing of inputs may allow an implicit reconstruction and operation on presynaptic membrane potential, circumventing information bottlenecks otherwise imposed by binary spikes [10].

**Intracellular Dynamics Modify Intercellular Processing** While simple LIF-like units respond with monotonically increased firing rate in response to input spikes, additional intracellular dynamics enable the modulation of intercellular spike-based communication. For example, units with a controllable inter-spike-interval pattern, which requires at least fourth-order differential equations, carry higher information density than firing rates alone [11]. This increased information density can allow for multiplexing of multiple streams of information [12], which then supports supervised [13], unsupervised [14], and one-shot [15] learning. Local populations of neurons can also become synchronized due to small-magnitude oscillations in local-field potentials coupling with intracellular responses, which then have causal effects on intracellular communication, by suppressing or increasing intracellular responses to incoming spikes. Oscillations of various frequencies and underlying physical origin have been implicated in multiple computational roles, including gating of information, spatial attention in visual systems [16], role-binding [17], and working memory in prefrontal cortex [18], and prevents catastrophic interference in hippocampus [19].

**Training** In recent years, LIF-like units have been successfully trained using backpropagation methods, similar to those in standard deep-learning approaches [20]. However, such approaches fail to optimize systems which require multiple temporal scales of interactions, as will be the case in systems which utilize intracellular and intercellular processing. Therefore, to train such systems, we will need to integrate biological learning rules which operate over multiple temporal scales, along with recent advances in dynamic systems approaches [21], [22] which fit dynamics of systems rather than function approximations, akin to physics-informed neural networks.

**Implications for Analog Design and Scaling** While the number of candidate dynamics outlined above prohibits bio-mimicry approaches, the intracellular behaviors outlined above are all second-order equations, suggesting that neuromorphic systems implement neural units as a mixture of universal-oscillators. Combined with negative feedback mechanisms, this class can create stable attractors and other behaviors that mitigate the intrinsic noise and variability of analog systems. Simultaneously, utilizing locally complex and dense computation within individual neurons can minimize the number of intercellular connections that typically accompanies population encoding. This effect can be compounded by imposing general topological and functional motifs that emphasize local communication [23], resulting in a hierarchy of extremely dense intracellular analog interactions, locally analog but mean-field oscillatory interactions, and long-range interactions via spiking activity only. Such minimization of long-range communication is critical for allowing high throughput of flexible digital routing [24], and enables relatively seamless interaction with HPC-based systems for scaling.

**Necessary Datasets** To continue to develop a comprehensive view of the interaction of spike-based and intracellular dynamics, we must continue to collect datasets which record both behaviors. Recently, systems neuroscience has emphasized methods which enable recording of large populations of neural spikes, but which do not record intracellular dynamics provided by earlier methods. However, novel and emerging techniques have allowed for recording of large populations, while also providing information on the intracellular responses of dendritic arborizations [25], including neurotransmitter specific responses [26], and relating these responses to systems-level activity [27].

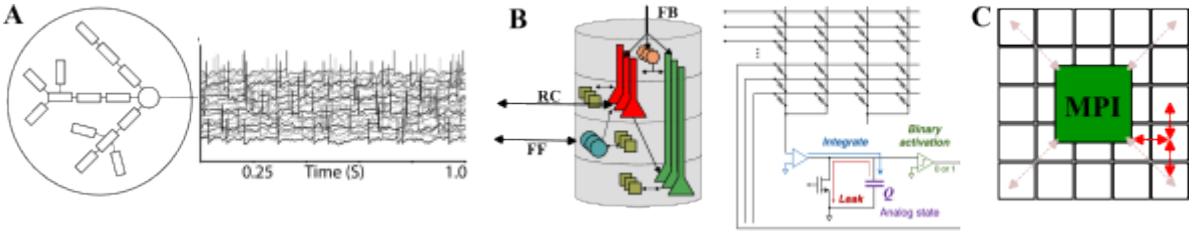



Fig. 1. A multi-scale approach to processing. **A** An individual neuron is made of multiple analog compartments, each of which has a unique combination of dynamics, chosen dependent on the use-case. **B** Local cortical structures, which are highly interconnected and transmit analog and spike values locally. Such a population could be manufactured into a single chip. **C** For larger scale systems, individual analog chips (from B) communicate to other distance chips only by spikes. These spikes can be routed by microprocessors (eg: MPI), allowing integration into existing HPC architectures.

SNL is managed and operated by NTESS under DOE NNSA contract DE-NA0003525

## REFERENCES


[1] C. Eliasmith and C. H. Anderson, *Neural Engineering: Computation Representation and Dynamics in Neurobiological Systems*. Cambridge, MA: MIT Press, 2004.

[2] G. W. Chapman, C. Teeter, S. Agarwal, T. P. Xiao, P. Hays, and S. S. Musuvathy, "Biological dynamics enabling training of binary recurrent networks," in *2024 Neuro Inspired Computational Elements Conference (NICE)*, 2024, pp. 1–7.

[3] D. Sussillo and L. F. Abbott, "Generating Coherent Patterns of Activity from Chaotic Neural Networks," *Neuron*, vol. 63, no. 4, pp. 544–557, Aug. 2009.

[4] E. K. W. Brennan and O. Ahmed, "Hyperexcitable Neurons Enable Precise and Persistent Information Encoding in the Superficial Retrosplenial Cortex," p. 24, 2020.

[5] M. Pals, T. C. Stewart, E. G. Akyürek, and J. P. Borst, "A functional spiking-neuron model of activity-silent working memory in humans based on calcium-mediated short-term synaptic plasticity," *PLoS computational biology*, vol. 16, no. 6, p. e1007936, 2020.

[6] F. S. Chance and S. G. Cardwell, "Shunting inhibition as a neural-inspired mechanism for multiplication in neuromorphic architectures," in *Proceedings of the 2023 Annual Neuro-Inspired Computational Elements Conference*, 2023, pp. 41–46.

[7] N. Takahashi, C. Ebner, J. Sigl-Glöckner, S. Moberg, S. Nierwetberg, and M. E. Larkum, "Active dendritic currents gate descending cortical outputs in perception," *Nature Neuroscience*, Aug. 2020.

[8] M. Ghosh, G. Béna, V. Bormuth, and D. F. M. Goodman, "Nonlinear fusion is optimal for a wide class of multisensory tasks," *PLOS Computational Biology*, vol. 20, no. 7, pp. 1–20, Jul. 2024.

[9] W. A. Wybo, M. C. Tsai, V. A. K. Tran, B. Illing, J. Jordan, A. Morrison, and W. Senn, "NMDA-driven dendritic modulation enables multitask representation learning in hierarchical sensory processing pathways," *Proceedings of the National Academy of Sciences*, vol. 120, no. 32, p. e2300558120, 2023.

[10] B. B. Ujfalussy, J. K. Makara, T. Branco, and M. Lengyel, "Dendritic nonlinearities are tuned for efficient spike-based computations in cortical circuits," *Elife*, vol. 4, p. e10056, 2015.

[11] Z. Friedenberger, E. Harkin, K. Tóth, and R. Naud, "Silences, spikes and bursts: Three-part knot of the neural code," *The Journal of Physiology*, vol. 601, no. 23, pp. 5165–5193, 2023.

[12] R. Naud and H. Sprekeler, "Sparse bursts optimize information transmission in a multiplexed neural code," *Proceedings of the National Academy of Sciences*, vol. 115, no. 27, pp. E6329–E6338, Jul. 2018.

[13] A. Payeur, J. Guerguiev, F. Zenke, B. A. Richards, and R. Naud, "Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits," *Nature Neuroscience*, vol. 24, no. 7, pp. 1010–1019, Jul. 2021.

[14] G. W. Chapman and M. E. Hasselmo, "Predictive learning by a burst-dependent learning rule," *Neurobiology of Learning and Memory*, p. 107826, 2023.

[15] K. C. Bittner, A. D. Milstein, C. Grienberger, S. Romani, and J. C. Magee, "Behavioral time scale synaptic plasticity underlies CA1 place fields," *Science (New York, N.Y.)*, vol. 357, no. 6355, pp. 1033–1036, Sep. 2017.

[16] R. F. Helfrich, I. C. Fiebelkorn, S. M. Szczepanski, J. J. Lin, J. Parvizi, R. T. Knight, and S. Kastner, "Neural Mechanisms of Sustained Attention Are Rhythmic," *Neuron*, vol. 99, no. 4, pp. 854–865.e5, 2018.

[17] J. E. Hummel, K. J. Holyoak, C. Green, L. A. A. Doumas, D. Devnich, and D. J. Kalar, "A Solution to the Binding Problem for Compositional Connectionism," p. 4, 2004.

[18] M. Lundqvist, P. Herman, M. R. Warden, S. L. Brincat, and E. K. Miller, "Gamma and beta bursts during working memory readout suggest roles in its volitional control," *Nature Communications*, vol. 9, no. 1, 2018.

[19] N. a Ketz, S. G. Morkonda, and R. C. O'Reilly, "Theta Coordinated Error-Driven Learning in the Hippocampus," *PLoS Computational Biology*, vol. 9, no. 6, p. e1003067, 2013.

[20] J. K. Eshraghian, M. Ward, E. O. Neftci, X. Wang, G. Lenz, G. Dwivedi, M. Bennamoun, D. S. Jeong, and W. D. Lu, "Training spiking neural networks using lessons from deep learning," *Proceedings of the IEEE*, 2023.

[21] A. Gu, K. Goel, and C. Ré, "Efficiently Modeling Long Sequences with Structured State Spaces," Aug. 2022.

[22] M. Schöne, N. M. Sushma, J. Zhuge, C. Mayr, A. Subramoney, and D. Kappel, "Scalable event-by-event processing of neuromorphic sensory signals with deep state-space models," *arXiv preprint arXiv:2404.18508*, 2024.

[23] A. M. Bastos, W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries, and K. J. Friston, "Canonical Microcircuits for Predictive Coding," *Neuron*, vol. 76, no. 4, pp. 695–711, 2012.

[24] A. Subramoney, K. K. Nazeer, M. Schöne, C. Mayr, and D. Kappel, "Efficient recurrent architectures through activity sparsity and sparse back-propagation through time," Mar. 2023.

[25] M. E. Sheffield and D. A. Dombeck, "Calcium transient prevalence across the dendritic arbour predicts place field properties," *Nature*, vol. 517, no. 7533, pp. 200–204, 2015.

[26] A. Aggarwal, R. Liu, Y. Chen, A. J. Ralowicz, S. J. Bergerson, F. Tomaska, B. Mohar, T. L. Hanson, J. P. Hasseman, D. Reep *et al.*, "Glutamate indicators with improved activation kinetics and localization for imaging synaptic transmission," *Nature methods*, vol. 20, no. 6, pp. 925–934, 2023.

[27] N. Cheng, Q. Dong, Z. Zhang, L. Wang, X. Chen, and C. Wang, "Egocentric processing of items in spines, dendrites, and somas in the retrosplenial cortex," *Neuron*, 2023.

# Computational Elements Underpinning the Emergence of Complex Behavior

Holger Dannenberg
George Mason University

**Position paper on neuroscience algorithms**

Despite significant advancements in neuroscience over the past decades, the field still lacks a comprehensive theory explaining how intelligent behavior emerges from neural activity. Although extensive datasets detailing neural correlates of behavior are available, the underlying causal structures remain largely unknown. To bridge this gap, *we need a theoretical framework that elucidates the minimal set of rules from which computational functions emerge*. These rules, once identified, can be translated into computational algorithms and implemented into analog circuits performing functions analogous to those executed by the brain.

An exemplary demonstration of this approach is our recent collaborative work [1] on an axiomatic framework, where we provide mathematical proof that spatially periodic firing patterns of grid cells in the entorhinal cortex arise from a simple sequence code of trajectories. This code is straightforward, interpretable, and intelligible. When implemented into a network simulation, such a code is expected to generate complex network activity patterns that perform the computational functions of grid cells. This indicates that identifying the simplest rules from which higher-order neural activity patterns emerge is crucial for reverse-engineering the simplest computational algorithms that perform complex brain functions.

This code leverages neural sequences, sometimes referred to as neural syntax, as a fundamental mode of brain function found in nearly all cortical regions, including the neocortex, entorhinal cortex, and hippocampus [2–5]. Implementing appropriate reader circuits, sequential activity of neurons can be used to store and consolidate memories and plan future actions. Despite their significance, current models rarely utilize sequential activity of neurons [6, 7]. Even when sequential activity emerges in simulations, it is often not read out by downstream readers to guide action, missing an opportunity to leverage the computational power and information content of neural sequences.

Sequential activity of neurons is embedded in brain rhythms, such as the hippocampal theta rhythm (6 10 Hz in rodents) [8]. Theta-rhythmic temporal organization, play a pivotal role in organizing brain activity [9]. These rhythms can separate encoding and retrieval intervals, modulate synaptic plasticity, bind different brain regions through synchronization, and synchronize activity within a brain region [10]. Despite their importance, the computational advantages of rhythmic activity are rarely leveraged in artificial systems. Utilizing these rhythms as computational mechanisms will enhance the power and efficiency of artificial networks by employing self-organizing principles in the temporal domain.

Brain rhythms are also a defining feature of brain states. Neuromodulators, such as acetylcholine, serotonin, and dopamine, provide a powerful means to modulate brain states [11]. They influence network dynamics by modulating meta-learning rules, affecting synaptic wakefulness, arousal, alertness, motivation, and decision-making. However, neuromodulation is not widely incorporated into simulations of neural networks [6]. Integrating neuromodulators into artificial networks would enable these systems to adapt to novel challenges and requirements, such as switching between learning new features and retrieving previously learned features, thus preventing catastrophic interference—a common issue in artificial deep learning algorithms and a significant issue in changing environments or task settings. A theoretical framework is needed to explain the principles governing how novel information is stored in the brain while not erasing previously learned concepts. One such theory, introduced by Michael Hasselmo [12], suggests that the temporal organization of brain states by neuromodulation allows for the separation of learning and retrieval of information at long timescales, preventing the overwriting of previously stored information. On finer timescales, theta rhythmic activity separates encoding and retrieval intervals, enabling brain circuits to retrieve and predict information on an internal model of the world in one cycle and update this model in the next cycle. Implementing pacemaker circuits and/or neuromodulation would enable such computational functions.

Brain activity can be studied across multiple spatial scales, yet artificial neural networks do not replicate the various connectivity patterns across spatial dimensions, such as dendritic versus somatic sites and cross-regional connections. Understanding the connectome and how architectural designs constrain activity while allowing flexible switching between different activity states will be critical in designing a circuit primitive.

To advance the field, it is essential to establish a common benchmark for testing and comparing the efficiency and performance of neural networks and artificial agents with that of animals and human subjects. This would facilitate meaningful comparisons and drive progress toward developing computationally efficient neuro-inspired algorithms. Meaningful progress can be made toward the standardization of common behavioral tasks with both neural and behavioral metrics comparable across agents, animals, and humans [13].

In conclusion, identifying the simplest rules from which complex brain dynamics emerge is crucial for the development of computational algorithms that emulate brain functions. *By leveraging the power of brain rhythms, neuromodulation, neural sequences, and understanding the spatial and temporal scales of brain activity, we can make significant strides toward creating more powerful and efficient artificial systems.*

**References**


1. R.G. R, Ascoli GA, Sutton NM, Dannenberg H (2024) Spatial periodicity in grid cell firing is explained by a neural sequence code of 2-D trajectories. *eLife*, 13https://doi.org/10.7554/eLife.96627.1

2. Dragoi G (2024) The generative grammar of the brain: a critique of internally generated representations. *Nature Reviews. Neuroscience*, 25(1):60–75. https://doi.org/10.1038/s41583-023-00763-0

3. Buzsáki G, Tingley D (2018) Space and Time: The Hippocampus as a Sequence Generator. *Trends in Cognitive Sciences*, 22(10):853–869. https://doi.org/10.1016/j.tics.2018.07.006

4. Rueckemann JW, Sosa M, Giocomo LM, Buffalo EA (2021) The grid code for ordered experience. *Nature Reviews Neuroscience*, 22(10):637–649. https://doi.org/10.1038/s41583-021-00499-9

5. Zhou S, Masmanidis SC, Buonomano DV (2020) Neural Sequences as an Optimal Dynamical Regime for the Readout of Time. *Neuron*, 0(0)https://doi.org/10.1016/j.neuron.2020.08.020

6. Hasselmo ME, Alexander AS, Hoyland A, Robinson JC, Bezaire MJ, Chapman GW, Saudargiene A, Carstensen LC, Dannenberg H (2020) The unexplored territory of neural models: Potential guides for exploring the function of metabotropic neuromodulation. *Neuroscience*, https://doi.org/10.1016/j.neuroscience.2020.03.048

7. Dannenberg H, Alexander AS, Robinson JC, Hasselmo ME (2019) The Role of Hierarchical Dynamical Functions in Coding for Episodic Memory and Cognition. *Journal of Cognitive Neuroscience*, 31(9):1271–1289. https://doi.org/10.1162/jocn_a_01439

8. Buzsáki G (2002) Theta oscillations in the hippocampus. *Neuron*, 33(3):325–340.

9. Buzsáki G, Moser EI (2013) Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nature neuroscience*, 16(2):130–138. https://doi.org/10.1038/nn.3304

10. Hasselmo ME, Stern CE (2014) Theta rhythm and the encoding and retrieval of space and time. *NeuroImage*, 85:656–666. https://doi.org/10.1016/j.neuroimage.2013.06.022

11. Grossman CD, Cohen JY (2022) Neuromodulation and Neurophysiology on the Timescale of Learning and Decision-Making. *Annual Review of Neuroscience*, 45:317–337. https://doi.org/10.1146/annurev-neuro-092021-125059

12. Hasselmo ME (2006) The role of acetylcholine in learning and memory. *Current Opinion in Neurobiology*, 16(6):710–715. https://doi.org/10.1016/j.conb.2006.09.002

13. Cothi W de, Nyberg N, Griesbauer E-M, Ghanamé C, Zisch F, Lefort JM, Fletcher L, Newton C, Renaudineau S, Bendor D, Grieves R, Duvelle É, Barry C, Spiers HJ (2022) Predictive maps in rats and humans for spatial navigation. *Current biology: CB*, :S0960-9822(22)01095–8. https://doi.org/10.1016/j.cub.2022.06.090

# Optimize Design Cost and Enhance Performance of a Neuromorphic System with Hardware-Software Co-Design

Hritom Das
Electrical and Computer Engineering,
Oklahoma State University, StillWater, OK
hritom.das@okstate.edu

## ABSTRACT

Neuromorphic computing is bio-inspired and energy-efficient. Various materials like $HfO_2$ and TaOx are used to construct neuromorphic circuitry like synapses, neurons, dot product engines, spike time-dependent-plasticity, reservoir computing, etc. The main challenge is to design and evaluate the system for reliability, energy efficiency, and compact design area. Most emerging devices and systems contain higher inherent variability, which is a big challenge to overcome without proper hardware design. At the time, to optimize the design cost and time, a hardware-software co-design can be beneficial. At the same time, a targeted performance is also achievable with the software framework.

## KEYWORDS

Memristor, synapse, inherent variation, process variation, programming levels, system-level optimization, fine-tune.

## 1 INTRODUCTION

Neuromorphic research and system implementations are expanding rapidly with significant success. Neuromorphic systems are developed with conventional CMOS technology with area and energy overhead.[1]. Moreover, it provides a reliable functionality with higher frequency. On the other hand, emerging materials are used to construct neuromorphic hardware with optimized design area and energy consumption [2–6]. At the same time, emerging technology shows a reliability issue with higher inherent process variation[2, 3]. However, analog neuromorphic circuits and systems play an important role in reducing peripherals like encoders and decoders. Spikes are treated as spikes in the neuromorphic cores and communication networks. In addition, analog neuromorphic systems are a good fit with various sensors. Analog data from a sensor can be directly provided to a neural core. This approach is useful to reduce the data prepossessing with peripherals. Various applications are covered with neuromorphic design such as dot product engine (DPE)[2], spike-time-dependant-plasticity (STDP)[5, 7–10], homeostatic plasticity[11], reservoir computing[12, 13], and so on. These implementations with emerging materials show optimized energy [5] and higher design density[4]. The main building block of this kind of architecture is memristive devices, which are coined with emerging materials like $HfO_2$ and TaOx. There are various challenges to using these emerging materials for large-scale design. a proper and guided design method can optimize the design cost and time with targeted performance.

## 2 CHALLENGE

Neuromorphic systems with emerging materials are highly sensitive to their functionality, endurance, retention time, weight precision, operating voltage, and system-level integration & scaling[2, 3, 14]. The initial challenges of an emerging device development are to maintain a higher endurance, longer retention time, the number of programming levels, and a lower operating voltage. Whenever the emerging devices are integrated with CMOS technology, the CMOS devices also introduce variation on top of the emerging device variation. CMOS compatibility is required to make the device suitable for system-level integration. Otherwise, expensive equipment will be required to operate the device by itself. In addition, the CMOS variation also reduces the programming levels. Most of the emerging materials require higher forming or switching voltage. Due to higher operating voltage, sub-micron CMOS technologies are not a good fit for emerging devices. Another change in designing an analog neuromorphic system is to scale up the number of synapses and neurons. The system will introduce reliability issues with the scaling. It would be great if we could evaluate the system-level design with the characteristics of the emerging devices and circuits.

## 3 OPPORTUNITY

There are various challenges to making an analog neuromorphic system functional and reliable. At first, the issue with endurance, retention time, and number of programming levels can be solved with device or physics-level research. After that, reliable and fully functional memristive circuitry like synapse, STDP, DPE, neural core, and so on can be designed with proper scaling of CMOS devices and supply voltage. Finally, the system-level integration and performance evaluation can be done with an optimized network. Evolutionary optimization for neuromorphic systems (EONS) and liquid state machines (LSMs)[15] can be utilized to optimize the spiking neural networks (SNN) for neuromorphic applications. At the same time, a hardware framework called RAVENS is useful to utilize for system-level performance evaluation[1]. The characteristics of the neuromorphic hardware are mapped with look-up tables and equations to observe the performance of a large-scale system. The parameters of hardware can be fine-tuned to get a better accuracy of various applications. Different parameters are useful for different applications[13]. A dynamically adaptive system can be designed to get a better performance at run time.

## 4 CONCLUSION

Analog neuromorphic system design and evaluation is a time-consuming and expensive process. To mitigate the design time and cost, a hardware-software co-design can be beneficial for rapid development. The parameters of the hardware design such as latency, power, leak, refractory, and number of weight levels can be tuned according to the needs of the software or application's performance. A trade-off between performance and design cost can be predetermined by the hardware-software co-design, which is useful for ASIC design.

# REFERENCES


[1] Adam Z. Foshie, Charles Rizzo, Hritom Das, Chaohui Zheng, James S. Plank, and Garrett S. Rose. Benchmark comparisons of spike-based reconfigurable neuroprocessor architectures for control applications. In *Proceedings of the Great Lakes Symposium on VLSI 2022*, GLSVLSI '22, page 383–386, New York, NY, USA, 2022. Association for Computing Machinery.

[2] Hritom Das, Rocco D Febbo, Sree Nirmillo Biswash Tushar, Nishith N Chakraborty, Maximilian Liehr, Nathaniel C Cady, and Garrett S Rose. An efficient and accurate memristive memory for array-based spiking neural networks. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2023.

[3] Hritom Das, Rocco D Febbo, Charles P Rizzo, Nishith N Chakraborty, James S Plank, and Garrett S Rose. Optimizations for a current-controlled memristor-based neuromorphic synapse design. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2023.

[4] Hritom Das, Catherine Schuman, Nishith N Chakraborty, and Garrett S Rose. Enhanced read resolution in reconfigurable memristive synapses for spiking neural networks. *Scientific Reports*, 14(1):8897, 2024.

[5] Nishith N Chakraborty, SNB Tushar, Hritom Das, and Garrett S Rose. Energy efficient and high-performance synaptic operating point evaluation for snn applications. In *2023 IEEE 66th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 918–922. IEEE, 2023.

[6] SNB Tushar, Hritom Das, and Garrett S Rose. Hfo 2-based synaptic spiking neural network evaluation to optimize design and testing cost. In *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2024.

[7] Nishith N Chakraborty, Hritom Das, and Garrett S Rose. A mixed-signal short-term plasticity implementation for a current-controlled memristive synapse. In *Proceedings of the Great Lakes Symposium on VLSI 2023*, pages 179–182, 2023.

[8] Nishith N Chakraborty, Hritom Das, and Garrett S Rose. Spike-driven synaptic plasticity for a memristive neuromorphic core. In *2023 IEEE 66th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 644–648. IEEE, 2023.

[9] Nishith N Chakraborty, Shelah O Ameli, Hritom Das, Catherine D Schuman, and Garrett S Rose. Hardware software co-design for leveraging stdp in a memristive neuroprocessor. *Neuromorphic Computing and Engineering*, 4(2):024010, 2024.

[10] Ryan Weiss, Hritom Das, Nishith N Chakraborty, and Garrett S Rose. Stdp based online learning for a current-controlled memristive synapse. In *2022 IEEE 65th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 1–4. IEEE, 2022.

[11] Nishith N Chakraborty, Hritom Das, and Garrett S Rose. Homeostatic plasticity in a leaky integrate and fire neuron using tunable leak. In *2023 IEEE 66th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 738–742. IEEE, 2023.

[12] Manu Rathore, Rocco Febbo, Adam Foshie, Sree Nirmillo Biswash Tushar, Hritom Das, and Garrett S Rose. Reliability analysis of memristive reservoir computing architecture. In *Proceedings of the Great Lakes Symposium on VLSI 2023*, pages 131–136, 2023.

[13] Catherine D Schuman, Hritom Das, James S Plank, Ahmedullah Aziz, and Garrett S Rose. Evaluating neuron models through application-hardware co-design. In *2023 57th Asilomar Conference on Signals, Systems, and Computers*, pages 537–542. IEEE, 2023.

[14] Hritom Das, Manu Rathore, Rocco Febbo, Maximilian Liehr, Nathaniel C Cady, and Garrett S Rose. Rfam: Reset-failure-aware-model for hfo2-based memristor to enhance the reliability of neuromorphic design. In *Proceedings of the Great Lakes Symposium on VLSI 2023*, pages 281–286, 2023.

[15] Catherine D Schuman, J Parker Mitchell, Robert M Patton, Thomas E Potok, and James S Plank. Evolutionary optimization for neuromorphic systems. In *Proceedings of the 2020 Annual Neuro-Inspired Computational Elements Workshop*, pages 1–9, 2020.

**NeuroAI: Neuromorphic Computing for Edge AI**

Position Paper: Neuromorphic Computing for Science Workshop

Technical area: Modeling and simulation approaches

Prasanna Date (Lead Author), Thomas E. Potok (Advising Author)
Oak Ridge National Laboratory

**Abstract**

We believe there is a need to develop an end-to-end approach for advancing AI at the edge for scientific applications through neuromorphic computing. Our approach would train neuromorphic AI models on quantum computers, simulate them in real-world scenarios on supercomputers, and deploy them in scientific applications on neuromorphic computers. We envision several key objectives: (1) Train neuromorphic AI models efficiently on universal quantum computers; (2) Develop an exascale simulator on the Frontier supercomputer for simulating neuromorphic AI models from analog circuity on real-world scenarios; (3) Develop a simulator for simulating digital, analog, and mixed-signal neuromorphic circuits and architectures; (4) Design and implement a digital neuromorphic hardware platform optimized for scientific edge AI applications; (5) Design a mixed-signal neuromorphic hardware platform; (6) Build a test bench for testing neuromorphic hardware platforms; and (7) Demonstrate the efficacy of our neuromorphic edge AI approach in two scientific applications: self-driving robotic cars and autonomous drones.

I. **Train SNNs on quantum computers**: Quantum computers are good at solving such problems and are well poised to accelerate the training of SNNs. In our previous work, we have shown that it is possible to get an order of magnitude faster performance using a quantum computer to train machine learning models over a classical computer. However, SNNs have never been trained on quantum computers.
II. **Simulate SNNs in the SuperNeuroMAT simulator at exascale**: Simulators for simulating neuromorphic algorithms have not been developed for HPC systems. Previously, we developed the SuperNeuroMAT simulator [3], which simulates neuromorphic algorithms and runs on CPUs on desktops and laptops.
III. **Simulate neuromorphic circuits and architectures in the SuperNeuroABM simulator**: Use SuperNeuroABM, a simulator that uses agent-based modeling, to simulate neuromorphic architectures [3]. SuperNeuroABM runs in a multi-node and multi-GPU fashion on NVIDIA GPUs. We plan to develop SuperNeuroABM on Frontier using a first principles approach.
IV. **Develop an FPGA-based digital neuromorphic hardware platform**: We will program a neuromorphic processor on a widely available, low-cost FPGA. FPGA-based implementations are faster, cheaper, and more reprogrammable than their analog counterparts. This will be helpful for users to implement and test SNN architectures quickly.
V. **Design a mixed-signal neuromorphic hardware platform**: We will design neuromorphic chips using mixed-signal substrates in this project. Use mixed-signal implementation to design and rigorous simulation across process corners. In mixed-signal neuromorphic chips, the synapse circuitry occupies the most silicon real- estate. We will use non-volatile memory devices to design synapse circuits. This will significantly improve the integrated circuit density.

This position paper lays out concepts to advance AI at the edge for scientific applications using neuromorphic computing by developing an end-to-end approach. Our plan involves training neuromorphic AI models on quantum computers, simulating them on supercomputers, and deploying them in scientific applications. Key objectives include: efficient training on quantum computers, developing exascale simulators on Frontier supercomputer, creating simulators for various neuromorphic circuits, designing digital and mixed-signal neuromorphic hardware platforms, building a test bench, and demonstrating the approach in self-driving robotic cars and autonomous drones.

## Supplemental Information

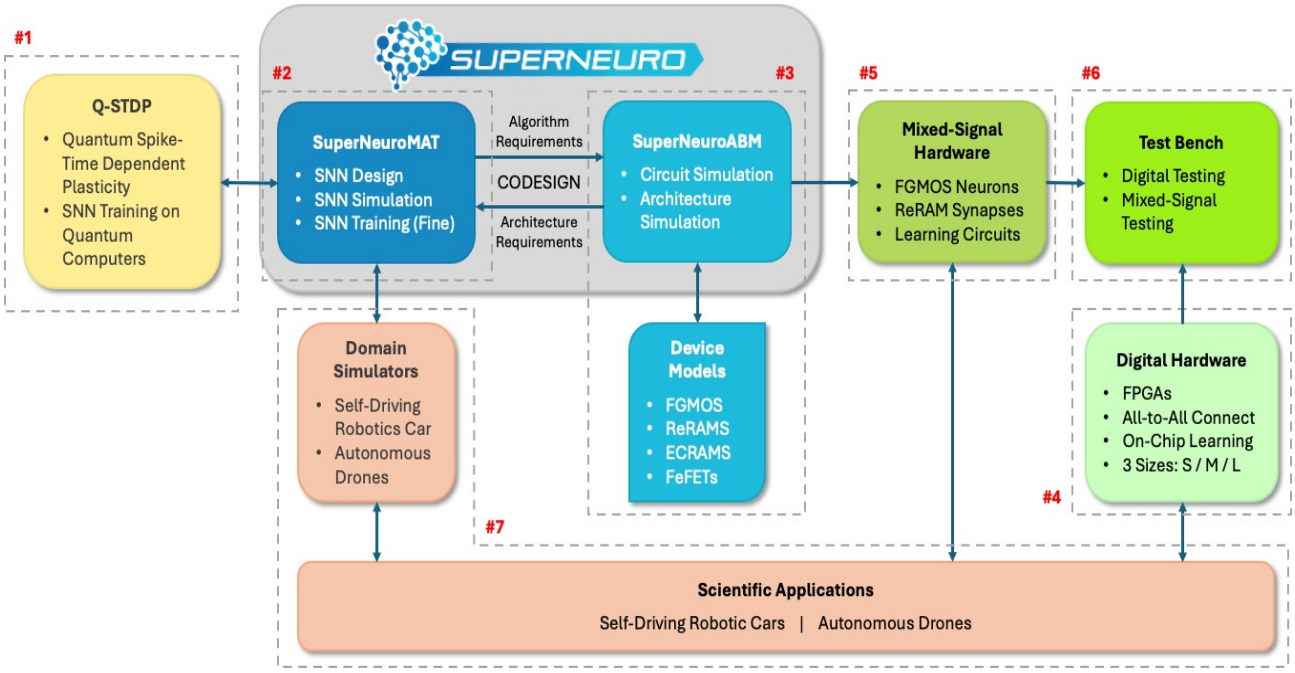



*Figure 1 A notional concept of the architecture of the proposed system based on prior research in these areas. Overview diagram of the project. Each objective is shown in a different color.*

## References


1. Date, Prasanna, Davis Arthur, and Lauren Pusey-Nazzaro. "QUBO formulations for training machine learning models." Scientific reports 11.1 (2021): 10029
2. Date, Prasanna, et al. "Neuromorphic computing is Turing-complete." Proceedings of the International Conference on Neuromorphic Systems 2022. 2022.
3. Date, Prasanna, et al. "Superneuro: A fast and scalable simulator for neuromorphic computing." Proceedings of the 2023 International Conference on Neuromorphic Systems. 2023.
4. Schuman, Catherine D., et al. "Opportunities for neuromorphic computing algorithms and applications." Nature Computational Science 2.1 (2022): 10-19.
5. Gautam, Ashish, and Takashi Kohno. "Biomimetic analog silicon synaptic circuit with tunable reversal potential." J. Robotics Netw. Artif. Life 7.1 (2020): 22-26.

Brain-inspired Neuromorphic Computing

Rogene Eichler West
Pacific Northwest National Laboratory

**Brain-inspired Neuromorphic Computing**

For neuromorphic computing to incorporate more of the salient features of information processing in the brain, several additional capabilities should be considered beyond the concept of integrate-and-fire neurons: excitatory-inhibitory circuits, functional connectivity and small world modularity, rate and burst encoding, neuromodulation, and direct electrical coupling. The McCulloch and Pitts neural model (1943), developed concurrently with Hodgkin and Huxley's Nobel Prize winning characterization of the "integrate and fire" action potential in the squid giant axon (1952) yielded the individual processing units of artificial neural networks. The recognition that learning occurs in a network because of adaptive weights between those functional units was first proposed by Hebb (1949) colloquially as "neurons that fire together, wire together", and then was experimentally verified later in studies of long-term potentiation by Bliss and Lono (1973), gave rise to the deep learning algorithms that are ubiquitous today. Our understanding of brain function has advanced significantly in the past 75 years, but the transfer of core principles to algorithmic design has not.

Excitatory-Inhibitory circuits enable the brain to efficiently process and respond to important stimuli, while filtering out unnecessary details. This ability to focus is essential for effective perception, decision-making, and adaptive behavior in complex environments. It is a well-studied phenomena in the retina for edge detection, contrast enhancement, and noise reduction. In the visual cortex, these microcircuits refine which microcolumns participate in the orientation selectivity, which themselves become the building blocks of perception. In motor cortex, it allows for the precision, coordination, and synchronization of movement. At a systems level, such as the loop that exists between the thalamus and cortex, the balance between these two contributions regulates states of consciousness. The computational concept is that the center of an activation inhibits the lateral processing units, or the "surround").

It has been known since the ablation studies of Lashley in the 1930s that the brain is divided into specialized processing centers. More recently, functional studies of the brain have revealed that the organization within and between those centers are characterized by a high degree of clustering known as small world networks. This emergent property of the developmental process yields more efficient processing, striking a balance between local specialization and global integration, while minimizing the wiring cost of connectivity. But further, these networks are not simply hard-wired, but functionally connected in a dynamic, meta-stable manner such that rapid adaptive switching is possible between, for example, salience to outside stimuli, executive function such as working memory, decision-making, and problem-solving, and a default mode that is self-referential and reflects on past and future events. This stands in contrast with the "layers" motif, which continues to be the organization principle of deep learning algorithms.

Many classes of neurons exist, and they are often characterized by the one or more distinct firing patterns they can produce: tonic, bursting, phasic, adaptive, rhythmic, or irregular. Many are electrically, directly, and dynamically coupled to each other (and to other non-spiking brain cells through gap junctions), which can result in synchronization at a very rapid time scale. Neuromodulators, chemicals which act on times scales several orders of magnitude longer than the duration of a typically "spike", have the impact of shifting neurons between these patterns. The consequence of neural populations shifting their firing patterns is the resulting change in the excitatory-inhibitory balance in circuits, which in turn perturbates the metastability of the functional networks.

To summarize, a neuromorphic wish list from a neuroscience perspective would be to have devices with firing pattern-configurable "neurons", modeled with chemical and electrical synapses. At the local network scale, activations result in a lateral inhibition of neighboring processing units. At a global scale, architectures should be physically connected using small world networks (which also reduces the demand on fan in/out requirements). These building blocks, drawn from canonical concepts in brain information processing, will lead to entirely new concepts in engineering efficient and effective analog information processing tools.

## References


Beck, Sandra, and Mark Hallett. "Surround inhibition in the motor system." *Experimental brain research* 210 (2011): 165-172.

Bliss, Tim VP, and Terje Lømo. "Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path." *The Journal of physiology* 232, no. 2 (1973): 331-356.

Hebb, Donald O. "The organization of behavior." *New York* (1949).

Hodgkin, Alan L., and Andrew F. Huxley. "A quantitative description of membrane current and its application to conduction and excitation in nerve." *The Journal of physiology*117, no. 4 (1952): 500.

McCulloch, Warren S., and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity." *The bulletin of mathematical biophysics* 5 (1943): 115-133.

Mitra, Anish, and Marcus E. Raichle. "How networks communicate: propagation patterns in spontaneous brain activity." *Philosophical Transactions of the Royal Society B: Biological Sciences* 371, no. 1705 (2016): 20150546.

Muldoon, Sarah Feldt, Eric W. Bridgeford, and Danielle S. Bassett. "Small-world propensity and weighted brain networks." *Scientific reports* 6, no. 1 (2016): 22057.

Raichle, Marcus E. "The brain's default mode network." *Annual review of neuroscience* 38, no. 1 (2015): 433-447.

Segall, K., C. Purmessur, A. D'Addario, and D. Schult. "A superconducting synapse exhibiting spike-timing dependent plasticity." *Applied Physics Letters* 122, no. 24 (2023).

Series, Peggy, Jean Lorenceau, and Yves Frégnac. "The “silent” surround of V1 receptive fields: theory and experiments." *Journal of physiology-Paris* 97, no. 4-6 (2003): 453-474.

Tatti, Roberta, Melissa S. Haley, Olivia K. Swanson, Tenzin Tselha, and Arianna Maffei. "Neurophysiology and regulation of the balance between excitation and inhibition in neocortical circuits." *Biological psychiatry* 81, no. 10 (2017): 821-831.

Tschirhart, Paul, and Ken Segall. "Brainfreeze: Expanding the capabilities of neuromorphic systems using mixed-signal superconducting electronics." *Frontiers in Neuroscience* 15 (2021): 750748.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

# Optimizing Mixed-Signal Neuromorphic Circuits: Bridging Computational Gaps

*Ashish Gautam, Oak Ridge National Laboratory*

**Introduction:** A key application for neuromorphic chips is low-power intelligent edge AI systems, enabling sensors and devices to operate without internet connectivity. The critical design metrics are ultra-low power consumption and on-chip learnability. Digital implementations offer faster design cycles and higher CMOS integration, but mixed-signal chips with analog neuromorphic cores excel in low power consumption and, with emerging non-volatile memory, are expected to achieve high integration density in the near future. This paper explores future directions for designing mixed-signal neuromorphic chips.

**Implementation Challenges & Opportunities:**

**Analog Neuromorphic core:** One promising direction to minimize energy consumption in next generation chips is designing core circuits to operate at a lower power supply voltage of 200 mV —compared to the typical 1V—which also reduces the size of the membrane capacitor and the circuit footprint. However, the lower voltage limits the headroom for stacked devices to operate in saturation, making the design challenging. An example neuron circuit achieves lower voltage operation but generates only limited spiking behaviors [1]. Different spiking behaviors enable neurons to represent diverse information and perform complex computations [2, 3]. Similarly, while current-based log-domain integrator (LDI) synapse circuits [4] can operate with lower power supply, designing a conductance-based synapse necessary for implementing shunting inhibition remains challenging [5]. Therefore, new design techniques are needed. These techniques could exploit the non-idealities of MOS transistors at lower technology nodes and characteristics of emerging devices being explored for neuromorphic circuits [6].

**On-chip Learning**: Mixed-signal neuromorphic chips often suffer from fabrication mismatches deteriorating performance. On-chip learning mechanisms are essential for enabling spiking neural networks (SNNs) to adapt to these non-idealities without compromising accuracy. While local brain-inspired spike-timing dependent plasticity (STDP) type learning rules have been implemented [7], they are costly in silicon area and power due to the need for high-precision synapses and still lag artificial neural networks (ANNs) in performance. Developing novel algorithms with binary or low-precision synapses, such as those using memristors, is essential. One promising direction is combining local plasticity rules with reduced compartment neuron models. This is inspired by the observation that local learning mechanisms in a neuronal cell differ depending on the synapse location in the dendritic tree [8]. Studies show that a neuron with active dendritic compartments can match the computational power of a multi-layer ANN [9], revealing a gap in neuromorphic research and indicating a need for further exploration of these models. Compartmental neuron models also open avenues to explore novel connectivity motifs between neurons in different layers and when combined with compartment specific local learning mechanism can create new computing primitives [10, 11]. Compartmental models could further benefit from probabilistic synapse circuits, with studies showing binary probabilistic synapses excel in machine learning tasks [12]. Synapse circuits using devices such as magnetic tunnel junction (MTJs) —which have been demonstrated to generate desired probability distributions [13]—for controlled probabilistic spike transmission could improve performance while reducing energy consumption and footprint.

**Benchmarking:** The targeted datasets are output of neuromorphic event-based sensors that encode information spatiotemporally [14]. However, these sensors may not capture the brain's spike encoding mechanisms, which could be crucial for leveraging SNNs' computational power. Therefore, exploring datasets that reflect brain spike trains—either through sensory pathway modeling or Multi Electrode Array (MEA) recordings from brain organoid simulations [15]—is essential for developing next-generation neuromorphic sensors.

**Acknowledgement:**

The author thanks Takashi Kohno, Thomas Potok, and Robert Patton for their guidance, and Shruti Kulkarni, Prasanna Date, and Chathika Gunaratne for stimulating discussions and their input on the concepts described in this paper.

**References:**

1. Sourikopoulos, Ilias, Sara Hedayat, Christophe Loyez, François Danneville, Virginie Hoel, Eric Mercier, and Alain Cappy. "A 4-fJ/spike artificial neuron in 65 nm CMOS technology." Frontiers in neuroscience 11 (2017): 123.
2. Bellec, Guillaume, Darjan Salaj, Anand Subramoney, Robert Legenstein, and Wolfgang Maass. "Long short-term memory and learning-to-learn in networks of spiking neurons." Advances in neural information processing systems 31 (2018).
3. Kohno, Takashi, Munehisa Sekikawa, Jing Li, Takuya Nanami, and Kazuyuki Aihara. "Qualitative-modeling-based silicon neurons and their networks." Frontiers in NEUROSCIENCE 10 (2016): 273.
4. **Gautam, Ashish,** and Takashi Kohno. "A Conductance-Based Silicon Synapse Circuit." Biomimetics 7, no. 4 (2022): 246.
5. Bartolozzi, Chiara, and Giacomo Indiveri. "Synaptic dynamics in analog VLSI." *Neural computation* 19, no. 10 (2007): 2581-2603.
6. Chakraborty, Indranil, A. Jaiswal, A. K. Saha, S. K. Gupta, and K. Roy. "Pathways to efficient neuromorphic computing with non-volatile memory technologies." Applied Physics Reviews 7, no. 2 (2020).
7. **Gautam, Ashish,** and Takashi Kohno. "Adaptive STDP-based on-chip spike pattern detection." Frontiers in Neuroscience 17 (2023): 1203956.
8. Ebner, Christian, Claudia Clopath, Peter Jedlicka, and Hermann Cuntz. "Unifying long-term plasticity rules for excitatory synapses by modeling dendrites of cortical pyramidal neurons." Cell reports 29, no. 13 (2019): 4295-4307.
9. Beniaguev, David, Idan Segev, and Michael London. "Single cortical neurons as deep artificial neural networks." *Neuron* 109, no. 17 (2021): 2727-2739.
10. Hawkins, Jeff, and Subutai Ahmad. "Why neurons have thousands of synapses, a theory of sequence memory in neocortex." Frontiers in neural circuits 10 (2016): 174222.
11. Lillicrap, Timothy P., Adam Santoro, Luke Marris, Colin J. Akerman, and Geoffrey Hinton. "Backpropagation and the brain." Nature Reviews Neuroscience 21, no. 6 (2020): 335-346.
12. Yousefzadeh, Amirreza, Evangelos Stromatias, Miguel Soto, Teresa Serrano-Gotarredona, and Bernabé Linares-Barranco. "On practical issues for stochastic STDP hardware with 1-bit synaptic weights." Frontiers in neuroscience 12 (2018): 665.
13. Misra, Shashank, Leslie C. Bland, Suma G. Cardwell, Jean Anne C. Incorvia, Conrad D. James, Andrew D. Kent, Catherine D. Schuman, J. Darby Smith, and James B. Aimone. "Probabilistic neural computing with stochastic devices." Advanced Materials 35, no. 37 (2023): 2204569.
14. Amir, Arnon, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak et al. "A low power, fully event-based gesture recognition system." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7243-7252. 2017.
15. Friston, Karl. "The sentient organoid?." Frontiers in Science 1 (2023): 1147911.

**ML-Assisted Neuromorphic Architecture Modeling and Simulation for Co-Design and Exploration**

Position Paper Submitted to DOE ASCR Workshop on Neuromorphic Computing for Science

Tech. area: Scalable integration for neuromorphic computing modeling**,** Theme: Modeling & Simulation

Prof. Andreas Gerstlauer, ECE Dept., UT Austin; gerstl@ece.utexas.edu; (512) 232-8294

**Key Challenges:** Many different spiking neuromorphic platforms ranging from fully digital to hybrid digital-analog architectures have been proposed and implemented. In particular, analog circuitry and novel devices have become increasingly popular due to their promise of ultra-energy-efficient computation while providing novel exciting behaviors. Future hybrid system architectures are expected to incorporate such analog compute blocks with a digital backend to handle control logic and scalable networking among compute cores (Figure 1). These types of hybrid digital-analog architectures open many new co-design and design space exploration questions in terms of how to integrate emerging devices and analog sub-blocks at the system architecture and algorithm levels. Despite this, there is a lack of modeling and simulation tools to support rapid, early co-design and design space exploration. Such tools are crucial in allowing the designer to quickly explore tradeoffs between power, performance, flexibility, and functionality. Existing approaches are either purely functional, lacking any hardware or implementation aspects [1,2], or are based on low-level hardware simulation and models that are too detailed and slow for system-level design-space exploration while also not capturing full-system effects [3,4].

**Research Directions:** Effectively co-designing next-generation neuromorphic systems requires novel tools that enable fast and accurate modeling and simulation of large-scale hybrid analog/digital architectures. Specifically, there is a need for system-level neuromorphic architecture simulators at high abstraction levels that model energy and performance of full system designs when executing real neuromorphic benchmarks, while also being flexible and extensible to support vast configurations of different architectures. In recent work, we have developed SANA-FE as a novel high-level simulator of advanced neuromorphic architectures for fast exploration [5]. SANA-FE uses an abstract and high-level execution model that simulates a given neuromorphic architecture executing an SNN-based application to accurately estimate performance and energy at time-step granularity (Figure 2). Similar to other architecture simulators, SANA-FE is, however, currently limited to digital architectures while also requiring manual configuration and calibration against target implementation technologies.

Further research is needed on automating the modeling process and support for modeling of analog sub-blocks and sub-components integrated into CMOS+X based hybrid architectures. A promising approach is the use of machine learning (ML) for modeling and simulation. Such methods allow to automatically derive ML-based surrogate models that replace traditional simulators for fast performance and energy estimation through data-driven machine learning methods. In preliminary work, we have investigated such data-driven methods using various advanced regression models to automatically derive coarse-grain surrogate models for analog sub-blocks with energy and performance prediction [6]. Such models are fast and can easily integrate into the simulation backplane of existing digital simulators via a function call. ML-based models have the further advantage of potentially being differentiable, opening up opportunities for gradient-based co-design and exploration methods in finding optimal design points.

A further potential research direction is the use of more robust ML techniques such as neural ODEs, which have garnered attention for circuit simulation due to their accuracy in predicting time-varying time-series data, and Bayesian models, which provide prediction intervals on energy and performance estimation. This methodology, using machine learning, also raises questions about the granularity of splitting analog blocks into sub-blocks, and how to mitigate additive errors in sequences of interconnected models during simulation. This work can be further extended to capture energy and performance behaviors of circuits composed of novel devices as well. The development of fast and accurate modeling and simulation methodologies for analog circuitry is required to effectively co-design next-generation hybrid digital-analog neuromorphic architectures.

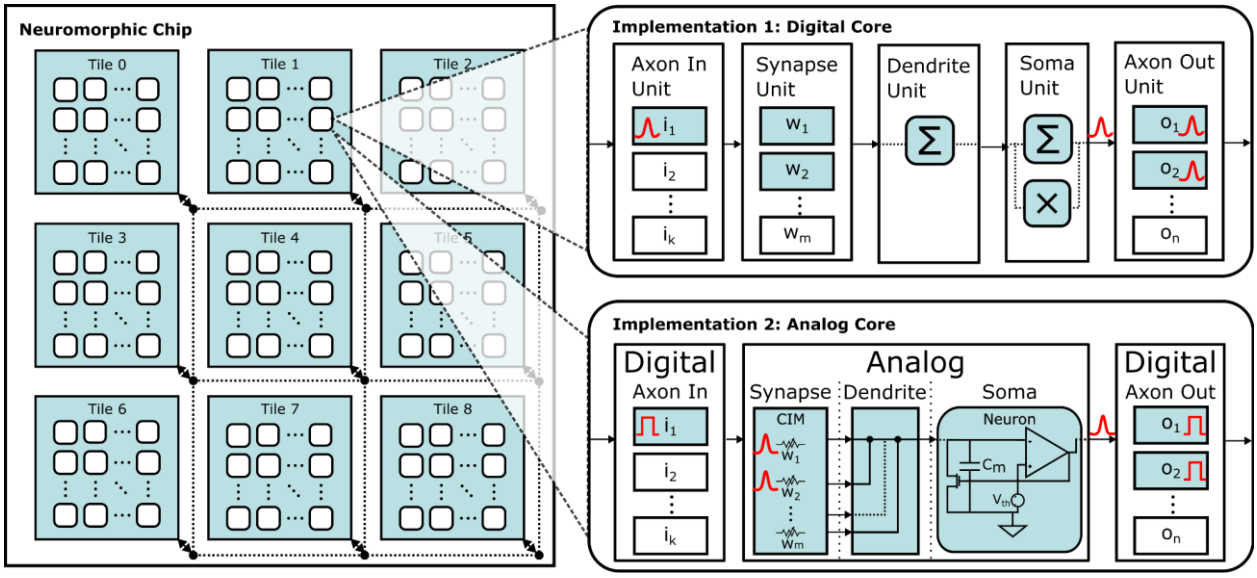



**Figure 1: Example of a scalable neuromorphic architecture.**

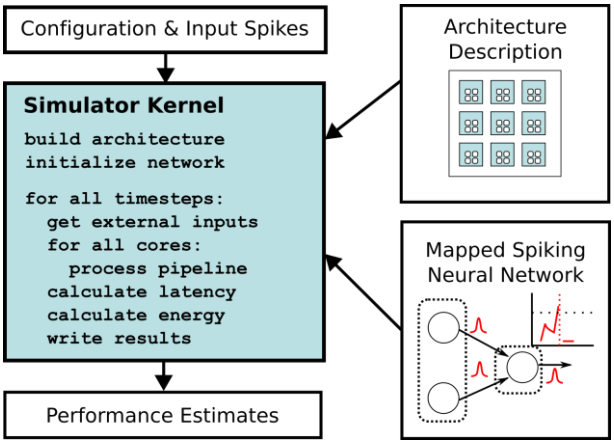



**Figure 2: SANA-FE simulator overview.**

## References


[1] M. Stimberg, R. Brette, and D. F. Goodman, "Brian 2, an intuitive and efficient neural simulator," *eLife*, vol. 8, p. e47314, Aug. 2019.

[2] T. Bekolay *et al.*, "Nengo: a Python tool for building large-scale functional brain models," Front. Neuroinform., vol. 7, p. 48, 2014.

[3] M. Plagge, C. D. Carothers, E. Gonsiorowski, and N. Mcglohon, "NeMo: A massively parallel discrete-event simulation model for neuromorphic architectures," *ACM Trans. Model. Comput. Simul. (TOMACS)*, vol. 28, no. 4, Sep. 2018.

[4] M. Plagge et al., "ATHENA: Enabling codesign for next-generation AI/ML architectures," in *IEEE Int. Conf. Rebooting Comput. (ICRC)*, 2022.

[5] J. Boyle, M. Plagge, S. Cardwell, F. Chance, and A. Gerstlauer, "Performance and Energy Simulation of Spiking Neuromorphic Architectures for Fast Exploration," in *ICONS*, Santa Fe, NM, Aug. 2023.

[6] J. Ho, J. Boyle, and A. Gerstlauer, "LASGNA: Large-scale Analog Surrogate Modeling for General Neuromorphic Architectures", submitted for publication.

# Energy-efficient neuromorphic hardware using analog and unary computing

Patricia Gonzalez-Guerrero
Lawrence Berkeley National Laboratory

**Energy-efficient neuromorphic hardware using analog and unary computing**

Patricia Gonzalez-Guerrero, Research Scientist, lg4er@lbl.gov

**Motivation:** Increasing the energy efficiency for AI training and inference is vital for scientific discovery. Biologically inspired, neuromorphic computing can be the key to solving the energy bottleneck limiting scientific AI workloads, given that the brain operates on a 20W budget [1]!

**Methods:** To address the AI energy bottleneck using neuromorphic computing we need a hardware architecture that complies with four design criteria (DC): (i) massive parallel operation, (ii) collocation of processing and memory, (iii) scalability, and (iv) event/data-driven computation [2]. Moreover, the brain processes and communicates information using a train of sparse electrical spikes. **A brain-inspired hardware architecture (Fig. 3) that combines (i) unary data representation alongside (ii) analog voltage/current instead of digital words, has the potential to surpass conventional digital computing in both energy efficiency and throughput because it can follow the aforementioned four design criteria. Moreover, in unary data representations, information is mapped to the time, frequency, or pulse width of a stream of "ones" and "zeros," similar to the electrical spikes the brain uses to compute and transmit information.** This brain-inspired hardware architecture, could achieve for the first time a neuromorphic hardware capable of using unary and analog data representations from end-to-end (Fig. 1-3) thus improving the energy efficiency of AI/ML algorithms such as Convolutional, Graph, and Spiking Neural Networks (CNN, GNN, SNN).

**Challenges:** Unary data representation is particularly well suited for mimicking the spiking nature of communication in the brain. The problem with unary computing is that individual components have been studied in isolation without system level or algorithm considerations. To give one example, unary computing effectively addresses the massive parallelism (DC-i, DC-iii) and low energy requirements because multipliers can be implemented with only 6 transistors (Fig. 2) as opposed to the 3000 required for conventional computing, saving more than 90% in passive/active power and area. However, the spectacular savings offered by the unary computing paradigm usually fade away when considering the complete dataflow. In the previous example, the cost of moving and transforming the data to the multiplier's inputs with the expected format is prohibitively expensive [3].

**Potential:** One of the most tantalizing promises of brain inspired computing (neuromorphic) is the ability to do complex processing within an ultra-low power budget. For AI to be able to mimic this energy efficiency, we need to part ways from conventional digital Von-Neuman paradigms (one-fits-all) that were implicitly co-designed for fetch-execute data flows. Thus, if we demonstrate that we can combine analog and unary computing in never-before ways that meet the neuromorphic design criteria (i-v), we will get up to 50% savings in latency and 127X power savings thanks to the asynchronous (clockless) computing data flow and the simplified analog to streams interface [3,4,6]. This will be a disruptive technology that can pave the way for energy-efficient AI for real-time and intelligent computing at the edge, and in-memory computing in HPC systems. An experimental evidence of the approach will open new research avenues such as (i) Co-design of unary/analog architectures for real-time, intelligent, edge computing for applications such as microfluidics (Fig. 3), genomic sequencing, and micro-tomography (u-CT). (ii) Exploration of mixed resolution for accuracy-latency-energy tradeoffs. (ii) Unary/analog computing at supercomputing facilities for applications such as ML-density functional theory (ML-DFT) and linear algebra kernels. (iii) Unary computing for neuromorphic wireless and wired communications channels. (iv) Emergent technologies such as Magnetic Transfer Junctions (MTJs) for probabilistic computing. (v) Modeling of alternative computing paradigms to obtain metrics such as power, latency, throughput and process, voltage, and temperature variation.

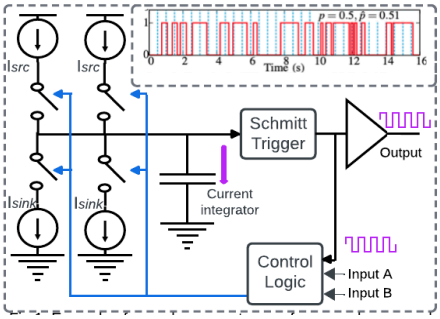



*Figure 1. Example of asynchronous stream of ones and zeros, and energy efficient mixed-signal exact-adder for streams. The adder is a current integrator (analog), but the output (asynchronous stream) has a digital amplitude [5]. It exploits the best of both the analog and digital worlds.*

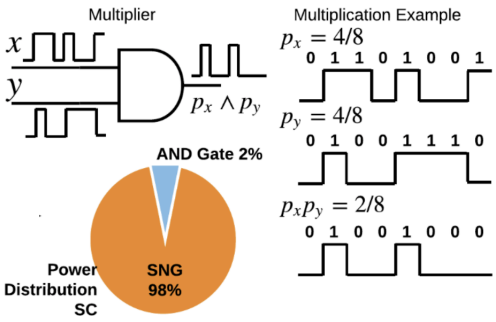



*Figure 2. In stochastic computing (SC), one of the unary computing variants, an approximate multiplier is implemented with an AND gate (top-left). 98% of the power consumption is due to the binary to streams conversion (SNG) [3]. Our approach aims to minimize/eliminate this conversion.*

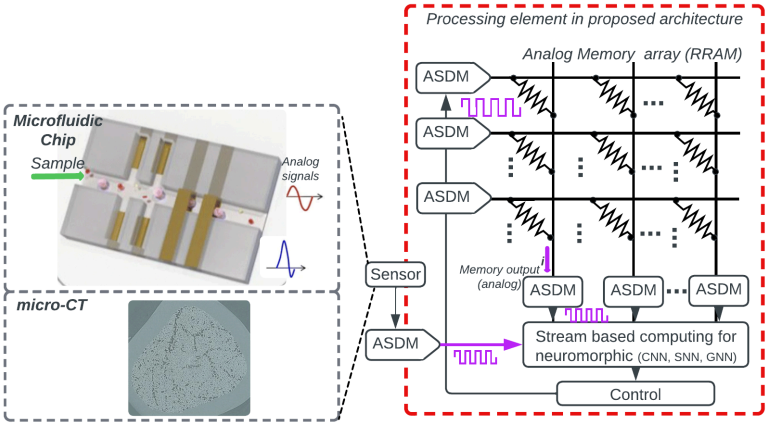



*Figure 3. An example of the proposed architecture used for edge processing. The analog memory's outputs and inputs are asynchronous streams. The memory's analog outputs are converted to streams with asynchronous sigma delta modulators (ASDM). The key innovations involve end-to-end streams and analog datapath (never in the digital domain) and collocation of memory and processing units, avoiding data transfer across memory hierarchies.*

**References**

**[1]** Balasubramanian. Proceedings of the National Academy of Sciences (2021).
**[2]** Schuman, Kulkarni, Parsa, Mitchell, & Kay. *Nature Computational Science* (2022).
**[3]** Gonzalez-Guerrero, Guo, & Stan. *IEEE ICRC* (2018).
**[4]** Gonzalez-Guerrero, Tracy, Guo, Sreekumar, Lenjani, Skadron, & Stan, *ACM JETC* (2020).
**[5]** Gonzalez-Guerrero, Guo, & Stan. *IEEE LASCAS* (2019).
**[6]** Gonzalez-Guerrero & Stan. IEEE *Asilomar (2019).*

# Unsupervised Online Learning in Photonic Neural Networks

Qing Gu, qgu3@ncsu.edu
Electrical and Computer Engineering, North Carolina State University, Raleigh, 27606, USA

Primary theme: Neuroscience-inspired computing principles
Secondary theme: Modeling and simulation approaches

## Introduction

With the aim to mimic the energy-efficient and robust information processing capabilities in the human brain, neuromorphic-inspired analog computing systems promise significant advantages in terms of computational power, parallelism, and energy efficiency for high-performance computing. Although light has not yet been widely used in information processing and computing, photonic hardware has substantial benefits over its electronic counterpart in that photonics enhances parallelism, increases bandwidth, and provides an energy-efficient solution for data movement[1].

## Approach

Over the years, various optical neurons and synapses have been proposed at the device level. However, demonstrations of system-level photonic neural networks (PNN) are rare. For large-scale integration, a system compatible with silicon photonics and CMOS foundry technology processes is preferable[2].

Recently, several PNNs utilizing Mach-Zehnder interferometer (MZIs)[3], phase-change materials (PCMs)[4,5], PIN attenuators[2], and metastructures[6] have been demonstrated at the system level. However, most of these PNNs can only perform inference or use the backpropagation method for supervised learning. Furthermore, they require additional electronic circuitry for loss function computations during training. This necessitates optical-electrical-optical (O-E-O) transitions, which significantly reduce energy efficiency[7]. In the meantime, training in the electrical domain presents two primary disadvantages: (1) it is highly dependent on the model accuracy of the physical system, and (2) the speed and power efficiency of electronic circuitry can become system bottlenecks, negating the advantages of photonic hardware.

What's more, labeled datasets used in supervised learning are not always readily available[4]. Therefore, in some application scenarios, e.g., when using drones, we have to opt for unsupervised learning as an alternative method. During unsupervised learning, the network automatically updates its synaptic weights, gradually adapting to specific patterns over time without relying on labeled data. To fully leverage the capabilities of photonic hardware, it is essential to develop an all-optical PNN with a suitable online learning method.

Based on the above arguments, we propose an all-optical PNN (Fig. 1), which includes local feedback and is capable of both supervised and unsupervised online learning. Our crossbar network can be constructed with non-volatile and reconfigurable PCM synapses and neurons. The local feedback effectively updates the network's synaptic weights according to the Hebbian rule, so unsupervised online learning can be realized.

## Conclusion

By eliminating the O-E-O transitions and training-associated electronic circuitry, our all-optical neural network – optimized for unsupervised online learning – is expected to achieve the projected high efficiency and compute density of PNN and explore the full potential of photonic hardware[1].

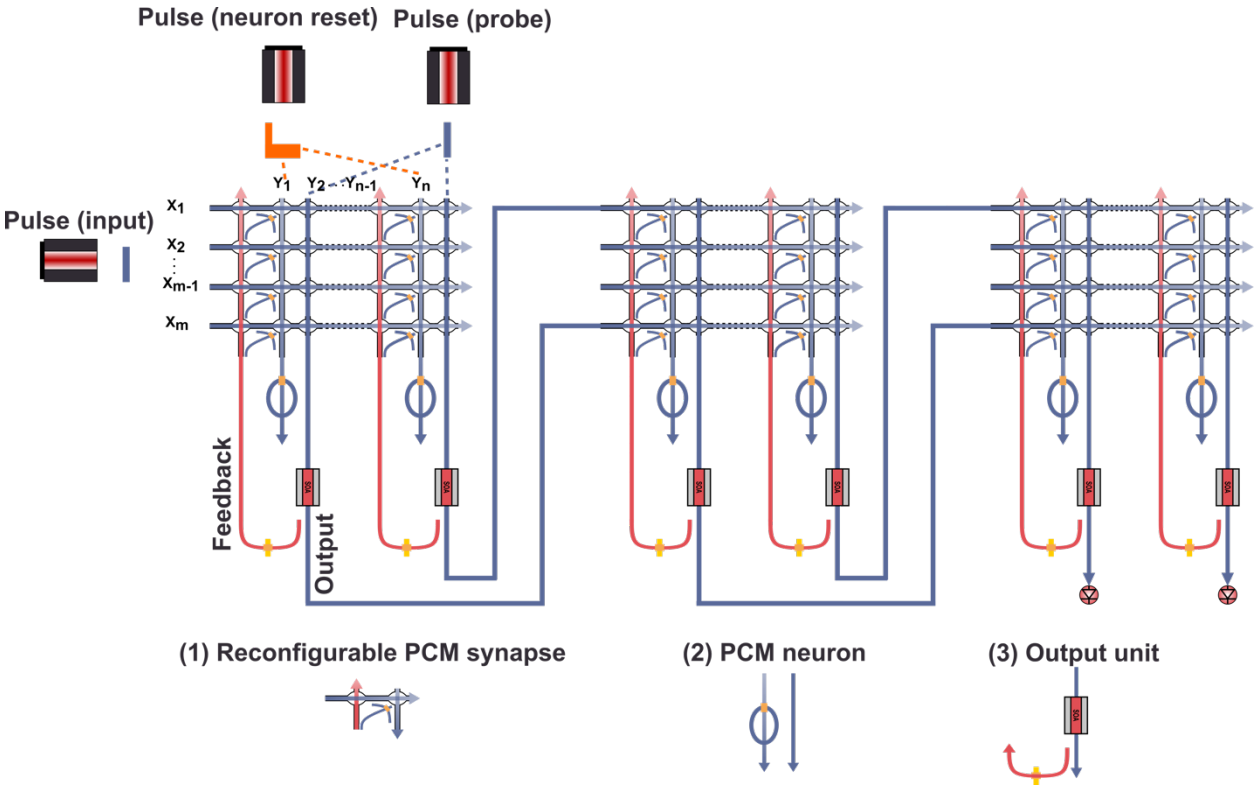



Fig. 1 Crossbar PNN for unsupervised learning that learns on the fly. The PCM cells are depicted as orange rectangles on top of the waveguide. Input signals (rows) are illustrated by blue arrows with decreasing color intensity, output signals (columns, downward direction) are depicted by blue arrows with increasing color intensity, and feedback signals (columns, upward direction) are represented by red arrows with decreasing color intensity. The decreasing color indicates a reduction in signal strength, while the increasing color indicates an increase in signal strength. The special two-step pulses (orange color) reset the PCM cell in the PCM neurons back to the crystalline state (off state).

**References**


1. Nahmias, M. A. *et al.* Photonic Multiply-Accumulate Operations for Neural Networks. *IEEE Journal of Selected Topics in Quantum Electronics* **26**, (2020).
2. Ashtiani, F., Geers, A. J. & Aflatouni, F. An on-chip photonic deep neural network for image classification. *Nature* **606**, 501–506 (2022).
3. Shen, Y. *et al.* Deep learning with coherent nanophotonic circuits. *Nat Photonics* (2017) doi:10.1038/nphoton.2017.93.
4. Feldmann, J., Youngblood, N., Wright, C. D., Bhaskaran, H. & Pernice, W. H. P. All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature* (2019) doi:10.1038/s41586-019-1157-8.
5. Feldmann, J. *et al.* Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, (2021).
6. Nikkhah, V. *et al.* Inverse-designed low-index-contrast structures on a silicon photonics platform for vector–matrix multiplication. *Nat Photonics* **18**, (2024).
7. Ma, X. *et al.* Photonics Multiply-Accumulation Computations System Based on Residue Arithmetic. *ACS Photonics* (2023).

# Harnessing Agent-Based Modeling and Evolutionary Algorithms for Scalable Heterogeneous Spiking Neural Network Co-Design

*Chathika Gunaratne, Oak Ridge National Laboratory*

**Introduction:** Neuromorphic computing has paved the way for energy-efficient and massively parallel computing, inspired by the neocortex, that challenge the scaling limits and "end of Moore's law" experienced by the Von Neumann architecture and CMOS technology [3]. Several neuron and synapse models for designing spiking neural networks (SNNs) on neuromorphic chips have been developed across a broad spectrum of biological plausibility and difficulty of implementation [10]. With the advent of beyond-CMOS circuit components and increasing scalability of neuromorphic chips [8, 5], it has become essential to explore the large space of possibilities regarding neuron and synapse circuit design, alongside neuron and synapse model selection and connectivity when co-designing heterogeneous SNNs. Mechanistic modeling and simulation methodologies such as agent-based modeling (ABM) combined with evolutionary algorithms (EAs) are well suited and scalable solutions for this task.

**Challenges:** An ideal neuromorphic computing simulator would be highly scalable and capable of accommodating primitives that describe neuron and synapse circuit components and model SNNs at two resolutions: 1) the neuron and synapse models resulting from neuron and synapse circuit components, and 2) the SNN architecture resulting from the networking of resulting neuron and synapse models. Such a simulator would produce highly scalable heterogenous SNNs, with multiple neuron and synapse models within the same network. The performance characteristics of the SNN, such as energy consumption, estimated hardware execution time, and accuracy, would be traceable to the circuit component primitives in addition to neuron and synapse model connectivity to enhance explainability. However, such an endeavor presents challenges in scalability, multi-tier modeling and simulation, and model interpretability that the current generation of neuromorphic simulators are not designed to accomplish. Unfortunately, the capabilities of extant simulators to seamlessly model heterogenous networks is highly limited, and the lack of explicit circuit mechanisms inhibits the ability to trace SNN performance characteristics to neuron and synapse circuit components. Furthermore, existing simulators do not make use of distributed GPU computation, limiting their ability to capitalize on high-performance computing (HPC) systems.

**Opportunities:** As neuromorphic hardware develops rapidly, large-scale heterogeneous neuromorphic architectures such as Darwin 3 [6], Loihi [2], and SpiNNaker [4] are now capable of computing millions of neurons with greatly improved synapse capacity, in addition to accommodating flexible neuron and synapse models. Searching the vast space of possible neuron and synapse circuits alongside optimal heterogenous SNN architectures is essential to optimizing the use of these platforms in the dimensions of scale, energy consumption, execution time, and accuracy. A recent neuromorphic simulation suite, SuperNeuro [1], employs ABM of spiking neural networks with the capability to scale on distributed GPU systems such as the Summit supercomputer, while providing the ability to model neuron and synapse circuit level mechanisms within agents. Mechanistic representations of neuron and synapse circuits based on hardware-emulating primitives, as provided by the ABM simulation paradigm, allow for device-scale traceability of SNN performance characteristics. This approach formulates a two-tier combinatorial problem that addresses both: 1) neuromorphic device selection for circuit co-design and, and 2) neuron and synapse model connectivity for SNN architecture search. EAs are highly scalable and suitable for combinatorial optimization and tools such as EONS use EAs for SNN architecture search on HPC [7, 9]. By expanding such EA frameworks to incorporate device-scale primitives to evolve heterogenous SNNs of mechanistic neuron and synapse models, the gap between algorithmic optimization and neuromorphic device selection during co-design may be further bridged.

**Acknowledgements**

The author is grateful to Thomas Potok for his guidance and to Ashish Gautam, Shruti Kulkarni, Prasanna Date, Robert M. Patton, and Mark Coletti for their input to this article and their efforts with developing SuperNeuro.

**References**

[1] Date, P., Gunaratne, C., R. Kulkarni, S., Patton, R., Coletti, M., & Potok, T. (2023, August). Superneuro: A fast and scalable simulator for neuromorphic computing. In *Proceedings of the 2023 International Conference on Neuromorphic Systems* (pp. 1-4).

[2] Davies, M., Wild, A., Orchard, G., Sandamirskaya, Y., Guerra, G. A. F., Joshi, P., Plank, P., & Risbud, S. R. (2021). Advancing neuromorphic computing with loihi: A survey of results and outlook. *Proceedings of the IEEE*, *109*(5), 911-934.

[3] Finocchio, G., Bandyopadhyay, S., Lin, P., Pan, G., Yang, J.J., Tomasello, R., Panagopoulos, C., Carpentieri, M., Puliafito, V., Åkerman, J. and Takesue, H. (2023). Roadmap for unconventional computing with nanotechnology. *Nano Futures* 8, no. 1 (2024): 012001.

[4] Höppner, S., Yan, Y., Dixius, A., Scholze, S., Partzsch, J., Stolba, M., Kelber, F., Vogginger, B., Neumärker, F., Ellguth, G., Hartmann, S., Schiefer, S., Hocker, T., Walter, D., Liu, G., Garside, J., Furber, S., & Mayr, C. (2021). The SpiNNaker 2 processing element architecture for hybrid digital neuromorphic computing. *arXiv preprint arXiv:2103.08392*.

[5] Hu, X., Hassan, N., Brigner, W. H., Chauwin, M., & Friedman, J. S. (2020, October). Device modeling and circuit design for scalable beyond-cmos computing. In *2020 IFIP/IEEE 28th International Conference on Very Large Scale Integration (VLSI-SOC)* (pp. 210-211). IEEE.

[6] Ma, D., Jin, X., Sun, S., Li, Y., Wu, X., Hu, Y., Yang, F., Tang, H., Zhu, X., Lin, P., & Pan, G. (2024). Darwin3: a large-scale neuromorphic chip with a novel ISA and on-chip learning. *National Science Review*, 11(5), nwae102.

[7] Plank, J. S., Schuman, C. D., Bruer, G., Dean, M. E., & Rose, G. S. (2018). The TENNLab exploratory neuromorphic computing framework. *IEEE Letters of the Computer Society*, *1*(2), 17-20.

[8] Potok, T., Schuman, C., Patton, R., Hylton, T., Li, H., & Pino, R. (2016). *Neuromorphic Computing, Architectures, Models, and Applications. A Beyond-CMOS Approach to Future Computing, June 29-July 1, 2016, Oak Ridge, TN*. USDOE Office of Science (SC)(United States). Advanced Scientific Computing Research (ASCR).

[9] Schuman, C. D., Mitchell, J. P., Patton, R. M., Potok, T. E., & Plank, J. S. (2020, March). Evolutionary optimization for neuromorphic systems. In *Proceedings of the 2020 Annual Neuro-Inspired Computational Elements Workshop* (pp. 1-9).

[10] Yamazaki, K., Vo-Ho, V. K., Bulsara, D., & Le, N. (2022). Spiking neural networks and their applications: A review. *Brain Sciences*, *12*(7), 863.

**Enablers for Analog 3D Spatio-Temporal Reconfigurable Computing in Neuromorphic Systems**

Technical Area: Technologies and Prototyping of Neuromorphic Analog Primitives

Akhilesh Jaiswal

Assistant Professor, Electrical and Computer Engineering, University of Wisconsin-Madison

Rich analog dynamic behavior of neurons and dense 3D analog interconnects are among the most critical features of biological brain systems. This position paper highlights key technology enablers to achieve dense reconfigurable analog neuromorphic computing and a potential pathway to achieve 'beyond-biology' capabilities in neuromorphic hardware systems.

**3D Spatio-Temporal Routing:** Novel 2.5D and 3D heterogeneous integration schemes are being actively explored for traditional digital computing. While these technologies are being hailed as data bandwidth-boosters for digital computing, they can also be repurposed for enabling dense 3-dimensional *analog* spatio-temporal data transfer in neuromorphic systems, beyond the capabilities of current 2D semiconductor platforms. 3-dimensional connections between layers of neurons are ubiquitously found both in sensory neural systems as well as brain systems. Leveraging 3D interconnections for analog data routing could eliminate the need for typical address event representation systems used in existing neuromorphic hardware. An initial example of such a neuromorphic 3D system is a new proposal on retina-inspired sensors call IRIS (Integrated Retinal Functionality in Image Sensors [1]) wherein multiple functional layers of retina can be mapped into 3D integrated chips using hybrid Cu-Cu bonding leveraging direct 3D analog data communication (this includes both analog and spike data) between various layers of neurons. Seamless communication of analog data onto 3D interconnects could be a potential hallmark of the next generation of neuromorphic systems.

Furthermore, recent advances in back-end-of-line (BEOL) devices including memory and transistors allow for reconfigurable interconnects embedded between 3D stacked chips. This in turn leads to the capability to dynamically modify the receptive field of neurons and synapses according to the needs of cognitive task being solved. For example, modifying the interconnections between bipolar and amacrine cells in a retinal neuromorphic hardware allows the hardware to rapidly switch between different visual features being extracted. This also leads to beyond-biology capabilities wherein the interconnections between neurons can be rapidly reconfigured in micro-/milli-second scale to suit the evolving nature of cognitive task at hand as opposed to biological counterparts that are adaptive but cannot be drastically reconfigured at rapid speeds.

**2.5D/3D Integrated Electro-Optics for Analog Neuromorphic Computing:** While small scale analog neuromorphic processors have been demonstrated, large scale analog neuromorphic processors that can rival the scale of computing provided by modern digital systems like GPUs have remained elusive. Among other issues long-distance analog communication of information is a critical bottleneck for scalable analog neuromorphic systems. Analog Optical Interconnects (AOI) can provide the much-needed technological pathway to enable long-distance analog interconnects [2]. As opposed to existing optical interconnects that are used as high-speed synchronous digital links, neuromorphic systems need novel asynchronous analog optical interconnects. Commercial advanced photonic platforms like Globalfoundries 45nm Photonics SPCLO technology allows *monolithic* co-integration of MOS transistors with active Silicon photonics devices providing the needed platform to explore co-integrated electro-optic solutions for neuromorphic computing. Further, new optical SRAM (static random access memory) using foundry compatible technology [3] can be used for high-speed retrieval of synaptic weights while also enabling high-bandwidth analog photonic in-memory computing.

Overall, future large-scale analog neuromorphic systems would deal with symbols (spikes or otherwise characteristic features of neuronal cells) as opposed to digital bits. 3D integration augmented with long-distance analog optical interconnects and high-speed memory could pave the pathway for scaling analog neuromorphic systems by facilitating seamless transfer of neuromorphic symbols. Key technology agnostic metrics like energy per symbol per unit length for a given fan-out would represent the efficiency of these complex neuromorphic systems.

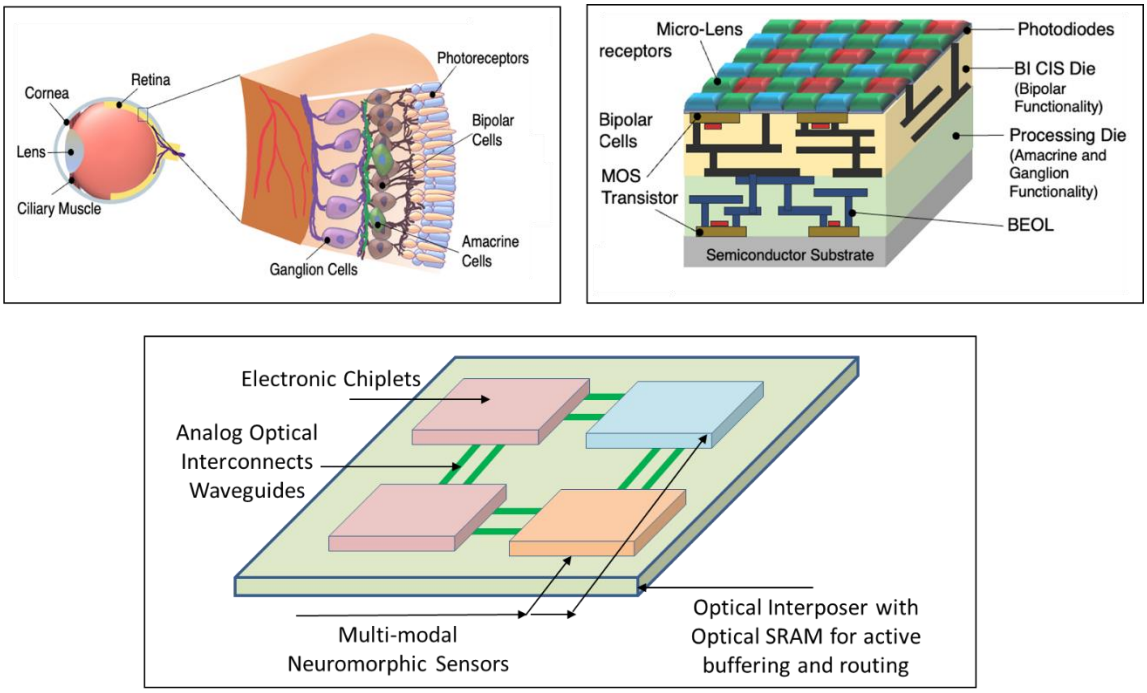



Figure: (Top Left) Biological Retina with its layered structure, (Top Right) Mapping of retinal layers into 3D integrated chip, (Bottom) Multi-chiplet neuromorphic hardware with long-distance analog optical interconnects.

**References:**

[1] Yin Z, Kaiser MA, Camara LO, Camarena M, Parsa M, Jacob A, Schwartz G, Jaiswal A. IRIS: Integrated retinal functionality in image sensors. Frontiers in Neuroscience. 2023 Sep 1;17:1241691.

[2] Jaiswal AR, Jacob AP, Bian Y, Rakowski M, inventors; GlobalFoundries US Inc, assignee. Optical neuro-mimetic devices. United States patent US 11,537,866. 2022 Dec 27.

[3] Kudalippalliyalil R, Chandran S, Jacob AP, Jaiswal A. Towards scalable, energy-efficient and ultra-fast optical sram. arXiv preprint arXiv:2111.13682. 2021 Nov 25.

# Flow-based Crossbar Computing and Neuronal Stochastic Dynamics

**Author:** Sumit Kumar Jha, Professor, Florida International University **Email:** sumit.jha@fiu.edu

**Theme(s):** Neuroscience algorithms and translation to neuromorphic analog circuits, Scalable integration for neuromorphic computing modeling

## Neuroscience-inspired Computing Principles

The intersection of flow-based computing in nanoscale 2-D and 3-D crossbars and stochastic neuroscience models is a fertile ground for innovations in neuromorphic computing. The research can leverage detailed models of neuronal dynamics to replicate the complexity and efficiency of biological computing systems. This involves moving beyond traditional integrate-and-fire models to incorporate Class 1, Class 2, and Class 3 neuronal behaviors. These behaviors correspond to different types of neuronal dynamics observed in biological systems:

- **Class 1 Neuronal Dynamics:** Neurons that move from their resting state to an active state through saddle node on an invariant circle bifurcation, encoding stochastic inputs.
- **Class 2 Neuronal Dynamics:** Neurons that lose stability through saddle-node off invariant cycle or Andronov-Hopf bifurcation, resulting in oscillatory responses with preferred frequency patterns.
- **Class 3 Neuronal Dynamics:** Neurons with a stable resting state that remains consistent over a range of inputs, essential for robust processing.

These principles can guide the development of neuromorphic architectures capable of performing probabilistic and approximate computations, which are critical for applications in machine learning, data analytics, and other error-tolerant domains.

## Translation to Analog Microelectronic Circuits

The translation of these neuroscience principles into analog microelectronic circuits involves the use of emerging nanoscale memory technologies, such as magnetic RAM (MRAM) as well as optical elements that can be fabricated in a crossbar structure. These technologies are characterized by their high speed, energy efficiency, and density, making them suitable for neuromorphic computing applications. The key steps for such a research agenda may include:

- **Design of Memristive Neuronal Circuits:** Creating circuits that replicate neuronal dynamics using memristors, which can exhibit various forms of oscillatory and stable behaviors that serve as building blocks for higher levels of abstractions.
- **Development of Stochastic Neuronal Computing Systems:** Utilizing flow-based computing through sneak paths in nanoscale crossbars coupled with memristive neuronal circuits to implement high-level tasks on compact 2-D and 3-D crossbar arrays.
- **Prototyping Neuromorphic Systems:** Integrating these circuits into functional prototypes capable of performing computational tasks with high efficiency, such as machine learning on scientific data sets.

## Modeling and Simulation Approaches

Modeling and simulation are crucial for understanding and optimizing the behavior of such bio-inspired neuromorphic circuits. This involves:

- **Algorithmic Synthesis:** Developing methods to automatically synthesize networks of neuronal stochastic dynamical systems that can be realized on nanoscale devices.
- **Simulation of Neuronal Dynamics:** Using stochastic differential equations and Andronov-Hopf bifurcations to model the behavior of memristive circuits and their interactions.
- **Performance Evaluation:** Simulating the performance of these neuromorphic systems on standard computational tasks, such as image processing and machine learning, to assess their efficiency.

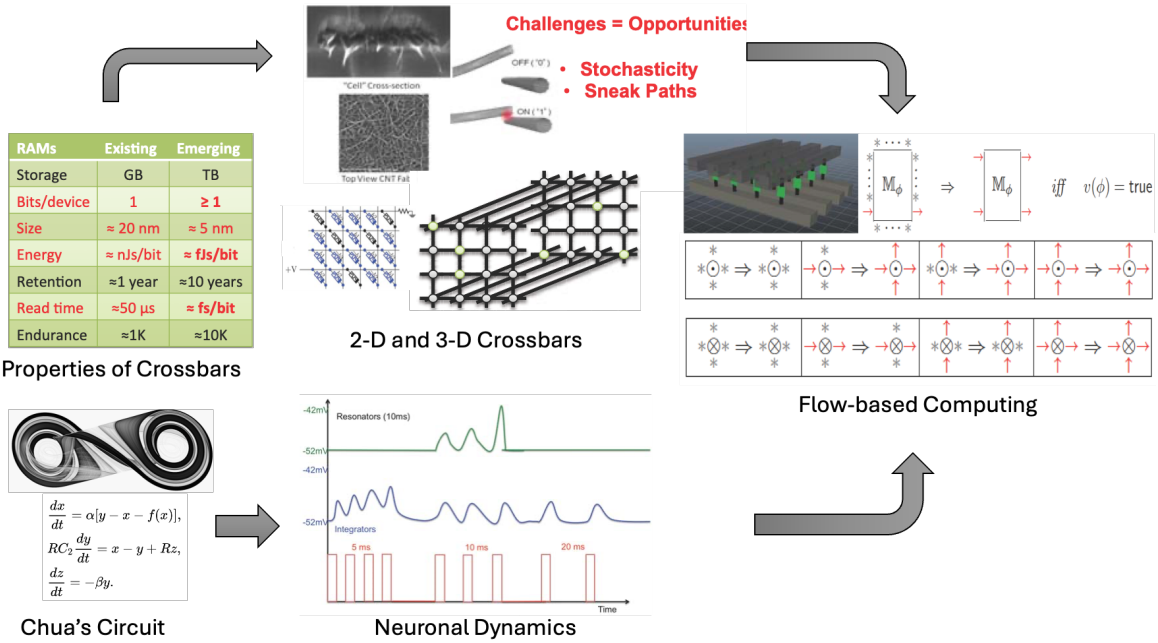



*Figure 1: High-density 2-D and 3-D crossbars with nonlinear elements like memristors can provide stochastic dynamics to replicate neural behaviors and enable high-level intelligence, exploiting device-level parallelism through sneak paths.*

References


1. Trappenberg, Thomas P. (2010). Fundamentals of Computational Neuroscience. United States: Oxford University Press Inc. pp. 2. ISBN 978-0-19-851582-1.
2. Patricia S. Churchland; Christof Koch; Terrence J. Sejnowski (1993). "What is computational neuroscience?". In Eric L. Schwartz (ed.). Computational Neuroscience. MIT Press. pp. 46–55. Archived from the original on 2011-06-04. Retrieved 2009-06-11.
3. Dayan P.; Abbott, L. F. (2001). Theoretical neuroscience: computational and mathematical modeling of neural systems. Cambridge, Mass: MIT Press. ISBN 978-0-262-04199-7.
4. Gerstner, W.; Kistler, W.; Naud, R.; Paninski, L. (2014). Neuronal Dynamics. Cambridge, UK: Cambridge University Press. ISBN 9781107447615.
5. Fan, Xue; Markram, Henry (2019). "A Brief History of Simulation Neuroscience". Frontiers in Neuroinformatics. 13: 32. doi:10.3389/fninf.2019.00032. ISSN 1662-5196. PMID 31133838.
6. Bharathwaj Muthuswamy, "Implementing memristor based chaotic circuits", International Journal of Bifurcation and Chaos, Vol. 20, No. 5 (2010) 1335–1350, doi:10.1142/S0218127410026514.
7. Leonov G. A.; Vagaitsev V. I.; Kuznetsov N. V. (2011). "Localization of hidden Chua's attractors" (PDF). Physics Letters A. 375 (23): 2230–2233.
8. Chua, Leon O.; Matsumoto, T.; Komuro, M. (August 1985). "The Double Scroll". IEEE Transactions on Circuits and Systems. CAS-32 (8). IEEE: 798–818. doi:10.1109/TCS.1985.1085791.
9. Thijssen, S., Rashed, M., Jha, S. K., & Ewetz, R. (2024). Equivalence Checking for Flow-Based Computing using Iterative SAT Solving. In 43rd International Conference on Computer-Aided Design (ICCAD) 2024.
10. Thijssen, S., Rashed, M., Jha, S. K., & Ewetz, R. (2024). Synthesis of Compact Flow-based Computing Circuits from Boolean Expressions. In 61st ACM Design Automation Conference (DAC), 2024.

# Robust Autonomy via Reward-Modulated Insect Circuitry

Erik Johnson
Johns Hopkins University Applied Physics Laboratory

**Robust Autonomy via Reward-Modulated Insect Circuitry**

Due to investments such as the US BRAIN Initiative, consortia of neuroscientists[1] are collecting nanoscale connectivity alongside functional and genomic data. These data give insight into the neural structure and function underlying model systems, such as the sensing, learning, and navigation systems of *Drosophila*[2]. Such behaviors require the interaction of many sub-systems for sensing, memory, and reward prediction, with different connectivity and learning rules, to enable behavior such as context-dependent navigation[3]. This includes circuitry for continual learning of sensory patterns in the mushroom body and the heading direction circuit, which fuses inertial and visual information to precisely estimate the organism's heading and state. An approach is required to bring together neuroinformatics, machine learning design, low-power robotics systems, and hardware/software co-design approaches to demonstrate feasibility of this goal[4]. A key challenge will be achieving continual learning, including context modulation, transfer learning, and avoiding catastrophic forgetting[5]. The availability of large-scale Insect connectomes and datasets provide an opportunity to study the connectivity between systems involved in sensory, memory, goal-directed action selection.

***Technical Approach.*** By building an integrated model of sensory processing, learning and memory, state estimation, and reward-modulated learning, we aim to demonstrate robust and adaptable autonomy suitable for low-power neuromorphic hardware. We focus on modeling heterogenous system components, including stereotyped connections in small neuron populations in the central complex and high-dimension random connections in the mushroom body. In particular, this approach relies upon: 1) large-scale connectivity analysis of *Drosophila* circuitry at the neuron-synapse level, 2) incorporation of connectivity insight into functional models of *Drosophila* subcircuits, 3) design of online learning rules suitable for implementation via neuromorphic hardware, and 4) testing and evaluation in continual learning scenarios.

Three large *Drosophila* hemisphere volumes are now available[2,6], containing critical circuits for learning (i.e., mushroom body) and state estimation (i.e., heading direction circuit). Large scale connectivity can be investigated with query tools, including discovery of repeated circuit motifs[7]. Results of these analyses can be incorporated into functional models of key insect subcircuits; for instance, we have previously demonstrated continual learning through a replay-based model incorporating details of mushroom body connectivity[8], and heading direction estimation through a dynamic neural circuit implementing ring attractor dynamics[9]. Critical extensions are required to incorporate context-dependent dopaminergic reward pathways[3] and to introduce inter-region connectivity. In order to utilize existing neuromorphic processor technologies, these networks need optimized dynamics and learning rules. We have shown the feasibility of our mushroom body model with binarized representations and local synaptic learning rules[10]. Further efforts are required to target existing neuromorphic hardware, including evolutionary design of hardware implementations utilizing specific platform constraints. An integrated system should be tested in sequences of tasks with different contexts (sensory cues, navigation paths, and distractors) and demonstrated on current neuromorphic processors.

***Benchmarks and Metrics.*** Key baselines include approaches for state estimation such as extended Kalman filters and visual inertial odometry, and action selection via shallow reinforcement learning policies. Individual subsystems can be benchmarked, such as the mushroom body continual learning circuit for image classification, to get single-system metrics. Overall testing can be conducted with a context-specific navigation task with changing sensory cues. Metrics include task performance (mean square error, classification accuracy, and reward), but also continual learning metrics built on single-task performance[11].

***Potential for Impact.*** The goal is to demonstrate insect-inspired autonomy with low complexity but high adaptability, which may be adapted to other sensing and decision-making tasks in dynamic and changing environments. This may show a pathway from large-scale neural data analysis to the implementation of heterogenous algorithms on low power hardware, providing a compelling alternative to the increasing complexity of monolithic machine learning systems.

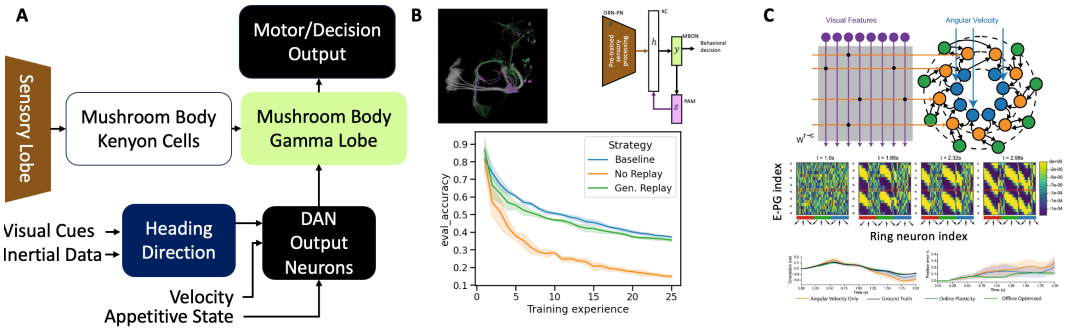



Figure 1. Panel A, overview of proposed insect-inspired neuromorphic system for continual learning in navigation tasks, including heading direction circuitry, with black boxes highlighting components for further development. Panel B represents results for continual learning for CIFAR 100 classification using a connectome-inspired replay strategy. Panel C shows results for heading direction estimation for an insect inspired network for visual feature and angular velocity estimation.


1. Johnson, E. C., Nguyen, T. T., Dichter, B. K., Zappulla, F., Kosma, M., Gunalan, K., ... & Yatsenko, D. (2023). A Maturity Model for Operations in Neuroscience Research. *arXiv preprint arXiv:2401.00077*.
2. Scheffer, L. K., Xu, C. S., Januszewski, M., Lu, Z., Takemura, S. Y., Hayworth, K. J., ... & Plaza, S. M. (2020). A connectome and analysis of the adult Drosophila central brain. *elife*, *9*, e57443.
3. Zolin, A., Cohn, R., Pang, R., Siliciano, A. F., Fairhall, A. L., & Ruta, V. (2021). Context-dependent representations of movement in Drosophila dopaminergic reinforcement pathways. *Nature neuroscience*, *24*(11), 1555-1566.
4. Johnson, E. C., Robinson, B. S., Vallabha, G. K., Joyce, J., Matelsky, J. K., Norman-Tenazas, R., ... & Hoffmann, J. A. (2023, June). Exploiting large neuroimaging datasets to create connectome-constrained approaches for more robust, efficient, and adaptable artificial intelligence. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications V* (Vol. 12538, pp. 394-405). SPIE.
5. Kudithipudi, D., Aguilar-Simon, M., Babb, J., Bazhenov, M., Blackiston, D., Bongard, J., ... & Siegelmann, H. (2022). Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence*, *4*(3), 196-210.
6. Dorkenwald, S., Matsliah, A., Sterling, A. R., Schlegel, P., Yu, S. C., McKellar, C. E., ... & FlyWire Consortium. (2023). Neuronal wiring diagram of an adult brain. *bioRxiv*.
7. Matelsky, J. K., Robinette, M. R., Wester, B., Gray-Roncal, W. R., Johnson, E. C., & Reilly, E. P. (2023). Data-driven motif discovery in biological neural networks. *bioRxiv*, 2023-10.
8. Robinson, B. S., Joyce, J., Norman-Tenazas, R., Vallabha, G. K., & Johnson, E. C. (2023). Informing generative replay for continual learning with long-term memory formation in the fruit fly. *bioRxiv*, 2023-01.
9. Robinson, B. S., Norman-Tenazas, R., Cervantes, M., Symonette, D., Johnson, E. C., Joyce, J., ... & Gray-Roncal, W. (2022). Online learning for orientation estimation during translation in an insect ring attractor network. *Scientific reports*, *12*(1), 3210.
10. Norman-Tenazas, R., Western, I., Vallabha, G., Roos, M. J., Johnson, E. C., & Robinson, B. S. (2023, August). Enabling local learning for generative-replay-based continual learning with a recurrent model of the insect memory center. In *Proceedings of the 2023 International Conference on Neuromorphic Systems* (pp. 1-7).
11. Baker, M. M., New, A., Aguilar-Simon, M., Al-Halah, Z., Arnold, S. M., Ben-Iwhiwhu, E., ... & Vallabha, G. K. (2023). A domain-agnostic approach for characterization of lifelong learning systems. *Neural Networks*, *160*, 274-296.

# Beyond Energy Efficiency: Neuromorphic Primitives for More General Metrics

Shubha Raj Kharel[1], Soumyajit Mandal[2], Sairam Sri Vatsavai[1], Shinjae Yoo[1], and Yihui Ren[1]
[1]CSI and [2]IO, Brookhaven National Laboratory
[1]{skharel,ssrivatsa,smandal,sjyoo,yren}@bnl.gov

**INTRODUCTION AND INSPIRATION:** While energy efficiency is showcased as the primary advantage of neuromorphic computing, with speed metrics like Matmul secondarily showcased, other practical metrics needed for realistic implementation—such as throughput, latency, chip area, scalability, reliability, noise tolerance, and integration complexity—are rarely considered [1]. Nonetheless, the advancement of scientific applications necessitates a comprehensive approach wherein these metrics are adequately optimized and synergistically balanced. For example, in particle physics, detectors require a fast detection of complex patterns within confined spaces. For practical use, brain-machine interfaces require increased speed and miniaturization. Scalability demands are also increasing, as evidenced by the recent increase in large language model use in science [2]. Neuromorphic systems are currently not sufficiently optimized for these practical performance metrics, which are becoming increasingly important. Neuroscience provides valuable insights into how to achieve these missing criteria. The brain's event-driven rapid identification of threats, responses, decisions, and motor signals, even in tiny insect brains, is one example of this. The brain's efficiency in tasks such as abstract reasoning and natural language processing is unparalleled by advanced artificial systems. This stark contrast emphasizes the need to dig deeper into neural mechanisms and build circuit primitives inspired by them.

**NEUROMORPHIC PRIMITIVES:** Existing neuromorphic primitives encompass a range of technologies, each with distinct advantages and limitations. Analog electronic circuits offer fast, energy-efficient processing but struggle with noise sensitivity and variability. Digital implementations are scalable but less efficient for low- and moderate-precision applications [3]. Interfacing between these domains remains challenging, with the energy and area budget of analog-digital and digital-analog converters often dominating overall resource usage [4]. Photonic integrated circuits excel in performing high-speed communication and highly parallel computation [5, 6] but have relatively poor area efficiency. Hybrid electro-photonic systems attempt to leverage strengths from both domains, but are often limited by the performance of the electro-photonic interfaces [7]. Computing in memory using non-volatile elements (such as RRAM) shows promise but lacks the maturity needed for complex applications. Despite these diverse approaches, no single solution currently achieves the combination of all practical metrics required to rival biological neural systems in complex, real-world applications, which motivates us to advocate the following research directions in different level of abstraction and implementation:

- **Neuronal primitives**: Identification of region-specific neural pathways in the brain for emulation is crucial. Research directions include 1) exploiting stochasticity in non-volatile memory [8], and 2) nonlinear photonics for performing complex, high-speed inference operations without sacrificing area [9].
- **Interface and encoding primitives**: Studies of noise resilience in biology [10] are required to improve sparse encoding and error correction methods for analog and hybrid analog-digital computing. Research into the attention and adaptation mechanisms of neurons in sensory pathways is also desirable [11].
- **Architectural primitives**: Development of improved 3D fabrication and integration techniques [12] is critical for emulating the brain's layered structure and massive neural connectivity. Other research directions include 1) implementing brain-inspired regional specialization within electronic circuits that emulates the heterogeneity of neural architectures [13], and 2) studying computational primitives based on large-scale synchronization, wave propagation, and other collective phenomena [14, 15].
- **Integration primitives**: One potential direction for research is to address the challenge of partitioning computations between analog, digital, and photonic domains and seamlessly interfacing them to leverage their respective strengths in speed, area efficiency, and scalability [16]. This could involve developing new signal conversion techniques, encoding methods, and hybrid computing paradigms.

**CONCLUSION:** While energy efficiency in neuromorphic computing remains crucial and promising, we advocate increased focus on neuromorphic computing primitives that collectively optimize implementation-focused metrics across multi-domain systems. This approach is crucial to address the rapidly evolving demands of scientific applications, which increasingly require neuromorphic solutions that are not only efficient, but also scalable and practical for real-world deployment.

## References


[1] Danijela Marković et al. Physics for neuromorphic computing. *Nature Reviews Physics*, 2020.

[2] Weixin Liang et al. Mapping the increasing use of LLMs in scientific papers. *arXiv preprint arXiv:2404.01268*, 2024.

[3] Rahul Sarpeshkar. Universal principles for ultra low power and energy efficient design. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2012.

[4] Boris Murmann. Mixed-signal computing for deep neural network inference. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2020.

[5] Sudip Shekhar et al. Roadmapping the next generation of silicon photonics. *Nature Communications*, 2024.

[6] Chaoran Huang et al. 3 - photonic computing: an introduction. In *Phase Change Materials-Based Photonic Computing*, Materials Today. Elsevier, 2024.

[7] Georgios Sinatkas et al. Electro-optic modulation in integrated photonics. *Journal of Applied Physics*, 2021.

[8] Robin Degraeve et al. Causes and consequences of the stochastic aspect of filamentary RRAM. *Microelectronic Engineering*, 2015.

[9] Massimo Borghi et al. Nonlinear silicon photonics. *Journal of Optics*, 2017.

[10] Pin Shu et al. The resilience and vulnerability of human brain networks across the lifespan. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2021.

[11] Clarissa J Whitmire et al. Rapid sensory adaptation redux: a circuit perspective. *Neuron*, 2016.

[12] Dingyou Zhang et al. 3D integration technologies: An overview. *Materials for Advanced Packaging*, 2017.

[13] Aaron Alexander-Bloch et al. Imaging structural co-variance between human brain regions. *Nature Reviews Neuroscience*, 2013.

[14] Jonathan Cannon et al. Neurosystems: brain rhythms and cognitive processing. *European Journal of Neuroscience*, 2014.

[15] Gyorgy Buzsaki. *Rhythms of the Brain*. Oxford University Press, 2006.

[16] Kaushik Sengupta et al. Terahertz integrated electronic and hybrid electronic–photonic systems. *Nature Electronics*, 2018.

# What are the Building Blocks for Neuro-Inspired Continual Learning?

Dhireesha Kudithipudi and Nicholas Soures
University of Texas at San Antonio

# What are the Building Blocks for Neuro-Inspired Continual Learning?

**Dhireesha Kudithipudi*, Nicholas Soures***

Biological processing has inspired the development of numerous learning algorithms and dedicated bio-inspired hardware. A more prominent example of this in neuromorphic computing is hebbian plasticity and spiking neurons that mimic local neuron and synapse dynamics to perform various degrees of intelligent behavior. However, the range of plasticity mechanisms observed in neuroscience is much broader and offers a fertile landscape for inspiration to neuromorphic computing researchers. Plasticity and learning are arguably among the most effective techniques the brain employs to perform processing. An active area of exploration is the ability to learn continually through experience. There is documented evidence on how structural and synaptic plasticity support continual learning in mammals and other species. The question then arises whether we can conceptualize such building blocks that facilitate learning. Can these mechanisms be practically implemented in hardware? This question is of great interest.

Various neural mechanisms have been identified that help retain knowledge while responding to new inputs without forgetting previous information. Capturing these mechanisms in hardware seems like a critical step towards increasing both immediate application of neuromorphic computing towards real-world problems, as well as advancing the level of intelligence displayed by these systems. Here, we present plasticity mechanisms that have been shown to support continual learning.

**Metaplasticity**, the plasticity of plasticity [1] aids in continual learning based on the premise that memories are stored by adjusting the strength of synapses in the brain [5]. This is achieved by regulating changes to synapses to prevent overwriting of prior knowledge.

**Synaptic consolidation** refers to the transition between early-phase long term potentiation(LTP) and late-phase LTP [7]. Biological synaptic weights involve multiple processes operating on different timescales. The interaction between these components can allow rapid integration of new information while deciding what to store permanently on a slower timescale.

**Neuromodulation** a phenemonon in which chemical signals modify how neurons process input signals [6] can enable context-dependent processing. This can facilitate sparse, distributed representations which are less prone to catastrophic interference.

**Replay** is the phenomenon that neuronal activity patterns that had previously occurred during waking are re-occurring during later sleep or rest. This replay was first observed in the hippocampus [10], and subsequently synchronously in the hippocampus and neocortical areas [3].

Each of these mechanisms can give rise to continual learning behaviors individually, but also integrated to form a more robust system, which protects and transfers knowledge throughout the lifetime of learning. Furthermore, combining these biological primitives for continual learning with homeostatic mechanisms such as synaptic scaling can allow learning in more challenging scenarios where the model needs to infer task boundaries between current data and previously experienced data to which it no longer has access (see 1). Based on these principles, preliminary neuromorphic hardware for continual learning [4] and analog designs that take advantage of the physics of memristor devices [11].

A new array of learning algorithms for neuromorphic computing systems can be generated by integrating these plasticity mechanisms inspired by biological phenomena.

# References

[1] W. C. Abraham. Metaplasticity: tuning synapses and networks for plasticity. *Nature Reviews Neuroscience*, 9(5):387–387, 2008.

[2] A. Daram and D. Kudithipudi. Neo: Neuron state dependent mechanisms for efficient continual learning. In *Proceedings of the 2023 Annual Neuro-Inspired Computational Elements Conference*, pages 11–19, 2023.

[3] D. Ji and M. A. Wilson. Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature Neuroscience*, 10(1):100–107, 2007.

| Method | DI MNIST | CI MNIST | DI FMNIST | CI FMNIST |
|---|---|---|---|---|
| SNN | 64.91% | 19.25% | 82.83% | 19.93% |
| NEO [2] | 78.14% | NA | 86.82% | NA |
| NACHOS [9] | 81.27% | 66.29% | 92.37% | 33.95% |
| PM [11] | 83.69% | NA | 93.23% | NA |
| Replay + Metaplasticity [8] | 74.98% | NA | 88.46% | NA |

Table 1: Overview of proposed neuro-inspired mechanisms for continual learning with spiking neural networks compared to state-of-the-art baseline models, evaluated for the MNIST and Fashion-MNIST datasets in domain incremental, class incremental continual learning settings. NACHOS is a task-agnostic continual learning framework that integrates metaplasticity, consolidation, decay, neuromodulation, and synaptic scaling using only local information. PM is a hardware-oriented design of metaplasticity that modulates the probability of a synapse undergoing plasticity based on inherent characteristics of memristor devices. NEO is a neuromodulation scheme for forming compartmentalized clusters of neurons and regulating their learning based on a neuron-level state of importance. Finally, a limited replay with a buffer of 2000 samples was used to evaluate the combination of replay and metaplasticity for efficient continual learning.

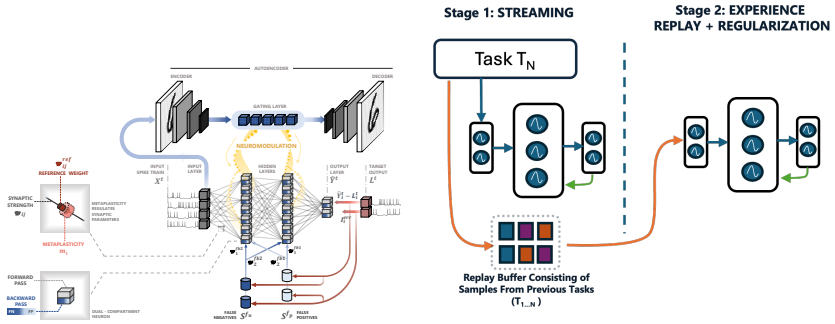



Figure 1: Left) General overview of continual learning SNN incorporating metaplasticity, consolidation, and neuromodulation to regulate learning. Right) High-level overview of replay with continual learning SNN. Local mechanisms are active during streaming learning. Replay adds a second learning stage during which knowledge is consolidated across all experiences


[4] V. Karia, F. T. Zohora, N. Soures, and D. Kudithipudi. Scolar: A spiking digital accelerator with dual fixed point for continual learning. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1372–1376. IEEE, 2022.

[5] J. J. Langille and R. E. Brown. The Synaptic Theory of Memory: A Historical Survey and Reconciliation of Recent Opposition. *Frontiers in Systems Neuroscience*, 12, 2018.

[6] E. Marder and V. Thirumalai. Cellular, synaptic and network effects of neuromodulation. *Neural Networks*, 15(4-6):479–493, 2002.

[7] R. G. M. Morris. Long-term potentiation and memory. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, 358(1432):643–647, Apr. 2003.

[8] S. N. Patel, Raghav and D. Kudithipudi. Does replay suffice for online continual learning in spiking networks? In *Cognitive Computational Neuroscience 2024*, 2024.

[9] N. Soures. *Lifelong Learning in Spiking Neural Networks Through Neural Plasticity*. Rochester Institute of Technology, 2023.

[10] M. A. Wilson and B. L. McNaughton. Reactivation of hippocampal ensemble memories during sleep. *Science*, 265(5172):676–679, 1994.

[11] F. T. Zohora, V. Karia, N. Soures, and D. Kudithipudi. Probabilistic metaplasticity for continual learning with memristors. *arXiv preprint arXiv:2403.08718*, 2024.

# Transformers-enabled Discovery of Neuromorphic Circuit Primitives from Large Scale Network Simulations

*Shruti R. Kulkarni, Seung-Hwan Lim, Oak Ridge National Laboratory*

**Introduction:** Human brain is capable of efficiently carrying out complex cognitive processing, with nearly hundreds of billions of neurons and trillions of synapses as the basic processing primitives. Simulations with millions of spiking neurons and billions of connections with plasticity have been demonstrated by various neuromorphic simulators on supercomputers [1], [2]. While deep learning models typically have layered networks, which can leverage the parallelization capabilities of Graphical Processing Units (GPUs), Spiking Neural Networks (SNN) topologies have primarily been small-world or have random connectivity between the neurons and operate with spare events, which poses a challenge in using matrix-vector accelerators like GPUs. SNN simulations with limited bio-plausibility have been demonstrated on heterogeneous platforms with GPUs, and Field Programmable Gate Arrays (FPGAs) [3], [4]. However, tools for interpreting the data from these large models, simulated on High Performance Computing (HPC) systems, are currently lacking. One way is to model the causal structure of different regions of such networks [5], [6] so that interpretable reduced order learning models can be developed, which can then enable the realization of neuromorphic circuit primitives for applications at various scales.

**Challenges:** There are two key challenges in the current neuromorphic simulation frameworks, one being the ability to simulate diverse topologies of large scale SNNs [7], and second, being able to parse these simulation results to develop interpretable learning models for real-world applications. There is always a performance mismatch between the software trained models to when they are realized on the hardware, thus, there is a growing need to codesign neuromorphic algorithms and hardware circuit topologies [8]. However, analog circuit design is a very laborious and time-consuming process, with dependance on costly design tools. Neuromorphic simulators, while being able to simulate large-scale SNNs on HPC systems, would greatly enhance the codesign effort by incorporating workflows to map the causal structure of these event based SNN dynamics with that of low-level analog circuits.

**Opportunities**: With the rapid emergence of generative Artificial Intelligence (AI) models, brings an opportunity to leverage these techniques for discovering novel learning models from large scale simulations of bio-plausible networks. Studies have shown that transformer models can learn the causal association between data sequences through their attention mechanisms [5] [9]. Transformer have also shown their potential to serve as surrogate models for analog circuits and accelerate the design search space [10], [11]. This presents a unique opportunity for the neuromorphic community to enhance the SNN simulation capabilities with causality modeling capabilities provided by transformers to solve the constrained optimization problem in analog design space with low-level circuit and post-CMOS device abstractions in an accelerated manner (see Figure 1). Recent simulators such as SuperNeuro [12], can simulate SNNs at scale by the agent-based modeling framework, which presents an opportunity to incorporate the dynamics of circuit elements as easily interpretable agent mechanisms, or as part of constraints within the neuronal dynamics. Hence, there is an opportunity to create a workflow wherein large-scale simulation tools are embedded with recent deep learning frameworks, that can infer the causal relationships between the simulated neural dynamics and those of analog circuit primitives, thereby, accelerating the codesign of efficient analog and mixed signal hardware platforms at scale.

**Benchmarking and Dataset Requirements:** For the comprehensive design of a simulation platform that accelerates codesign of future neuromorphic platforms, we would need to benchmark both the simulation software and the transformer attention model in terms of accuracy, scalability, and speedup. Further, the data generated from SNN simulations to train the transformer models, could be curated as a database for further studies in this direction. This effort also requires establishing a methodology to create a database of circuit and device dynamics from analog circuit solvers and their behavioral models that can be integrated in the neuromorphic software simulators.

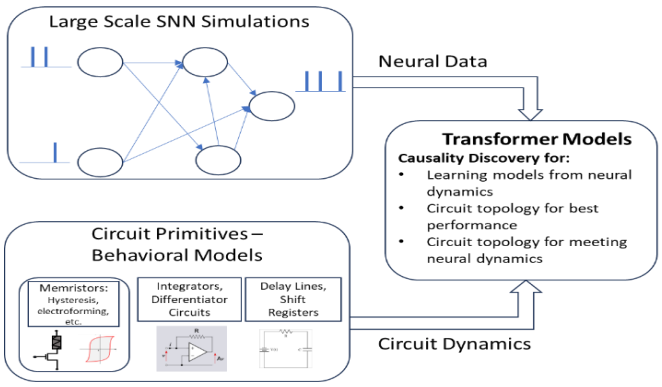



*Figure 1. Attention enhanced neuromorphic simulators with the behavioral models of circuits will aid in designing circuit primitives capable of demonstrating bio-inspired dynamics from large scale SNN simulations.*

**Acknowledgements**:
Authors greatly acknowledge Chathika Gunaratne, Prasanna Date, Anika Tabassum, Catherine Schuman, Robert Patton, and Thomas Potok for their valuable discussions and inputs to this paper.

**References:**

[1] F. Wang, S. Kulkarni, B. Theilman, F. Rothganger, C. Schuman, S. H. Lim and J. B. Aimone, "Scaling neural simulations in STACS," Neuromorphic Computing and Engineering, 2024.

[2] D. M. Gewaltig M-O, "NEST Neural Simulation Tool," Scholarpedia, 2007.

[3] J. C. Knight and T. Nowotny, "GPUs outperform current HPC and neuromorphic solutions in terms of speed and energy when simulating a highly-connected cortical model," Frontiers in neuroscience, 2018.

[4] J. P. Mitchell, C. D. Schuman, R. M. Patton and T. E. Potok, "Caspian: A neuromorphic development platform," in Proceedings of the 2020 Annual Neuro-Inspired Computational Elements Workshop, 2020.

[5] Z. Liu, A. Tabassum, S. R. Kulkarni, L. Mi, J. N. Kutz, E. Shea-Brown and S.-H. Lim, "Attention for Causal Relationship Discovery from Biological Neural Dynamics," *arXiv preprint arXiv:2311.06928.*, 2023.

[6] S. R. Kulkarni, A. Tabassum, S. H. Lim, C. D. Schuman, B. Theilman, F. Wang, F. Rothganger and B. Aimone, "Explaining Neural Spike Activity for Simulated Bio-plausible Network through Deep Sequence Learning," in *2024 Neuro Inspired Computational Elements Conference (NICE)*, 2024.

[7] S. R. Kulkarni, M. Parsa, J. P. Mitchell and C. D. Schuman, "Benchmarking the performance of neuromorphic and spiking neural network simulators," *Neurocomputing*, pp. 145-160, 2021.

[8] "DOE Basic Research Needs for Neuromorphic Computing Pre-workshop report," 2024.

[9] V. Melnychuk, D. Frauen and S. Feuerriegel, "Causal transformer for estimating counterfactual outcomes," in *International Conference on Machine Learning, PMLR*, 2022.

[10] S. Poddar, Y. Oh, Y. Lai and H. Zhu, "INSIGHT: Universal Neural Simulator for Analog Circuits Harnessing Autoregressive Transformers," *https://arxiv.org/html/2407.07346v2*, 13 July 2024.

[11] F. Jiao, H. Li and A. Doboli, "Modeling and Extraction of Causal Information in Analog Circuits," *IEEE TCAD*, pp. 1915-1928, 2018.

[12] P. Date, C. Gunaratne, S. R. Kulkarni, R. Patton, M. Coletti and T. Potok, "Superneuro: A fast and scalable simulator for neuromorphic computing," in *Proceedings of the 2023 International Conference on Neuromorphic Systems*, Sant Fe, 2023.

# The Need for Beyond-Backpropagation AI Training Algorithms

Suhas Kumar
Sandia National Laboratories

**POSITION PAPER**: THE NEED FOR BEYOND-BACKPROPAGATION AI TRAINING ALGORITHMS

Suhas Kumar

The energy spent on training increasingly large artificial intelligence (AI) models has grown at an alarmingly exponential rate (Fig. 1),[1,2] raising infrastructural, financial, and computational costs.[3–5] For example, training the model for Chat-GPT consumed >1000 MWh of energy (equivalent to >500 tons of $CO_2$ emissions),[3,4] a scale of cost that is unaffordable to most institutions. Despite growing demand for (and dependence on) AI, the high cost of training AI has slowed its growth since 2021.[1,6] Fundamentally, this inefficiency is due to the use of the backpropagation algorithm (i.e. high-precision, incremental, first-order analytical math) implemented on energy-hungry digital/CMOS graphics processing unit (GPU) hardware in massive data centers.[1,5,6] To support this crucially required technology and reduce its environmental footprint, we need a radical new **energy-efficient approach**.

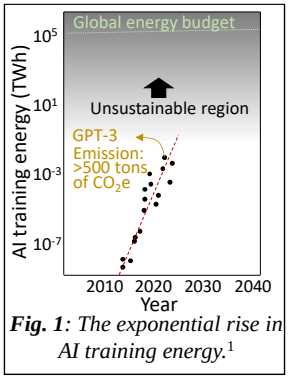



***Fig. 1**: The exponential rise in AI training energy.*[1]

Such an approach might be found in the human brain, which learns quickly at very low power (~20 W) and in a constrained substrate (e.g. low-precision data, shallow layers, presence of stochasticity). Neuroscience suggests that the brain's efficiency can be mimicked by numerical algorithms that possess additional degrees of freedom (e.g., by employing second-order derivatives/curvatures), thereby converging quickly and non-incrementally (Fig. 2).[7–14] We term this class **bio-plausible numerical optimization learning algorithms** (**BiNOLA**s). On a map of learning algorithms, backpropagation is a very small corner (Fig. 2), with BiNOLAs occupying the vast majority of an unexplored space. In other words, we have barely scratched the surface of AI learning algorithms.

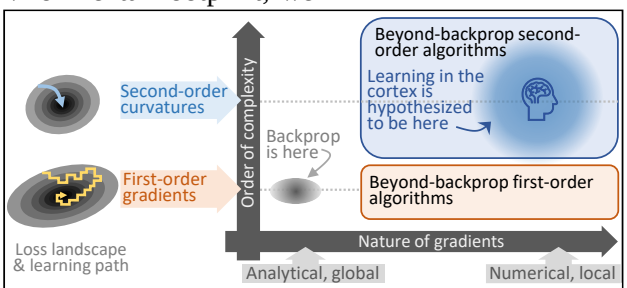



***Fig. 2**: Beyond-backpropagation learning algorithms.*

Although BiNOLAs have been studied by neuroscientists and mathematicians, they have not been commercially successful because their underlying math (such as matrix inversion) does not scale well in existing digital/CMOS hardware, which is bottlenecked by data movement. In a parallel development, inexorable new post-digital (especially **analog**) hardware has significantly resolved the data bottleneck by combining memory and processing, leading to more energy-efficient matrix multiplication[15] (and, recently, matrix inversion).[16,17] But such hardware also poses severe non-idealities (e.g., programming errors, nonlinearity, asymmetry, noise, and variability), which make it unsuitable for the established backpropagation algorithm.[18,19] Theoretically, post-digital hardware could efficiently support BiNOLAs (which are more resilient to noise and low precision), but there have been few efforts to tailor BiNOLAs to (and implement them on) post-digital hardware.

Ongoing reliance on backpropagation in CMOS hardware is untenable; to enable radically cheap and efficient AI training, we urgently need to develop BiNOLAs that offer efficiency via both algorithmic advantages and compatibility with efficient post-digital analog hardware.

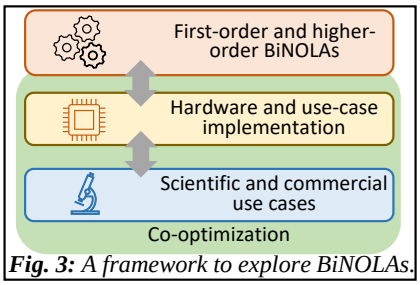



***Fig. 3:** A framework to explore BiNOLAs.*

**How do we do it?** Unlike the digital CMOS era, in the post-digital era, algorithms and hardware are going to be tailor-made to work with each other, and with a specific application. We identify three key challenges (C) that need to be addressed via co-design: **(C1)** discovering and scaling the fundamental math of first- and higher-order BiNOLAs, (**C2**) developing post-digital hardware architectures that work with BiNOLAs, (**C3**) identifying broad scientific and commercial use cases.

1. A. A. Conklin and S. Kumar, *Solving the Big Computing Problems in the 21st Century*, Nature Electronics 6, 464 (2023).
2. J. Sevilla, L. Heim, A. Ho, T. Besiroglu, et al., *Compute Trends Across Three Eras of Machine Learning*, in *2022 International Joint Conference on Neural Networks (IJCNN)* (2022), pp. 1–8.
3. D. Patterson, J. Gonzalez, Q. Le, C. Liang, et al., *Carbon Emissions and Large Neural Network Training*, arXiv:2104.10350 (2021).
4. K. G. A. Ludvigsen, *The Carbon Footprint of ChatGPT*, https://towardsdatascience.com/the-carbon-footprint-of-chatgpt-66932314627d.
5. A. de Vries, *The Growing Energy Footprint of Artificial Intelligence*, Joule 7, 2191 (2023).
6. N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, *Deep Learning's Diminishing Returns: The Cost of Improvement Is Becoming Unsustainable*, IEEE Spectrum 58, 50 (2021).
7. D. J. Felleman and D. C. Van Essen, *Distributed Hierarchical Processing in the Primate Cerebral Cortex*, Cereb Cortex 1, 1 (1991).
8. T. P. Lillicrap, A. Santoro, L. Marris, C. J. Akerman, et al., *Backpropagation and the Brain*, Nat Rev Neurosci 21, 6 (2020).
9. A. Meulemans, F. Carzaniga, J. Suykens, J. Sacramento, et al., *A Theoretical Framework for Target Propagation*, in *Advances in Neural Information Processing Systems*, Vol. 33 (2020), pp. 20024–20036.
10. R. Kanai, Y. Komura, S. Shipp, and K. Friston, *Cerebral Hierarchies: Predictive Processing, Precision and the Pulvinar*, Philosophical Transactions of the Royal Society B: Biological Sciences 370, 20140169 (2015).
11. J. G. Daugman and C. J. Downing, *Demodulation, Predictive Coding, and Spatial Vision*, J. Opt. Soc. Am. A, JOSAA 12, 641 (1995).
12. D. I. Fournier, H. Y. Cheng, S. Robinson, and T. P. Todd, *Cortical Contributions to Higher-Order Conditioning: A Review of Retrosplenial Cortex Function*, Front. Behav. Neurosci. 15, (2021).
13. A. Granier, M. A. Petrovici, W. Senn, and K. A. Wilmes, *Confidence and Second-Order Errors in Cortical Circuits*, arXiv:2309.16046 (2024).
14. T. P. Todd, R. Huszár, N. E. DeAngeli, and D. J. Bucci, *Higher-Order Conditioning and the Retrosplenial Cortex*, Neurobiology of Learning and Memory 133, 257 (2016).
15. J. D. Kendall and S. Kumar, *The Building Blocks of a Brain-Inspired Computer*, Applied Physics Reviews 7, 011305 (2020).
16. W. Zhang, B. Gao, J. Tang, P. Yao, et al., *Neuro-Inspired Computing Chips*, Nature Electronics 3, 7 (2020).
17. Z. Sun, G. Pedretti, E. Ambrosi, A. Bricalli, et al., *Solving Matrix Equations in One Step with Cross-Point Resistive Arrays*, Proceedings of the National Academy of Sciences 116, 4123 (2019).
18. Q. Xia and J. J. Yang, *Memristive Crossbar Arrays for Brain-Inspired Computing*, Nat. Mater. 18, 4 (2019).
19. C. Sung, H. Hwang, and I. K. Yoo, *Perspective: A Review on Memristive Hardware for Neuromorphic Computation*, Journal of Applied Physics 124, 151903 (2018).

# A Brain-Inspired-Attention-Based Spiking-Driven Neuromorphic System with Compute-in-Memory Design

**Hai (Helen) Li, Clare Boothe Luce Professor, Duke University**

Attention-based models, inspired by brain mechanisms of selective concentration and dynamic allocation, have shown their excellence in capturing relevant information and improving performance on various cognitive tasks with deep learning techniques. However, in energy-constrained scenarios like edge computing, conventional attention-based hardware adopting clock-driven computation suffers from power-consuming synchronous computation and dense data transmission. Alternatively, bio-inspired spiking neural networks (SNNs) enable sparse and asynchronous processing, and an optimized neuromorphic design promises the synergy effect for hardware efficiency. By fusing selective processing capabilities of attention mechanisms, a spiking-driven neuromorphic system will boost energy efficiency to an unprecedented level while preserving biological plausibility.

We aim an energy-efficient compute-in-memory (CIM) neuromorphic system for bio-inspired attention-based spiking neural networks (SNNs) with CMOS and emerging resistive random-access memory (RRAM) technology. Pursuing resource efficiency for data- and compute-intensive sequential processing applications in DOE, e.g., large-scale energy prediction, the proposed system incorporates cross-layer solutions at algorithm, architecture, and circuit levels including: (1) bio-plausible attention-based SNNs with efficient encoding schemes for substantial improvement of energy efficiency; (2) a dedicated neuromorphic CIM architecture that fully leverages the efficiency inherent in SNNs and thereby surpass conventional architecture; and (3) *in situ* RRAM/CMOS-based processing element (PE).

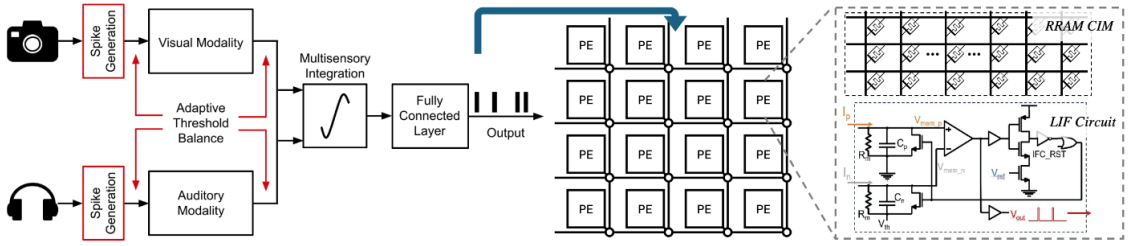



*Figure 1. Holistic system design with the cross-layer development of (a) an SNN algorithm with biological attention, (b) 2D mesh neuromorphic CIM architecture, and (c) CMOS/RRAM-based PE circuit.*

The multi-head attention mechanism has become a powerful solution for sequential data processing, such as language and bioinformatics. Bio-plausible SNNs have the potential as a breakthrough for further efficiency in the attention mechanism by embracing the sparse nature of biological computing: neurons in SNNs communicate with each other by discrete electrical spike signals and, correspondingly, can save hundreds of millions of operations per second. One of the key theories is that neurons are regulated by not a single but multiple learning rules. Hence, a possible approach for the implementation involves giving neurons adaptive spiking thresholds based on the maximum current the layer has received.

Tackling the inadequacy of existing computer architectures, which do not mirror the intricacies of our brain-like algorithm, our neuromorphic architecture tailored for SNNs with attention mechanisms, to fully harness the energy efficiency inherent in SNNs. A parallelizable 2D mesh interconnect of CIM architecture will constitute the neuromorphic system without the barrier between heterogeneous units. The adaptive execution by dynamic routers and controls with CIM will efficiently handle the challenges in attention-based SNNs, i.e., irregular sparsity and dynamics.

The neuromorphic PE is built with RRAM-CIM synaptic arrays and CMOS circuits that implement data encoding, neuron dynamics, and learning capability. CMOS neurons support various neuron models that effectively mimic the neuron dynamics in biological systems, such as the leaky integrate-and-fire (LIF) neuron with μW-scale power, taking spikes from excitatory and inhibitory synapses, integrating the membrane potential, and firing post-synaptic spikes. Such a circuit design supports differential operations from excitatory and inhibitory synapses, significantly simplifying the peripherals.

# Hypothesis-Driven Applications, Neuromorphic Circuits, and Technology Co-Design

Peng Li, Dept. of Electrical and Computer Engineering, UC Santa Barbara, Email: lip@ucsb.edu

Xiaoning Qian, Byung-Jun Yoon, Nathan Urban, Applied Mathematics, Computational Science Initiative, Brookhaven National Laboratory, Emails: xqian1@bnl.gov, byoon@bnl.gov, nurban@bnl.gov

**Primary Theme**: Translation to analog microelectronic circuits
**Secondary Theme:** Neuroscience-inspired computing principles

The capabilities and efficiency of analog neuromorphic circuits hold strong promise for revolutionizing brain-inspired computing. However, a significant challenge arises from a lack of systematic, joint exploration of end applications, neuromorphic circuits, and technology. Emulating the functionality of specific neural circuits and brain regions, such as the cortex, hippocampus, and thalamus, offers a rich playground for realizing various brain-inspired learning mechanisms [1, 2]. However, there are gaps in translating these learning principles into practical machine learning applications. Key challenges and research questions include: (1) how can formal optimization approaches be used to improve the co-design of hardware, circuits or networks, and algorithms over the enormous design space, and (2) how can uncertainty and probabilistic reasoning be incorporated into both the operation and design of neuromorphic computing platforms, and more generally (3) how can the neuromorphic computing design task be organized into a closed-loop process of hypothesis generation, testing, and revision?

As shown in Figure 1, to address the above challenges, it is crucial to develop methodologies that enable well-orchestrated co-design of applications, neuromorphic circuits, and technology, driven by brain-inspired learning hypotheses, for which fundamental brain learning principles can be mapped to a library of analog neuromorphic circuit primitives. The efficiency of these primitives can be enhanced through joint optimization of circuit design and the underlying device technology [3]. The resulting neuromorphic system, built from this library of optimized analog primitives, should then be assessed against real-world targets to identify potential performance gaps. Feedback from these assessments will drive revisions of the learning hypothesis, potentially coupled with *in-vivo* experimental data and evidence. Concepts from optimal experimental design and active learning [4, 5] can be used to more quickly identify limitations in the optimized neuromorphic system and accelerate the pace of design improvement.

Beyond optimizing the targeted analog primitives, the design of their interconnections-– potentially inspired by connectivity patterns in neural circuits such as those in the visual and auditory cortices and hippocampus—should also be considered. This integrated exploration strategy is essential for developing analog neuromorphic systems that embed brain-inspired learning principles with practical, cross-cutting implications. Generating and further optimizing such system designs can leverage recent advances in generative AI and machine learning (ML) for efficiently exploring the huge design space.

Uncertainties are inherent when co-designing neuromorphic circuits and technologies for specific applications. To address this, Bayesian learning provides a powerful approach. With the natural inherent stochasticity of analog neuromorphic circuits, e.g., based on the nanoscale behavior of memristors [3], Bayesian inference and learning capabilities can be developed for more energy-efficient uncertainty quantification (UQ). Brain-inspired Bayesian and non-Bayesian learning mechanisms and perceptual decision making [6-8], together with objective-driven uncertainty quantification, may open the door to more efficient probabilistic reasoning in neuromorphic systems and co-design of neuromorphic circuits, neuromorphic computing algorithms, AI/ML models, as well as underlying device/system technologies. It allows for a systematic and efficient exploration of information flow and co-optimization of key learning, circuit, and device parameters throughout the co-design process. The envisioned holistic methodology accounts for uncertainties arising from limited knowledge of learning principles, device characteristics and variability, and incomplete design exploration.

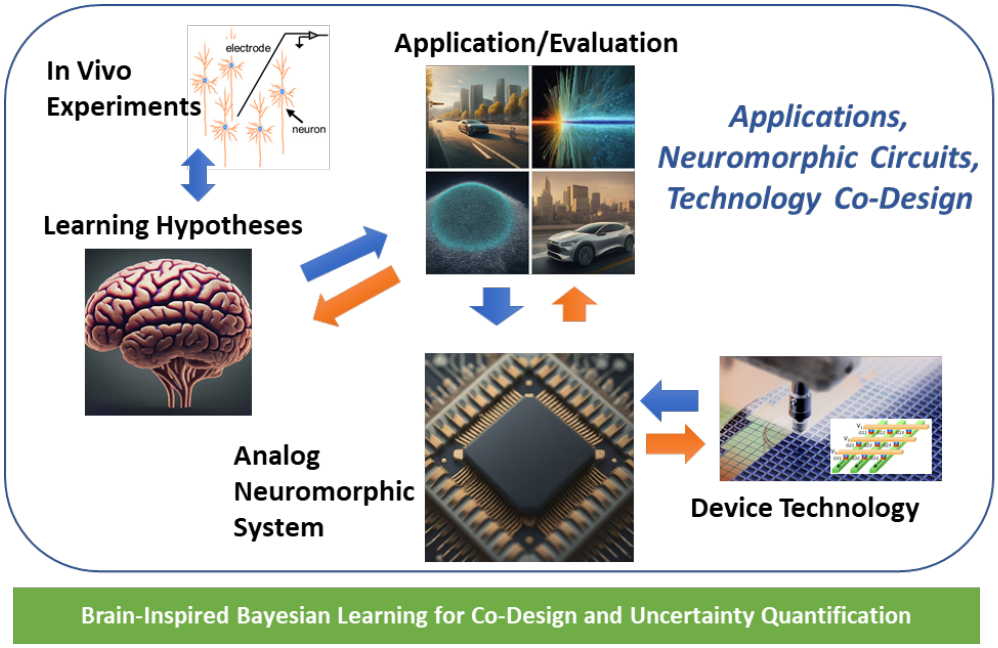



**Figure 1. Overview of Hypothesis-Driven Applications, Neuromorphic Circuits, Technology Co-Design with Bayesian Learning**

## References


[1] Gidon, A., et al., Dendritic action potentials and computation in human layer 2/3 cortical neurons. Science 367: 83-87(2020). DOI:10.1126/science.aax6239

[2] Poirazi, P., Brannon, T., Mel, B.W., Pyramidal Neuron as Two-Layer Neural Network. Neuron, 37(6): 989-999 (2003).

[3] Sarwat, S.G., Moraitis, T., Wright, C.D. et al. Chalcogenide optomemristors for multi-factor neuromorphic computation. Nat Commun 13: 2247 (2022).

[4] Qian, X., Yoon, B.J., Arróyave, R., Qian, X., Dougherty, E.R. Knowledge-Driven Learning, Optimization, and Experimental Design under Uncertainty for Materials Discovery. Patterns 4(11): 100863 (2023).

[5] Zhao, G., Dougherty, E.R., Yoon, B.J., Alexander, F., Qian, X. Uncertainty-aware Active Learning for Optimal Bayesian Classifier, in the 9th International Conference on Learning Representations (ICLR), May 4-8, 2021.

[6] Hasson, U. The neurobiology of uncertainty: implications for statistical learning. Philosophical Transactions of the Royal Society B: Biological Sciences 372: 1711 (2017).

[7] Drugowitsch, J., Pouget, A. Probabilistic vs. non-probabilistic approaches to the neurobiology of perceptual decision-making. Current Opinion in Neurobiology 22(6): 963-969 (2012).

[8] Rahnev, D. The Bayesian brain: What is it and do humans have it? Behavioral and Brain Sciences 42: e238 (2019).

Bio-inspired Emergent Intelligence for Scientific Computing

Siddharth Mansingh
Los Alamos National Laboratory

# Bio-inspired Emergent Intelligence for Scientific Computing

Backpropagation-based algorithms have shown state-of-the-art performance in the field of deep learning, when training artificial neural networks on traditional GPU-based hardware. However, they are computationally expensive to train. Neuromorphic computers mitigate this issue through event-driven computing, meaning that they compute only when new data becomes available. Neuromorphic systems are inherently parallel, where all the neurons and synapses perform computations simultaneously and asynchronously, as opposed to their von Neumann counterparts, which rely on a centralized processor [1]. However, most neuromorphic applications have been limited to backpropagation-based approaches because of their state-of-the-art performance in deep learning. Backpropagation on neuromorphic hardware often involves copying weights that were trained on a separate machine to the neuromorphic chip [2] or having separate copies of the weights on-chip [3] that affects the highly parallelizable, asynchronous attribute of the neurons. Common neuromorphic architectures are also not inspired by the brain despite their goal being to emulate them.

Top-down feedback and lateral inhibition are two brain-inspired components offering critical advantages but have often been overlooked in standard DNNs, not to mention their neuromorphic implementations. Earlier neuroimaging studies have found that bidirectional visual cortex activity has functional consequences such as directing spatial attention or enhancing illusory-contour coding in lower visual areas [4, 5]. The strength of feedback coupling in the early visual cortex has been found to enhance contextual [6] and perceptional effects [7]. Top-down connections are also known to turn inference into a dynamical process and attend to behaviorally relevant information [8]. Additionally, lateral inhibition/competition is thought to contribute to feature selectivity [9], contrast-invariant tuning [10], and noise filtering in the primary visual cortex[11]. The inclusion of lateral competition in deep-learning architectures has enabled the learning of sparse representations, thereby adding to robustness against attacks [12].

Recent developments at LANL have found several advantages to using an energy-based approach that incorporates top-down feedback into neural networks. These methods implement a dynamical energy-minimization rule where the representations are allowed to settle into attractors over time, based on a learning framework called equilibrium propagation. The algorithms lack a global update, which is otherwise memory intensive. Instead, weights are modified locally in time as well as in space, eliminating the need for separate forward and backward passes and making it more viable for scalable implementation in neuromorphic hardware. Furthermore, we also eliminated gradient computations by comparing pre- and post-synaptic activations, inspired by spike timing dependent plasticity [13], while learning the weights between neurons. While these advances promise energy-efficient computation, there are far-reaching consequences of including biologically inspired components. Energy-based models (EBMs) trained with local learning rules tend to be robust to adversarial attacks and natural corruptions , see Figure 1. This is hypothesized to be caused by the complex energy landscape where features are embedded [14]. Early investigation has also shown that EBMs exhibit state-space hysteresis, where the prediction of the model depends on the priors it was instantiated with, which allows for their applications in spatio-temporal datasets such as video classification and fluid modelling. This opens avenues for scientific computing and discovery, something that was not accessible in previous pursuits of neuromorphic computing.

In terms of practical implementation of energy-based models, there have been related non-spiking analog implementations in the form of decentralized physics-driven flow networks that are resistant to hardware damage [15]. The identification of materials that can realize local update rules is necessary to eliminate the need for a central processor and bring miniaturization and scalability of high-performance neuromorphic processors within reach. Fortunately, one may not have to look far since these features already exist in present day neuromorphic substrate candidates. Lateral artificial synapses have been materialized in several experiments, in the form of, but not limited to, lateral memristors [16] protonic/electronic hybrid transistors [17], metal oxide heterojunctions [18] and transition metal dichalcogenide layered channel tunnel-field-effect transistors [19] .

To conclude, the computational science community should not just adapt state-of-the-art deep learning solutions such as transformer architectures into neuromorphic hardware but attempt to surpass it with more brain-inspired emergent learning.

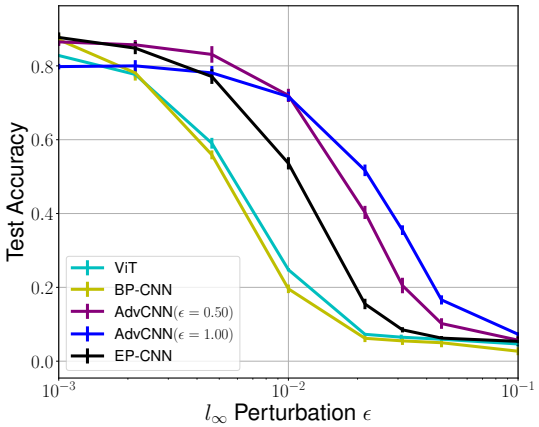



Figure 1: Accuracy under adversarial attacks for models trained with different learning frameworks, plotted against various strengths of attacks. **ViT**: Vision Transformers, **BP-CNN**: Convolutional neural networks trained with backpropagation, **Adv-CNN**: CNNs trained with backpropagation and adversarial training with different attack strengths and **EP-CNN**: CNNs trained with equilibrium propagation, with feedback connections. **EP-CNN** exhibits the best robust performance without any adversarial training, without any drop in clean accuracy. Figure adapted from [14].

## A References


[1] Marković, D., Mizrahi, A., Querlioz, D. & Grollier, J. Physics for neuromorphic computing. *Nature Reviews Physics* **2**, 499–510 (2020).

[2] Esser, S. K., Appuswamy, R., Merolla, P., Arthur, J. V. & Modha, D. S. Backpropagation for energy-efficient neuromorphic computing. In *Advances in Neural Information Processing Systems*, vol. 28 (2015).

[3] Renner, A., Sheldon, F., Zlotnik, A., Tao, L. & Sornborger, A. The backpropagation algorithm implemented on spiking neuromorphic hardware (2021). `arXiv:2106.07030`.

[4] Nielsen, M. *et al.* Reciprocal connectivity in visual cortex: evidence from fmri. In *IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.99CH37028)*, ICSMC-99 (IEEE).

[5] Dijkstra, N., Zeidman, P., Ondobaka, S., van Gerven, M. A. J. & Friston, K. Distinct top-down and bottom-up brain connectivity during visual perception and imagery. *Scientific Reports* **7** (2017).

[6] Czigler, I. & Winkler, I. *Unconscious Memory Representations in Perception: Processes and mechanisms in the brain*, vol. 78 (John Benjamins Publishing, 2010).

[7] Noudoost, B., Chang, M. H., Steinmetz, N. A. & Moore, T. Top-down control of visual attention. *Current Opinion in Neurobiology* **20**, 183–190 (2010).

[8] Desimone, R. & Duncan, J. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience* **18**, 193–222 (1995).

[9] Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996).

[10] Denève, S., Pouget, A. & Latham, P. Divisive normalization, line attractor networks and ideal observers. In *Advances in Neural Information Processing Systems*, vol. 11 (MIT Press, 1998).

[11] Chettih, S. N. & Harvey, C. D. Single-neuron perturbations reveal feature-specific competition in v1. *Nature* **567**, 334–340 (2019).

[12] Teti, M., Kenyon, G., Migliori, B. & Moore, J. LCANets: Lateral competition improves robustness against corruption and attack. In *Proceedings of the 39th International Conference on Machine Learning*, vol. 162 (PMLR, 2022).

[13] Bliss, T. V. P. & Collingridge, G. L. A synaptic model of memory: long-term potentiation in the hippocampus. *Nature* **361**, 31–39 (1993).

[14] Mansingh, S., Kucer, M., Kenyon, G. T., Moore, J. & Teti, M. How robust are energy-based models trained with equilibrium propagation? In *Associative Memory & Hopfield Networks NeurIPS workshop* (2023).

[15] Dillavou, S., Stern, M., Liu, A. J. & Durian, D. J. Demonstration of decentralized physics-driven learning. *Physical Review Applied* **18** (2022).

[16] Huh, W., Lee, D. & Lee, C. Memristors based on 2d materials as an artificial synapse for neuromorphic electronics. *Advanced Materials* **32** (2020).

[17] Zhu, L. Q., Wan, C. J., Guo, L. Q., Shi, Y. & Wan, Q. Artificial synapse network on inorganic proton conductor for neuromorphic systems. *Nature Communications* **5** (2014).

[18] Liu, Q. *et al.* All-in-one metal-oxide heterojunction artificial synapses for visual sensory and neuromorphic computing systems. *Nano Energy* **97**, 107171 (2022).

[19] Pal, A. *et al.* An ultra energy-efficient hardware platform for neuromorphic computing enabled by 2d-tmd tunnel-fets. *Nature Communications* **15** (2024).

Scalable Ultrafast Superconducting Neuromorphic Circuits (SNC)

Andres Marquez, Mukhanov, and Rogene Eichler West
Pacific Northwest National Laboratory

**Scalable Ultrafast Superconducting Neuromorphic Circuits (SNC)**

The main driver for analog neuromorphic circuits is the prospect of orders-of-magnitude energy-efficiency improvements. For years, data centers have remained somewhat stable in their power consumption, with efficiencies growing commensurate with workloads. With the AI revolution, however, the demands on computing capacity to train the largest AI models double every 3-4 months. It is expected that this growth will increase power demand by 160% over the next two years alone, consuming 3-4% of the world's electricity while doubling carbon dioxide emissions. Scalable ultrafast superconducting neuromorphic circuits promise to offer a low powered, alternative building block for both data centers and supercomputers, deployable in a heterogeneous environment alongside traditional CMOS and emerging quantum technologies.

Traditional analog circuits suffer from error accumulation as the length of the circuit increases, requiring high accuracy component biasing and sophisticated differential error compensation. Single flux quantum-based (SFQ) logic addresses many of these issues; it is spike based, providing the benefits of analog charge accumulation (in the form of currents), while also paired with threshold-based activation, thereby avoiding analog error propagation. Within a reasonable fan-out size between stages, this technology enables large scale energy-efficient circuit designs.

SFQ logic exploits Josephson junctions (JJ), which devices that consist of two superconductors separated by a thin insulating barrier. The Josephson effect, which occurs in these junctions, is a quantum mechanical phenomenon where a supercurrent (a current of paired electrons with zero electrical resistance) flows across the junction without any applied voltage. Magnetic flux discretization naturally lends itself to analog spiking signal activity, following the dynamics of a dampened pendulum. SNC can model neurons with polarizing and hyperpolarizing channels, coupled with modeled chemical synapses. Built into a circuit, this logic naturally lends itself to retention and spike rate encoding, with speeds ~$10^2$ GHz and switching energy per gate ~$10^{-19}$ J. Notably, this superconductive approach is an order of magnitude, to tens of gigahertz faster than neuromorphic chips developed commercially, such as IBM's TrueNorth.

Challenges SNC include variations in fabrication processes or environmental conditions that can lead to differences in critical current density, junction capacitance, and other parameters, all of which can affect the overall behavior of an analog circuit by increasing noise. There are a number of a near-term improvements which are being implemented or being considered for high-density, higher reliability fabrication process including Josephson junctions with higher critical current density, Josephson junctions with materials capable of withstanding up to 400 C processing temperatures, integrating multiple junction layers, self-aligned via shunts, high kinetic inductance layers, superconducting plugs, tunable HZO capacitors, etc.

In parallel with improvements in materials and process have been new approaches to real-time pre-decoding error correction for the inevitable analog noise, employing stage encoding to reduce data rate and latency when performing parity checks. One approach towards pre-decoding error correction capabilities have been based on a convolutional neural network, as such a model implements the multistage sliding window process by which a parity check might be performed locally.

Operating circuits at cryogenic temperatures require sophisticated cooling systems and insulation to maintain stable conditions and prevent thermal noise from affecting performance. However, as recently demonstrated by an IARPA funded study, the energy efficiency of these devices, even with the energy cost of a cryogenic system, is still three orders of magnitude greater than traditional CMOS processors. Further, as quantum devices already have cryogenic systems in place, SNC can be deployed next to qubits with energy-efficiency over six orders of magnitude better compared to CMOS processors which dissipate too much heat for co-locating.

Analog, scalable, and capturing the dynamics of chemical communication in biological neurons, we propose the SNC will emerge as a frontrunner in the next generation of neuromorphic computing.

## References


Barnell, Mark, Courtney Raymond, Matthew Wilson, Darrek Isereau, Eric Cote, Dan Brown, and Chris Cicotta. "Demonstrating Advanced Machine Learning and Neuromorphic Computing Using IBM's NS16e." In *Intelligent Computing: Proceedings of the 2020 Computing Conference, Volume 1*, pp. 1-11. Springer International Publishing, 2020.

Chalkiadakis, Dimitrios, and Johanne Hizanidis. "Dynamical properties of neuromorphic Josephson junctions." *Physical Review E* 106, no. 4 (2022): 044206.

Crotty, Patrick, Dan Schult, and Ken Segall. "Josephson junction simulation of neurons." *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* 82, no. 1 (2010): 011914.

Crotty, Patrick, Kenneth Segall, and Daniel Schult. "Biologically realistic behaviors from a superconducting neuron model." *IEEE Transactions on Applied Superconductivity* 33, no. 4 (2023): 1-6.

Delfosse, Nicolas. "Hierarchical decoding to reduce hardware requirements for quantum computing." *arXiv preprint arXiv:2001.11427* (2020).

Herr, Anna, Quentin Herr, Steve Brebels, Min-Soo Kim, Ankit Pokhrel, Blake Hodges, Trent Josephsen et al. "Scaling NbTiN-based ac-powered Josephson digital to 400M devices/cm $^ 2$." *arXiv preprint arXiv:2303.16792* (2023).

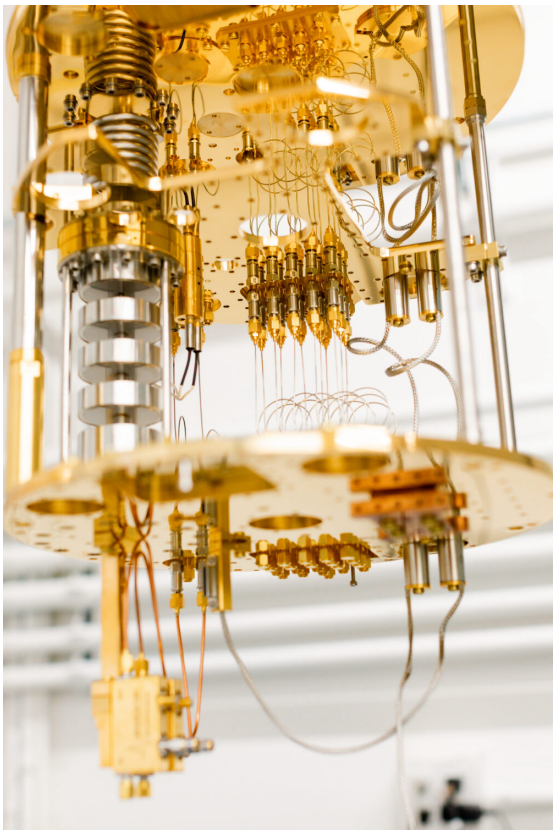
Holmes, D. Scott, Andrew L. Ripple, and Marc A. Manheimer. "Energy-efficient superconducting computing—Power budgets and requirements." *IEEE Transactions on Applied Superconductivity* 23, no. 3 (2013): 1701610-1701610.

Meinerz, Kai, Chae-Yeun Park, and Simon Trebst. "Scalable neural decoder for topological surface codes." *Physical Review Letters* 128, no. 8 (2022): 080505.

Segall, K., C. Purmessur, A. D'Addario, and D. Schult. "A superconducting synapse exhibiting spike-timing dependent plasticity." *Applied Physics Letters* 122, no. 24 (2023).

Tschirhart, Paul, and Ken Segall. "Brainfreeze: Expanding the capabilities of neuromorphic systems using mixed-signal superconducting electronics." *Frontiers in Neuroscience* 15 (2021): 750748.

Ueno, Yosuke, Masaaki Kondo, Masamitsu Tanaka, Yasunari Suzuki, and Yutaka Tabuchi. "QECOOL: On-line quantum error correction with a superconducting decoder for surface code." In *2021 58th ACM/IEEE Design Automation Conference (DAC)*, pp. 451-456. IEEE, 2021.

Cryogenic real-time pre-decoding error correction using a convolutional network with a multistage sliding window process for parity checking

# Optimal Data Encoding Methods for Neuromorphic Computing

Prashansa Mukim[1], Soumyajit Mandal[1], and Piotr Maj[1]
[1]Instrumentation Department, Brookhaven National Laboratory, {pmukim,smandal,pmaj}@bnl.gov

**INTRODUCTION:** Neuromorphic computing, which is inspired by the brain's organizational principles and dynamic capabilities, excels for specific workloads, such as complex spatio-temporal processing and real-time decision-making, that are challenging for traditional von Neumann architectures. However, their performance is strongly dependent on the methods used to encode information into asynchronous neural outputs (known as spikes). Our hypothesis is that neuromorphic paradigms are highly efficient and successful only when both the input data and neural outputs use appropriate encoding methods. For instance, we expect neuromorphic systems to benefit from mimicking the neural encoding schemes observed in brain regions such as the cortex, hippocampus, and thalamus during similar tasks. Thus, we propose a novel approach to neuromorphic computing that emphasizes the crucial role of data encoding as a foundational primitive needed to capture the full functionality of biological computing mechanisms.

To harness the full potential of neuromorphic circuits and systems, we advocate for a focused effort on developing and implementing spike-based encoding schemes that align with the intrinsic strengths of neuromorphic architectures while being resilient to noise and hardware errors. This targeted approach will enable us to identify and solve computational problems where neuromorphic computing can be most effective. Furthermore, we propose the development of encoding primitives tailored for applications in nuclear physics, high-energy physics, and basic energy sciences providing a pathway for practical and efficient implementation of neuromorphic systems with currently available technology. Our work aims to contribute to the ongoing discourse in the neuromorphic computing community by providing insights and practical solutions for translating neuroscience algorithms into functional neuromorphic circuits, ultimately advancing the field towards achieving true computational equivalence with biological systems.

**ENCODING METHODS:** Neural systems use a range of signal encoding methods, each with distinct advantages and limitations [1]. Broadly speaking, the two main approaches are rate coding and temporal coding. Rate codes embed information into the instantaneous or averaged rate at which one or more neurons generate spikes, while temporal codes encode information in their relative timing. Temporal codes can use a multitude of timing features, such as temporal contrast, latency and inter-spike intervals, correlation and synchrony, and timing with respect to global references, with the choice for a particular task determined by evolution. In fact, sensory systems have evolved highly efficient coding strategies to maximize the information conveyed to the brain while minimizing the required energy and neural resources [2–4]. For example, the coding used by auditory nerve fibers approaches an information theoretic optimum for natural sounds and human speech [5]. However, such data- and task-specific optimization of the encoding method has not been carried out for existing neuromorphic systems. Instead, the encoding method is typically determined by hardware constraints set by the neurons or input/output buses.

**Analysis of neuromorphic coding methods**: It is important to theoretically analyze the efficiency of encoding methods used by neuromorphic processors, analogous to the work already done for biological sensory pathways [6, 7], and then develop improved methods that approach the limits set by information theory. A key goal is to study adaptive coding methods, which are common in biology [8].

**Co-design of coding and hardware**: Development of tools for the co-design of coding strategies and neural hardware is critical for improving the performance of neuromorphic systems. Such tools should allow users to optimize the coding methods used for particular data sets under hardware constraints [9], such as speed, power, area, and the available parameter space for the neurons and synapses.

**Applications in scientific computing**: A key application of the proposed co-design tools lies in the design of neuromorphic processors and coding strategies optimized for the unique needs of scientific computing [10, 11], which typically include a combination of low latency, low power, and scalability.

**CONCLUSION:** While the data encoding methods used by biological neurons have been extensively studied, only a small subset of these methods have been applied to neuromorphic computing. Here we highlight the need to address this gap by pursuing a co-design approach in which the encoding methods are optimized for the expected workloads in a hardware-aware manner, i.e., while taking hardware constraints into account. The proposed approach is expected to be a key enabler for the development of practical neuromorphic processors that can meet the rapidly growing needs of scientific computing.

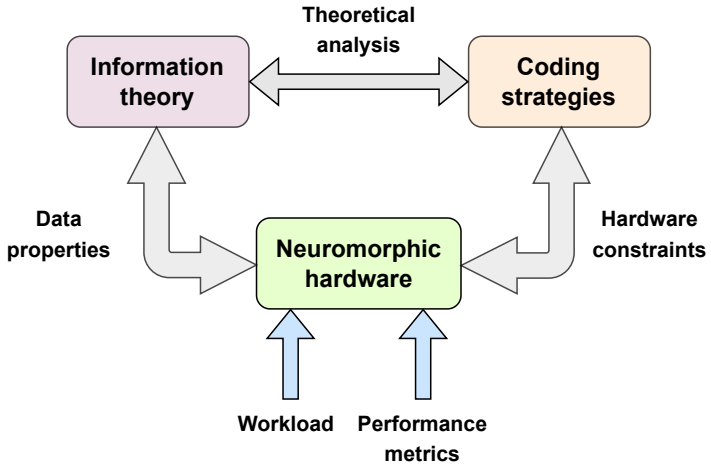



Figure 1: Overview of the proposed hardware-coding co-design approach for optimizing neurormorphic processors for use in emerging scientific applications.

## References


[1] Daniel Auge, Julian Hille, Etienne Mueller, and Alois Knoll. A survey of encoding techniques for signal processing in spiking neural networks. *Neural Processing Letters*, 53(6):4693–4710, 2021.

[2] Joseph J Atick. Could information theory provide an ecological theory of sensory processing? *Network: Computation in neural systems*, 3(2):213–251, 1992.

[3] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

[4] Simon B Laughlin and Terrence J Sejnowski. Communication in neuronal networks. *Science*, 301(5641):1870–1874, 2003.

[5] Evan C Smith and Michael S Lewicki. Efficient auditory coding. *Nature*, 439(7079):978–982, 2006.

[6] Eizaburo Doi and Michael S Lewicki. A simple model of optimal population coding for sensory systems. *PLoS computational biology*, 10(8):e1003761, 2014.

[7] Alexander Borst and Frédéric E Theunissen. Information theory and neural coding. *Nature Neuroscience*, 2(11):947–957, 1999.

[8] Alison I Weber and Adrienne L Fairhall. The role of adaptation in neural coding. *Current opinion in neurobiology*, 58:135–140, 2019.

[9] Rajit Manohar. Hardware/software co-design for neuromorphic systems. In *2022 IEEE Custom Integrated Circuits Conference (CICC)*, pages 01–05. IEEE, 2022.

[10] Catherine D Schuman, Shruti R Kulkarni, Maryam Parsa, J Parker Mitchell, Bill Kay, et al. Opportunities for neuromorphic computing algorithms and applications. *Nature Computational Science*, 2(1):10–19, 2022.

[11] James B Aimone, Ojas Parekh, and William Severa. Neural computing for scientific computing applications: more than just machine learning. In *Proceedings of the Neuromorphic Computing Symposium*, pages 1–6, 2017.

# Neuromorphic ionic computing for next-generation information processing and AI

Aleksandr Noy (LLNL), Seth B. Darling (ANL), Alberto Salleo (Stanford University)

Topic: "*Technologies and prototyping of neuromorphic analog primitives*"

**Introduction and Position**. The human brain is able to process information with astonishing efficiency by (i) combining information processing and storage at the same location and using very low energy single steps; (ii) using a large number of distinct information carriers; and (iii) building a massively 3D-interconnected and parallelized information processing network. Ionic computing,[1-3] which uses ions instead of electrons as information carriers, can mimic the brain computing paradigm. We believe that *ionic computing platforms can utilize multiple types of ions as independent information carriers to take advantage of novel phenomena such as ionic memory, ionic crowding, and ion accumulation and depletion to implement neuromorphic functionality and advanced artificial intelligence algorithms and deliver high computational efficiency and complexity.*

**Discussion.** Our brains are the prototypical ionic computers, which use ions and small molecules to carry information through a vastly parallel and complex neural network built entirely with soft materials and can process information at a mere fraction of the cost of modern supercomputing infrastructure. We argue that ionic computing[4, 5], which encodes information in ionic conductance states realized in nanostructured materials and assemblies that include nanofluidic channels,[1, 2] mixed ion/electron conductors,[6] and other platforms, has the potential to realize many of the advantages of neuromorphic computing. These pioneering examples have already demonstrated basic neuromorphic functionality, such as short- and long-term potentiation, spike-time dependent plasticity, and Hebbian learning, and implemented neuromorphic algorithms such as reservoir computing. Overall, ionic computing offers the following key advantages:

(i) The information can be encoded with multiple ion types, which would not only enable parallel information processing but would also open up an opportunity to creates mixed analog states that would further enhance the system information processing capabilities.

(ii) Even the first proof-of-principle ionic computing devices demonstrate impressive energy efficiency; for example, a nanofluidic synapse showed extremely low energy consumption of 0.66 pJ/spike.[2]

(iii) Ionic circuits are typically based on soft materials that simplify 3D integration and offer an opportunity to create massively interconnected architectures that mimic brain complexity.

To realize this potential, the following *fundamental knowledge gaps* must be closed:

1. Current theoretical models and simulations of iontronic systems are limited in their ability to accurately predict behavior in complex, real-world scenarios. We need to develop robust, multi-scale theoretical frameworks and computational models that can accurately describe the dynamics of ions in nanofluidic environments, including interactions with surfaces, solvents, and other ions.
2. We need to understand the physical origins of long-term and short-term ionic memory effects, their relationship to spatial confinement, chemical gradient, and electric field and how we can control them.
3. While the concept of using multiple ions as information carriers is central to the proposed vision for ionic computing vision, there is little precedent for simultaneous specific detection of multiple ion states and fluxes on the size and complexity required for multi-ion computing. Thus, we need to develop completely new detection schemes, materials, and hardware to monitor multiplexed information flows in nanofluidic networks.
4. A single biological neuron makes on average 7,000 synaptic connections to other neurons over a 3D network that packs trillions of such connections in a relatively small volume. We need to develop new approaches and materials that will support manufacturing or programmable self-assembly that could achieve the required complexity and integrate the required functionality.
5. We need to develop new computational paradigms, algorithms, and approaches for "multi-color computing" that can take advantage of ionic computing capabilities.

**Conclusion**. We believe that the concept of information processing with ions can deliver advanced complexity and energy efficiency that is a hallmark of neuromorphic computing. We also believe that addressing some of the fundamental knowledge gaps that we have outlined will bring us significantly closer to making neuromorphic ionic computing a reality.

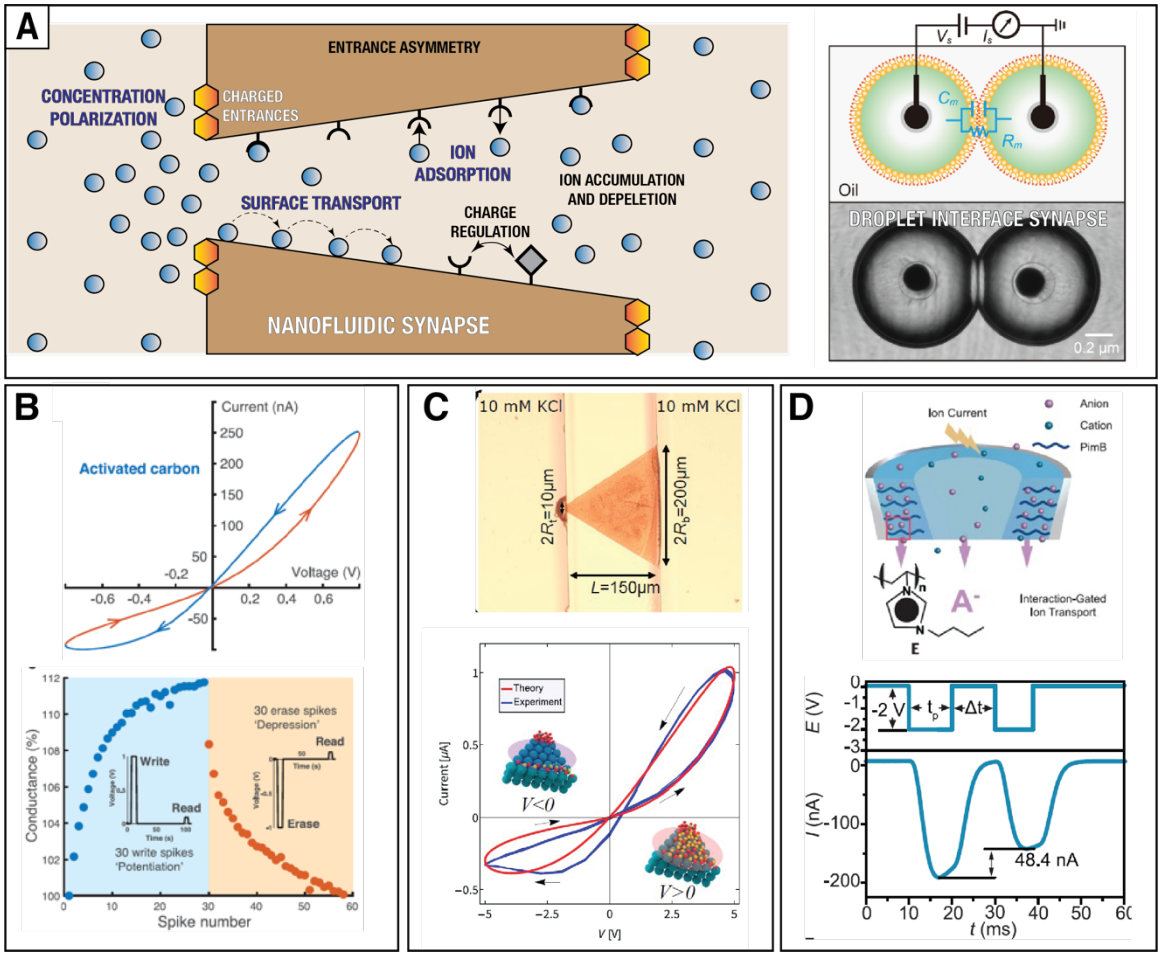



**Figure 1. A.** Fundamental mechanisms enabling neuromorphic behavior in nanofluidic channels and droplet interfaces.[5] **B-D**. Examples of neuromorphic functionality implemented in nanofluidic systems. From: Ref. 1(B); Ref. 3 (C); Ref. 2 (D).

**Author contacts**: *noy1@llnl.gov; darling@anl.gov; asalleo@stanford.edu.*

**Literature**.

(1) Robin, P.; Emmerich, T.; Ismail, A.; Nigues, A.; You, Y.; Nam, G.; Keerthi, A.; Siria, A.; Geim, A. K.; Radha, B.; et al. Long-term memory and synapse-like dynamics in two-dimensional nanofluidic channels. *Science* **2023**, *379*, 161-167.

(2) Xiong, T.; Li, C.; He, X.; Xie, B.; Zong, J.; Jiang, Y.; Ma, W.; Wu, F.; Yu, P.; Mao, L. Neuromorphic Functions with a Polyelectrolyte-Confined Fluidic Memristor. *Science* **2023**, *379*, 156-161.

(3) Kamsma, T. M.; Kim, J.; Kim, K.; Boon, W. Q.; Spitoni, C.; Park, J.; van Roij, R. Brain-inspired computing with fluidic iontronic nanochannels. *Proc. Natl. Acad. Sci. USA* **2024**, *121* (18), e2320242121.

(4) Noy, A.; Darling, S. B. Nanofluidic computing makes a splash. *Science* **2023**, *379* (6628), 143-144.

(5) Noy, A.; Li, Z.; Darling, S. B. Fluid learning: Mimicking brain computing with neuromorphic nanofluidic devices. *Nano Today* **2023**, *53*, 102043.

(6) van de Burgt, Y.; Lubberman, E.; Fuller, E. J.; Keene, S. T.; Faria, G. C.; Agarwal, S.; Marinella, M. J.; Alec Talin, A.; Salleo, A. A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing. *Nature Mater.* **2017**, *16* (4), 414-418.

**Title: Brain Inspired Large Scale Simulation of Spiking Neural Networks**

**Authors:** Robert M. Patton, Oak Ridge National Laboratory

**Introduction:**

The human brain is one of the most intricate and sophisticated organs, with approximately 86 billion neurons forming trillions of synaptic connections that enable memory, logic, sensory processing, and motor coordination. Neurons communicate through electrical and chemical signals, creating a dynamic web of activity underlying thoughts, emotions, and behaviors. The brain's specialized regions handle distinct functions such as memory and logic while its plasticity allows for lifelong adaptation and learning. This complexity makes the brain a center of intelligence and consciousness and a profound enigma for neuroscientists. Spiking neural networks (SNNs) are inspired by the brain's processing of information through electrical spikes, allowing for time-dependent data processing and dynamic neural interactions. By mimicking these natural mechanisms, SNNs offer efficiency and adaptability advantages in artificial intelligence, with promising applications in robotics and sensory processing.

**Challenges:**

Despite their inspiration from the human brain, spiking neural networks (SNNs) do not fully replicate its capabilities, especially in memory, logic, and behavior. SNNs mimic the brain's communication through electrical spikes but lack the complex, multi-layered processes of human memory and logical reasoning. Human memory relies on an elaborate network of neurons, synapses, and biochemical signals for precise information storage and recall. Logical thinking in the brain involves sophisticated neural circuits and dynamic interactions across regions, far surpassing the simple mechanisms of SNNs. Current neuromorphic hardware significantly constrains the scale of SNNs, which require specialized technology still in its infancy and not yet scalable for large applications. As a result, SNNs, while insightful, remain simplified approximations unable to fully capture the advanced functions of the human brain.

**Opportunities**:

Oak Ridge National Laboratory (ORNL) developed the 2023 R&D 100 Award winning SuperNeuro[1] for simulating SNN architectures. SuperNeuro is a python-based neuromorphic simulator capable of simulating diverse workloads of SNNs at different scales of network sizes and timesteps in an accelerated manner and has simulated a 5 million neuron SNN (0.006% of the human brain) on ORNL's Summit machine. This simulator leverages matrix operations, high performance computing (HPC), and agent-based simulation on GPUs for accelerating the underlying neuronal and the network-level spiking dynamics. This simulation framework provides an opportunity to explore SNNs at scales beyond the limits of current neuromorphic hardware as well as simulate new neuron models. Further, there is an opportunity to explore SNN architectures as they relate to functions. For example, structures that support functions of long-term and short-term memory, logic, and behaviors (e.g., fight or flight). With current DOE HPC platforms, such SNN architectures would be modeled at very large scales. For instance, using all of ORNL's Frontier, we could simulate different structures for memory, logic, behavior as well as their interactions and co-evolution. Consequently, a taxonomy of structures and their corresponding functions would be created enabling researchers to leverage macro-level structures to build different networks for applications (i.e., like Lego blocks of known size and function for building larger structures). Insights from this research could be used to drive neuromorphic hardware and application development.

---

[1] P. Date, et. al., 2023. SuperNeuro: A Fast and Scalable Simulator for Neuromorphic Computing. In Proceedings of the 2023 International Conference on Neuromorphic Systems (ICONS '23). Association for Computing Machinery, New York, NY, USA, Article 40, 1–4. https://doi.org/10.1145/3589737.3606000

# A plausible neuromorphic implementation of a novel coding scheme for memory of daily experience

Colin Porter[1], Raphael Heldman[1], Alex Roxin[2], Yingxue Wang[1]

Max Planck Florida Institute for Neuroscience, One Max Planck Way, Jupiter, 33458, FL, USA[1]
Centre de Recerca Matemàtica (CRM), Campus de Bellaterra
Edifici C, 08193 Bellaterra, Barcelona, Spain [2]

Our daily experience of the world is a continuous flow of information that is segmented into discrete and meaningful episodes during memory encoding (Ben-Yakov and Henson 2018). However, how the hippocampus, a brain region essential for memory encoding, mediates the encoding of segmented experience, has remained unclear. Recently, we found a hippocampal code where neurons respond at the boundaries between individual episodes and then exhibit continuously evolving population dynamics to encode the subsequent segmented experience (Heldman et al. 2023). These neurons form a novel coding scheme different from the well-known place cell code, where neurons respond at discrete locations or time points, encoding discrete moments of experience (O'Keefe and Dostrovsky 1971).

In our experiments, mice performed a repetitive task that requires the integration of self-motion information to infer distance traveled or time elapsed in order to receive a reward at the end of each trial (Fig. 1A). To perform the task successfully, mice must segment a continuous session into discrete episodes (trials) and integrate the subsequent experience following the boundary of each episode (Fig. 1B). In the hippocampal CA1 area, the majority of pyramidal neurons exhibit sharp changes in their activity at the episode boundary. Specifically, the neural activity of one subset of neurons rises sharply before decaying over distance or time (Fig. 1C, top). The neural activity of another subset of pyramidal cells falls sharply before ramping up toward the reward (Fig. 1C, bottom). These subsets of neurons thus exhibit distinct neural activity dynamics that together encode discrete episodes of experience, in this instance, tracking distance traveled or time elapsed in reference to their own response at the boundary. These two dynamic patterns form a novel coding scheme that complements the place cell code in encoding experience episodes with different levels of spatiotemporal precision.

Furthermore, optogenetic manipulations revealed specific interneuron subtypes that predominantly modulate either the subset of pyramidal neurons whose neural activity rises or the subset whose neural activity falls at the episode boundaries. Therefore, the pyramidal neurons and distinct interneuron subtypes form parallel circuit motifs to generate different dynamic patterns and support the new coding scheme.

We constructed a computational model of canonical CA1 circuits that can replicate our experimental observations (Fig. 1D). This model consists of a population of pyramidal neurons, each with a somatic and a dendritic compartment. We demonstrated that two inhibitory interneuron subtypes target distinct compartments of pyramidal cells to form two parallel circuit motifs, which may play critical roles in modulating one of the two subsets of pyramidal cells (Fig. 1E, F). These parallel circuit motifs together form a basic computational unit that can generate the observed CA1 neuronal dynamics and explain the experimental results from manipulating interneuron subtypes.

Hence, we suggest two ideas to contribute to future neuromorphic circuit design. First is the concept of complementary coding schemes. In the novel coding scheme we describe, subsets of pyramidal neurons exhibit abrupt changes in activity at episode boundaries, signaling the start of episode encoding. Additionally, the continuous neural dynamics following the boundary response may bind together the experience segment into an episode. By contrast, the place cell code displays specific tuning at discrete moments of experience. Together, these two complementary coding schemes can encode experience episodes with different spatiotemporal precision.

Second, we propose a basic computational unit that can contain two parallel circuit motifs within a single pyramidal neuron formed by its interaction with distinct types of interneurons. By leveraging dendritic compartmentalization and nonlinear summation of inputs, specific interneurons can support the generation of one of the two dynamic patterns in pyramidal neurons based on the excitatory inputs they receive. A population of such computational units enables the coding scheme consisting of two subpopulations with distinct dynamic patterns. These distinct subpopulations may control distinct components of memory encoding, such as binding ongoing experience or encoding reward.

Taken together, we propose that our computational model may be adapted for neuromorphic computing as a series of interconnected circuit motifs (Fig. 1G) and can provide a foundation for designing scalable neuromorphic circuits that emulate the brain's episodic memory encoding mechanisms.

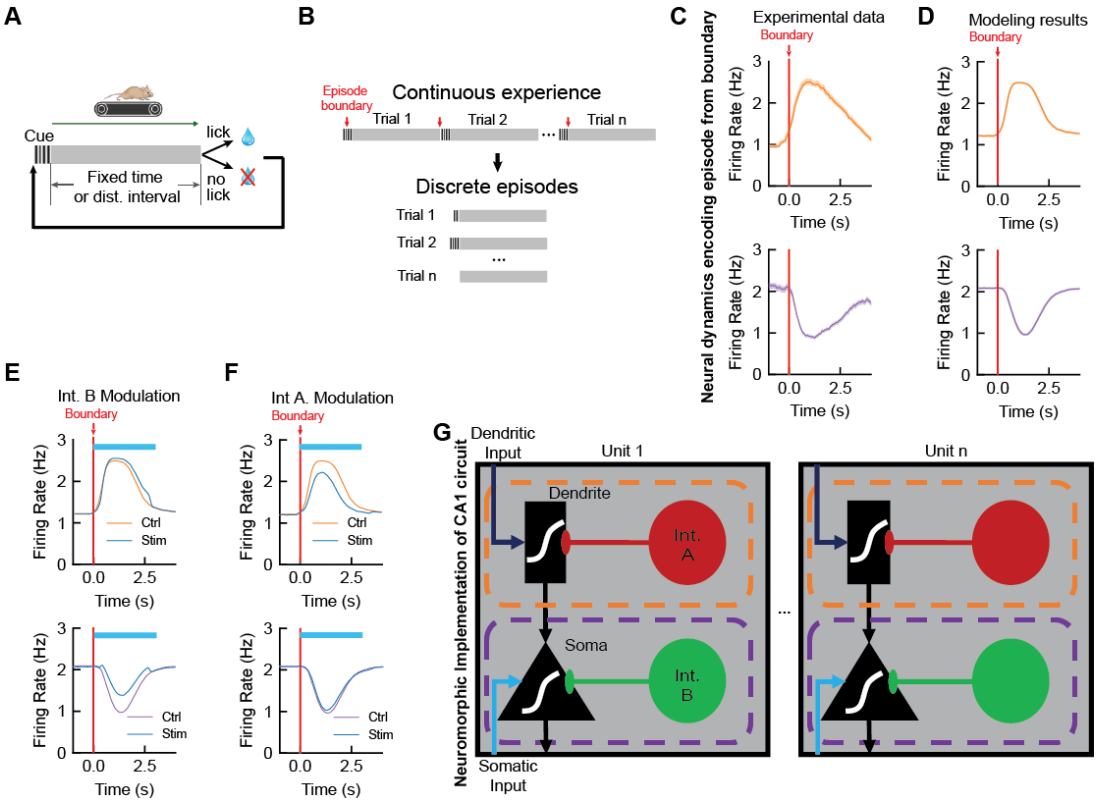



**Figure 1.**

A. Schematic of the behavioral task where head-fixed mice are required to run on a treadmill for a fixed time or distance interval after a start cue, and then lick actively at the end of the trial to trigger a drop of water reward. B. Diagram showing how continuous experience is segmented into discrete episodes (trials). The start of each episode can be different for each trial. C-F. The mean activity of the pyramidal neuron subset that abruptly increases activity at episode boundary (top) and decreases activity at episode boundary (bottom). Aligned to episode boundary. C. Experimental data from extracellular recordings. D. Modelling results of the data in C. E. Modelling results when inactivating interneurons subtype B shown in Figure 1G. F. Modelling results when inactivating interneurons subtype A shown in Figure 1G. G. Schematic of plausible modular implementation of the computational model with neuromorphic circuits. Pyramidal neuron in black with one dendritic and one somatic compartment. Two interneuron subtypes (A and B) that target either the dendritic or somatic compartment of pyramidal neurons and form two parallel circuit motifs (dotted outlines). Color of the outline indicates the type of dynamic pattern produced when the highlighted circuit motif receives strong inputs. Same color scheme as in C-F.

**References**

Ben-Yakov, A. and R. N. Henson (2018). "The Hippocampal Film Editor: Sensitivity and Specificity to Event Boundaries in Continuous Experience." The Journal of Neuroscience **38**(47): 10057-10068.

Heldman, R., et al. (2023). "A CA1 circuit motif that signals the start of information integration." bioRxiv: 2023.2003.2012.532295.

O'Keefe, J. and J. Dostrovsky (1971). "The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat." Brain Research **34**(1): 171-175.

# A multi-mode simulation framework for hardware-software co-development of neuromorphic systems

Prafull Purohit
*Brookhaven National Lab, ppurohit@bnl.gov*

**Topic:** This position paper addresses challenges and opportunities for scalable integration for neuromorphic computing modeling.

**Challenge:**
With inspiration from biology, a neuromorphic system aims to build an electronic computing system that clones the energy efficiency and computational capabilities of the human brain. Simulators are an essential element of the design and development cycle of such neuromorphic systems. They provide a platform to build a mathematical model that describes the dynamic behavior of neural circuits and compute evolution of such circuits through time. Using these simulation frameworks, we can verify our existing understanding in computational neuroscience and evaluate new hypothesis even with novel neuromorphic technologies which are not ready for large scale commercial production. Such tools for simulating neural networks fall into two categories: simulation software and neuromorphic hardware [1]. Software frameworks such as **Brian2** [2], **Nest** [3] and **NEURON** [4] generally target computational neuroscience [5] whereas hardware frameworks such as **TrueNorth** [6], **Loihi** [7], **Braindrop** [8] support millions of neurons using a multi-chip array and offers good power efficiency and performance [9].

While several features are available and often desired in a simulator, some features are especially necessary for replicating the benefits of brain-inspired computing. Such features include scalable connectivity, event-driven computing, high performance, etc. Each of these features of neuromorphic systems is inspired by a particular characteristic or a region of the brain.

**Opportunity:**
Co-design of novel, bio-realistic neuromorphic circuits and neuroscience-based algorithms requires a framework which allows large-scale, high-performance co-simulation of software primitives for computational neuroscience and microelectronic primitives for neuromorphic circuits. In order to capture vital biological computing mechanisms with the right level of abstraction we should have capabilities to simulate large-scale neuromorphic circuit primitives with different level or precision and abstraction. Eventually we should be able to efficiently and precisely simulate a human brain, or parts of it, to understand how changes in the physical connectivity or structure would impact its behavior.

An opportunity exists for the co-design of neuromorphic system and theoretical understanding in computational neuroscience. New computing architecture can be developed which make use of emerging analog primitives for flexible synapse connectivity, temporal delays, etc. Recent evidence suggests that the brain display a complex multi-scale temporal organization where different regions exhibit different timescales [10]. A new co-simulation environment supporting such muti-scale temporal organization would allow us to improve our understanding on how brain processes information. Such an integrated multi-mode simulation framework would allow for focused developed of new neuromorphic computing, communication, and sensing system along with opportunities for rich collaborations between diverse research groups in computational neuroscience, biology, and integrated circuits.

**Assessment:**

Modeling and simulation of large-scale neuromorphic systems are computationally intensive and should utilize High-Performance Computing (HPC) system for highly parallel and distributed implementation. We believe that a holistic simulation framework (see Figure 1), that includes discrete-event simulator for processing neural states and different timescales; logic simulator for digital primitives; and SPICE-like continuous time simulator for emerging analog primitives would provide a platform suitable for understanding the behavior of large-scale neural systems and hardware/software co-design of neuromorphic systems using existing technologies as well as new technologies/approaches where commercial fabrication has not matured enough.

It is our position that the co-design, co-optimization, and co-simulation of neuromorphic hardware and algorithm will enable the development of novel circuit primitives, accelerate discovery of new bio-realistic functionality, and capabilities to simulate large-scale neuromorphic systems that can match the cognitive capabilities of the brain.

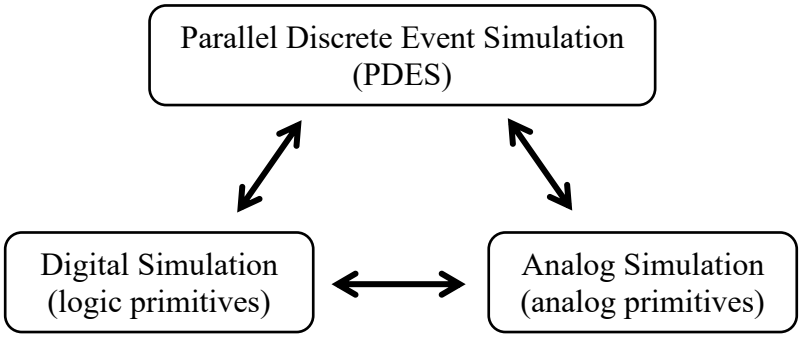



Figure 1. Multi-mode simulation framework for large-scale neuromorphic systems

**References:**

[1] Van Albada, Sacha J., et al. "Performance comparison of the digital neuromorphic hardware SpiNNaker and the neural network simulation software NEST for a full-scale cortical microcircuit model." *Frontiers in neuroscience* 12 (2018): 291.

[2] Stimberg, Marcel, Romain Brette, and Dan FM Goodman. "Brian 2, an intuitive and efficient neural simulator." elife 8 (2019): e47314.

[3] Gewaltig, Marc-Oliver, and Markus Diesmann. "Nest (neural simulation tool)." Scholarpedia 2.4 (2007): 1430.

[4] Carnevale, Nicholas T., and Michael L. Hines. The NEURON book. Cambridge University Press, 2006.

[5] Kulkarni, Shruti R., et al. "Benchmarking the performance of neuromorphic and spiking neural network simulators." *Neurocomputing* 447 (2021): 145-160.

[6] Akopyan, Filipp, et al. "Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip." *IEEE transactions on computer-aided design of integrated circuits and systems* 34.10 (2015): 1537-1557.

[7] Davies, Mike, et al. "Loihi: A neuromorphic manycore processor with on-chip learning." *Ieee Micro* 38.1 (2018): 82-99.

[8] Neckar, Alexander, et al. "Braindrop: A mixed-signal neuromorphic architecture with a dynamical systems-based programming model." *Proceedings of the IEEE* 107.1 (2018): 144-164.

[9] Nageswaran, Jayram Moorkanikara, et al. "Efficient simulation of large-scale spiking neural networks using CUDA graphics processors." 2009 International Joint Conference on Neural Networks. IEEE, 2009.

[10] Golesorkhi, Mehrshad, et al. "The brain and its time: intrinsic neural timescales are key for input processing." *Communications biology* 4.1 (2021): 970.

# Leveraging Circuit Dynamics for Temporal Applications and Event Driven Computing

Qinru Qiu, Department of Electrical Engineering and Computer Science, Syracuse University

The human cognitive process relies on the interaction between long-term and short-term memory. While long-term memory is based on persistent changes in molecular and cellular structures, short-term memory operates through sustained neural circuit activities [1]. The rapid advancement of machine learning has produced very large neural network models for image and language applications, which can be seen as long-term memories containing information engrams. For applications with temporal input, such as decision-making, sensing, and actuation, short-term memory is needed for retaining, integrating past information and extracting temporal features from time series.

Recurrent neural networks (RNN) are typically used for temporal applications. Emerging technologies such as Process in memory (PIM) and compute-in-memory (CIM) can significantly enhance the energy efficiency and throughput of neural networks by executing multiply-and-accumulate (MAC) operations near or within memory. Figure 1 illustrates the inference flow of a simple RNN implemented using memristor CIM. Despite the advantages of CIM and PIM, the system still faces significant challenges: (1) It suffers from dense data and intensive computing activities. (2) Updating the hidden state of an RNN requires costly **global operations**, including analog-to-digital (A/D) and digital-to-analog (D/A) conversions. (3) The conventional approach to training an RNN, which involves unrolling the network along the time axis and performing backpropagation through time (BPTT), is too expensive in terms of both hardware and energy.

Biological *leaky-integrate-and-fire* (*LIF*) neurons are stateful devices. The membrane potential charging process allows them to keep a record of historical inputs, while the leakage process serves as a timer to differentiate temporal patterns in the input. Recent neuroscience research indicates that each dendrite of the neuron exhibits its own dynamic behavior [2] and a single neuron itself is an RNN can function as an RNN, providing complex responses to various temporal patterns. When these neurons form a network, even without recurrency, **states are maintained locally within the network elements without global loops.** As spikes are only generated when membrane potential exceeds the threshold, **neurons have very sparse output**, maintaining low data traffic within the network. Furthermore, **biological neural networks do not undergo unrolling or BPTT for learning**. Instead, they use simple local rules such as three-factor Hebbian learning [3], which leverages a reward function modulated with pre-synaptic (i.e., input) and post-synaptic (i.e., output) spike timing dependent plasticity (STDP).

We argue that circuits with inherent device-level dynamics should be leveraged to implement stateful neurons in short-term memory. Our previous research has demonstrated that even feedforward SNNs exhibit performance comparable or superior to RNNs, showcasing sparse computing activity in sequential input processing and pattern generation [4, 5, 6, 7]. Additionally, we have developed a three-factor Hebbian learning algorithm to train the stateful SNN without using BPTT [8]. We also proposed a CIM-based inference circuit [9] where the LIF neuron is implemented using CMOS circuits. Through the use of hardware-based LIF neurons and with proper training, the global loop in Figure 1 can be reduced to local loops and the dense data is replaced by sparse spiking activities, as illustrated in Figure 2. Although the system is effectively still an RNN, the state update is performed *within the neuron*, hence significantly reducing the amount of register reads, writes, and data movements. Since spikes are binary, no A/D and D/A conversions are needed to interface with the memristor array. The voltage-based CMOS LIF neuron can also be replaced by a 1M1T1R diffusive memristor, which not only performs leaky integration and firing but also spontaneously returns to its resting state afterward [10]. There is an urgent need to study the performance of learning and inference of such stateful SNN at large scale for its potential as a mechanism for short-term memory.

References


[1] J. Sweatt, "Introduction," in *Mechanisms of Memory*, Academic Press, 2010.

[2] C. Koch and I. Segev, "The role of single neurons in information processing," *Nature neuroscience,* vol. 3, no. 11, 2000.

[3] W. Gerstner, M. Lehmann, V. Liakoni, D. Corneil and J. Brea, "Eligibility traces and plasticity on behavioral time scales: experimental support of neohebbian three-factor learning rules," *Frontiers in neural circuits,* vol. 12, no. 53, 2018.

[4] H. Fang, A. Shrestha, Z. Zhao, Y. Li and Q. Qiu, "An Event-Driven Neuromorphic System with Biological Plausible Temporal Dynamics," in *proceeding of International Conference on Computer-Aided Design (ICCAD)*, 2019.

[5] H. Fang, A. Shrestha and Q. Qiu, "Multivariate time series classification using spiking neural networks," in *International Joint Conference on Neural Networks (IJCNN)*, 2020.

[6] H. Fang, A. Shrestha, Z. Zhao and Q. Qiu, "Exploiting Neuron and Syn-apse Filter Dynamics in Spatial Temporal Learning of Deep Spiking Neural Network," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[7] H. Fang, Z. Mei, A. Shrestha, Z. Zhao, Y. Li and Q. Qiu, "Encoding, Model, and Archi-tecture: Systematic Optimization for Spiking Neural Network in FPGAs," in *IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, 2020.

[8] Z. Zhang, J. Jing and Q. Qiu, "SOLSA: Neuromorphic Spatiotemporal Online Learning for Synaptic Adaptation," in *Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2024.

[9] H. Fang, B. Taylor, Z. Mei, L. Ziru, H. Li and Q. Qiu, "Neuromorphic Algorithm-hardware Codesign for Temporal Pattern Learning," in *Design Automation Conference (DAC)*, 2021.

[10] Z. Wang, S. Joshi, S. Savel'Ev, W. Song, R. Midya, Y. Li and M. R. e. al, "Fully memristive neural networks for pattern classification with unsupervised learning," *Nature Electronics,* vol. 1, no. 2, 2018.

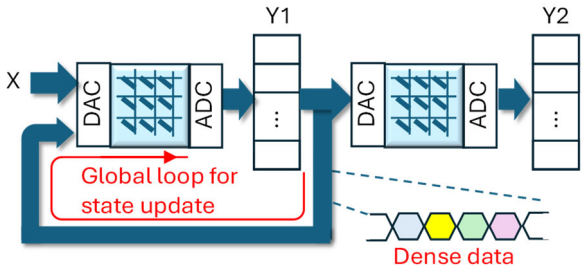




**Figure 1 RNN using conventional CIM.**

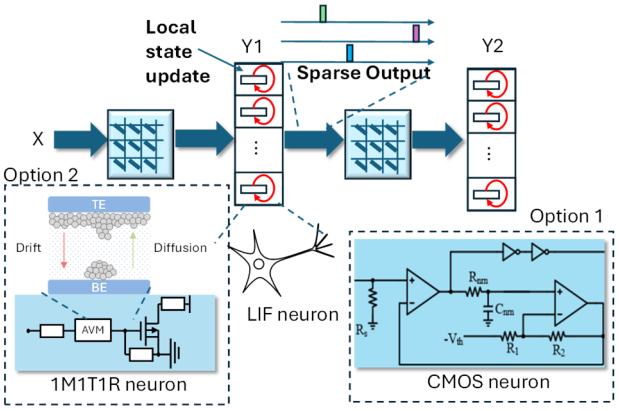



**Figure 2 Stateful SNN using LIF neuron.**

# Neuromorphic Learning with Over-Parameterized Generalized Feedback Networks

Brian Robinson
Johns Hopkins University

**Neuromorphic Learning with Over-Parameterized Generalized Feedback Networks**

The ability of online learning in biological systems to continuously update synapses is fundamental to the functionality and robustness of biological intelligence. Performant artificial neural networks are trained via backpropagation of error to calculate credit assignment for each parameter based on first-order gradients, which is not biologically plausible and not readily implementable in neuromorphic hardware. In particular, "weight transport," utilized in backpropagation is not amenable for neuromorphic implementation because it assumes two sets of identical and co-evolving weights between neurons in a forward pass for inference and a feedback pass for error-gradient-based credit assignment. In neuromorphic hardware that supports online learning [Davies et al. 2018, Denim et al. 2021, Pehle et al. 2022], the most common type of learning implemented are parsimonious rules, such as variants of spike-timing-dependent plasticity (STDP) [Cramer et al. 2020, Denim et al. 2021], which have not been demonstrated to scale across multiple neuron populations or layers, or to enable high performance on modern machine learning benchmarks. For calculating weight updates, a parsimonious rule is attractive from a modeling and neuroscientific perspective, however in biology, synaptic modification relies on complex biochemical processes and specialized network components to impose credit assignment for synaptic modification (including calcium dynamics, active and passive membrane potential dynamics, metaplasticity, and protein synthesis) [Lillicrap et al. 2020, Bliss 2022], which may not be able to be distilled to a parsimonious rule while maintaining robust online learning. We believe that neuromorphic computing systems should include flexible computation of online learning which goes beyond simplified learning rules or even direct approximation of first-order error gradient approximation. Instead, online learning should be over-parameterized to approximate the richness of biological synaptic modification.

***Technical Approach.*** We propose to investigate generalized feedback networks, which extends the feedback network structure pass utilized in backpropagation to encompass a broader set of functions to calculate network parameter updates while preserving the recursive structure of backpropagation. Generalized feedback networks extend parameter update calculation as utilized in backpropagation by still having the same inputs (pre- and post-synaptic activity and a collection of error signals) and outputs (estimated weight change and feedback error signals). However, instead of a linear input-output transformation with weights shared with the feedforward pass (as in backpropagation), generalized feedback networks can encompass a broader set of functions, including artificial neural network components such as a multi-layer perceptron (MLP) and can update a vector of state variables for each synapse. These generalized feedback networks can be directly optimized. Optimization approaches include utilizing an exemplar optimal trajectory of weight changes and utilizing an error-based objective function to directly optimize the feedback network. Thus, generalized feedback networks represent a flexible approach to capture the richness of biological processing mechanism for synaptic modification while maintaining a practical engineering approach targeted to enabling a performant system. Furthermore, generalized feedback networks can be optimized to approximate different forms of idealized network parameter updates, including conditioning on unseen data, higher-order error-gradients [Anil et al. 2020], continual learning regularization approaches [Kirkpatrick et al. 2017], and weight updates observed in biological systems.

***Potential for Impact***

Ultimately, biological neural networks have advantages over artificial neural networks in training efficiency while utilizing specialized network structures to impose credit assignment for synaptic modification. The proposed direct optimization of generalized feedback networks represents a promising path forward to enable scalable online learning in neuromorphic systems with enhanced performance while capturing neuroscience-based computing principles.

***References***


Anil, R., Gupta, V., Koren, T., Regan, K. & Singer, Y. Scalable Second Order Optimization for Deep Learning. *arXiv* (2020) doi:10.48550/arxiv.2002.09018.

Bliss, T. Neuroscience in the 21st Century, From Basic to Clinical. 3053–3075 (2022) doi:10.1007/978-3-030-88832-9_143.

Cramer, B. *et al.* Control of criticality and computation in spiking neuromorphic networks with plasticity. *Nat. Commun.* **11**, 2853 (2020).

Davies, M. *et al.* Loihi: A Neuromorphic Manycore Processor with On-Chip Learning. *IEEE Micro* **38**, 82–99 (2018).

Demin, V. A. *et al.* Necessary conditions for STDP-based pattern recognition learning in a memristive spiking neural network. *Neural Netw.* **134**, 64–75 (2021).

Kirkpatrick, J. et al. Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. 114, 3521–3526 (2017).

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J. & Hinton, G. Backpropagation and the brain. Nat. Rev. Neurosci. 21, 335–346 (2020).

Pehle, C. *et al.* The BrainScaleS-2 Accelerated Neuromorphic System With Hybrid Plasticity. *Front. Neurosci.* **16**, 795876 (2022).

# Recurrent quasi-Boolean circuits for neuromorphic primitives and micro-brain modeling

Matt Roos
Johns Hopkins University

**Recurrent quasi-Boolean circuits for neuromorphic primitives and micro-brain modeling**

Development of scalable neuromorphic computing hardware and algorithms is hampered by a lack of knowledge of the critical neural primitives of processing and learning, and by the long and costly cycle of novel hardware design, development, and fabrication at scales needed to realize practical use cases. We argue that "quasi-Boolean circuits" (QBCs), defined as those composed of unclocked and recurrently connected logic gates, can serve as a low-cost, presently available substrate for developing and testing neuromorphic computing hardware. QBCs can exhibit analog and spike-like transients at nanosecond time scale (Fig 1B), and require very low energy consumption when designed to have transient sparsity. Commercial FPGAs contain ~$10^6$-$10^7$ logic elements, or look-up tables (LUTs), and support reconfigurable implementation of QBCs. These circuits are compact, operate at high-speed, and are energy efficient compared to digital artificial neural network accelerators.

**Technical approach:** Due to the reconfigurability of circuit topology and LUT content, FPGAs could serve as an effective platform not only for the study and development of QBC-based neuromorphic hardware, but also as a deployable platform directly, with no additional hardware development necessary. QBCs used to date[1,2,3,4,5] have neuromimetic analogs including individual neurons (LUTs or modules of LUTs), spiking (transitions between logic states), neural transmission delay (series of inverter gates), and dynamics with a tight border between order and chaos[3]. To bring additional biological analogs to QBCs and study their impact on neurocomputation and learning, researchers could explore novel QBC subcircuits that are designed, synthetically evolved, or gleaned from biological connectomic data. Diversity of neuron types could be simulated using a set of distinct LUTs or LUT modules. Greater spiking realism (punctate low-high-low transitions and refractory periods) could be designed in canonical subcircuits. And while non-biological learning in the form of synthetic evolution[5,6] or gradient-based learning[7] could be studied, the recent commercial availability of online reconfigurable LUTs, updateable at microsecond time scales, provides a mechanism by which plasticity and online learning could be modeled in QBCs. In addition, models of entire micro-brains could be conceivably implemented in QBCs, up to the size of a fruit fly on current or near-future FPGAs, for study of the topology and computational mechanisms that give rise to observed biological behaviors. It is worth nothing that QBCs can support not only development of biofidelic neuromorphic models, but more general "neuro-inspired models" that can be applied to traditional machine learning task with greater energy efficiency, speed, and perhaps capability, relative to modern neural networks implemented on CPUs, GPUs, or even FPGAs under conventional use case (clocked and feedforward). For this latter application, existing ML benchmarks, metrics, and datasets are suitable, when also including measures of hardware platform size, weight, and power requirements. The addition of a suite of temporal-based tasks with spike-based inputs and outputs would better assess the advantages of more neuromimetic models. Example include the Spiking Heidelberg audio[8] and DVS gesture[9] datasets.

**Impact:** Given the market availability of low-cost FPGAs with ~$10^6$-$10^7$ logic units, we believe this approach could allow a large number of academic and industry research labs to do broad and deep experimentation into neuromorphic computation and learning, rapidly advancing the field and ultimately greatly reducing the cost and power requirements of AI/ML while allowing for extreme scales of computation. Implementations with greater brain fidelity may also allow for faster learning, online learning, and better generalization. Finally, while some outcomes may motivate fabrication of novel neuromorphic hardware that more efficiently implements neural mechanisms simulated in QBCs, it is likely that existing FPGA hardware will be suitable for a subset of large-scale QBCs for challenging and important problems, allowing rapid prototyping and deployment of neuromorphic solutions for this range of applications.

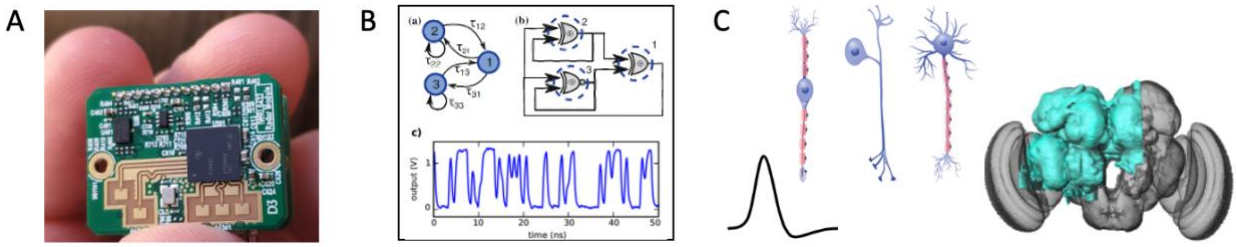



Figure 1: (A) FPGA development boards with >$10^6$ logic elements cost only a few hundred to a few thousand US dollars and are widely available. (B) Quasi-Boolean Circuits (QBCs) made up of unclocked, recurrent logic elements give rise to fast dynamics and spike-like analog signals. (C) QBCs modules could be designed to mimic neural primitives such a transient spike with refractory periods, diverse neuron types, and even real-time plasticity. Larger FPGAs could host QBC network models of entire micro-brains, built from the modules. Such models would be computationally fast and energy efficient, and enable broad study by academic and industry researchers.

**References**


1. Rosin, David P., et al. "Experiments on autonomous Boolean networks." *Chaos: An Interdisciplinary Journal of Nonlinear Science* 23.2 (2013).
2. Komkov, Heidi, et al. "The recurrent processing unit: Hardware for high speed machine learning." *arXiv preprint arXiv:1912.07363* (2019).
3. Apostel, Stefan, et al. "Reservoir computing using autonomous Boolean networks realized on field-programmable gate arrays." *Reservoir Computing: Theory, Physical Implementations, and Applications* (2021): 239-271.
4. Komkov, Heidi, et al. "RF signal classification using Boolean reservoir computing on an FPGA." *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021.
5. Norman-Tenazas, R., Kleinberg, D., Johnson, E. C., Lathrop, D. P., & Roos, M. J. (2023, October). Using evolutionary computation to optimize task performance of unclocked, recurrent Boolean circuits in FPGAs. In *2023 57th Asilomar Conference on Signals, Systems, and Computers* (pp. 1564-1568). IEEE.
6. Schuman, Catherine D., et al. "Evolutionary optimization for neuromorphic systems." *Proceedings of the 2020 Annual Neuro-Inspired Computational Elements Workshop*. 2020.
7. Neftci, Emre O., Hesham Mostafa, and Friedemann Zenke. "Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks." *IEEE Signal Processing Magazine* 36.6 (2019): 51-63.
8. James, Conrad D., et al. "A historical survey of algorithms and hardware architectures for neural-inspired and neuromorphic computing applications." *Biologically Inspired Cognitive Architectures* 19 (2017): 49-64.
9. Amir, Arnon, et al. "A low power, fully event-based gesture recognition system." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

# What is the basic element of neural computing?

Fred Rothganger, Sandia National Labs

In digital electronics, there is the notion of a *universal gate*. With that one kind of gate and sufficient wiring, you can make any other basic gate, and thus build any logic system. Examples include NAND and NOR. Is there a universal gate for biological computing?

For neural computing, the default answer is the leaky-integrate-and-fire (LIF) model, combined with event messages (spikes). Most existing neuromorphic platforms support this basic model in some form, and much of the algorithm research assumes it. However, LIF has its shortcomings. It must be augmented with a plausible learning mechanism such as spike-timing-dependent plasticity (STDP). Even then, it's not clear the LIF is suitable to "capture the *full* functionality of … biological computing".

To find a better answer, we need to pull back and ask: *What is the purpose of biological computing?* This brings us to the field of Cybernetics, that is, communication and control [Norbert Wiener, 1948]. Despite the overwhelming complexity seen in cellular-molecular systems, and particularly neural systems, we consistently observe them controlling other parts of the system, or conveying information for the purposes of control. Of course, these are not the only functions of a biological system, but they are frequently present even when the mechanism has some other purpose.

This suggests a higher-level abstraction for describing biological computation: the *feedback controller*. This device takes an input signal from its surrounding and outputs a control signal. The content of this black box can take several mathematical forms. It generally has some internal memory as state variables, some parameters, and perhaps a direct feedback pathway independent of the loop through the environment. A feedback controller tries to keep its input within some range defined by its parameters.

Notice two things. First, a basic feedback controller is mathematically no more complex than a LIF model. A basic circuit could be built, in digital or analog, that would have the general flexibility of an FPGA gate. Second, spikes are not necessary to the model, but it could generate them if needed by some other part of the system. The key advantage of event messages is that they would eliminate the need to continuously convey the output value. In a sense, spiking would make this a "bang-bang" controller.

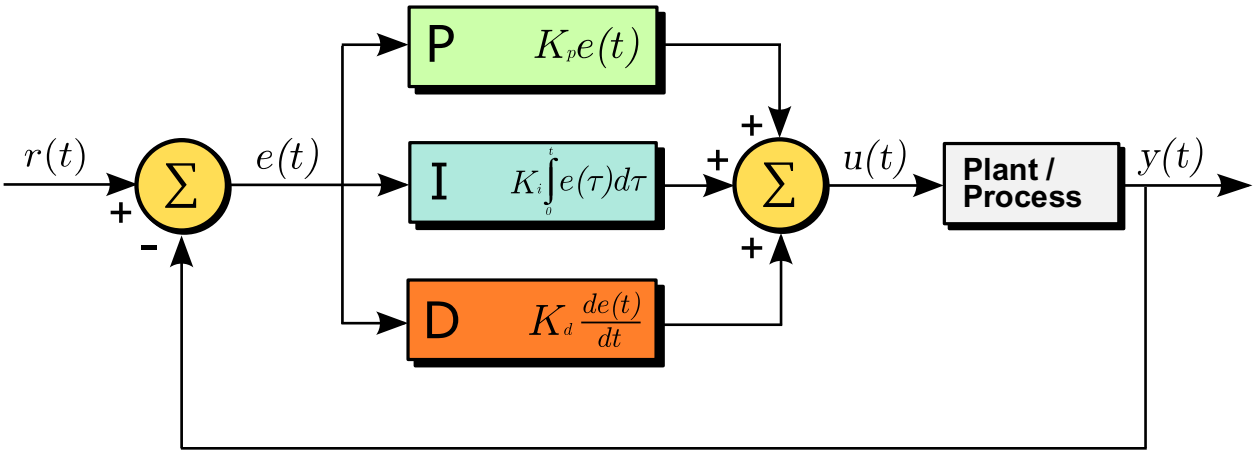
Suppose the output of one controller is coupled to a parameter in another controller. This creates a second-order system capable of adaptation and learning [W. Ross Ashby, 1952]. The input to the second-order controller is the reinforcement signal, while the first-order controller implements the actual behavior of the system. This kind of arrangement can be extended into a complex web or hierarchy.

The significance of such an approach is that, on the one hand, it has a concrete and practical implementation, while on the other hand, it allows the programmer to express higher-level concepts about the biological mechanism in view. By extension, it can express the concepts and goals of artificial technological systems. As such, it is an appropriate abstraction.

A feedback controller can encompass the behavior of multiple neurons, perhaps entire populations. It is also possible to build a detailed model of one biological neuron out of feedback controllers. As such, this device is scale free, truly a universal "gate" for biological computing.

**Disclosure:** *SNL is managed and operated by NTESS under DOE NNSA contract DE-NA0003525.*

**Figures**




[https://en.wikipedia.org/wiki/Closed-loop_controller]

The Proportional-Integral-Derivative (PID) controller is popular because of its simplicity and general usefulness. Subtypes can be created by setting some of the parameters *K* to zero. The full model requires only two state variables and a handful of operations. Of course, more sophisticated controllers also exist.

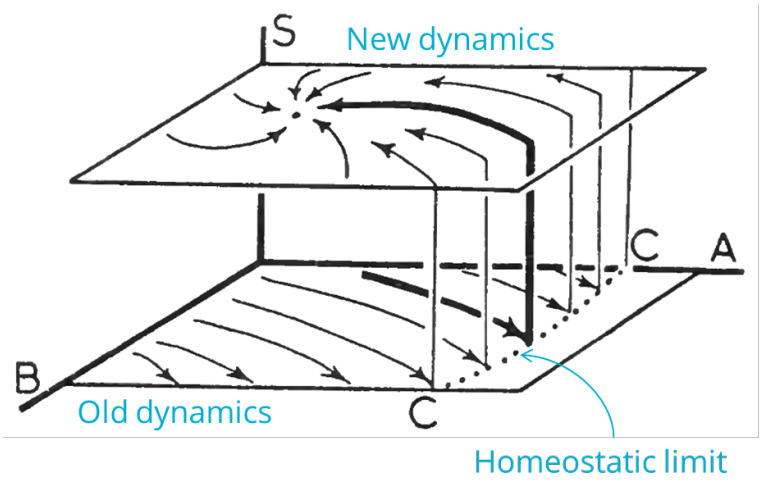



Diagram from [Ashby, 1952], with added notations in blue. The lower plane is a system that is drifting out of equilibrium. Once it hits the limit, a signal is triggered that starts “learning”. It then updates the parameters to new values that produce the dynamics on the upper plane. These are stable, since the attractor is inside the homeostatic limit.

**References**


W. Ross Ashby. *Design for a Brain: The Origin of Adaptive Behavior.* Chapman & Hall, Ltd. 1952.

Norbert Wiener. *Cybernetics: Or Control and Communication in the Animal and the Machine.* MIT Press, 1948.

| | |
|---|---|
| TITLE | ELECTRONIC PHOTONIC INTEGRATED CIRCUITS (EPICS) FOR NEUROMORPHIC COMPUTING |
| TECHNICAL AREA | Technologies and prototyping of neuromorphic analog primitives |
| AUTHOR | Vishal Saxena, *Professor, ECE Department, University of Delaware*<br>Email : vsaxena@udel.edu |

Photonic Integrated Circuits (**PICs**) and their co-integration with CMOS electronics have successfully replaced electrical links in the data centers that support artificial intelligence (**AI**) computation. PICs have recently demonstrated massive computing ability for analog neural networks (**NN**) by exploiting the very high bandwidth of photonic circuits. However, investigations of electronic photonic integrated circuits (EPICs) have lagged in neuromorphic computing. This is especially important as we look at rack-scale comprised of over a thousand neuromorphic processors recently deployed by DOE [1]. This position paper presents an overview of the challenges and opportunities for EPICs for neuromorphic computing.

**Optical Interconnects for Neuromorphic Computing:** Neuromorphic computing encompasses analog NN computing using subthreshold analog or digital integrated circuits, mainstream, and emerging memory arrays, and spiking neural networks (**SNNs**) [2]. A key expectation is that energy efficiency is gained from leveraging the massively parallel, sparse, and event-driven brain-inspired primitives [3]. Neuromorphic systems employ Address Event Representation (AER) or similar interconnects to support asynchronous communication between neuromorphic chips [2]. As the size of these systems scales beyond thousands of chips, the throughput of electrical interconnects will need a rethink. EPICs can play a pivotal role here by allowing extremely high throughput with low-loss communication over lightweight optical fibers.

EPICs feature compact microring modulators that leverage dense wavelength and mode division multiplexing (**WDM**/MDM) to scale data rates to Peta bits/s (***Fig. 1***) [4]. However, adapting these to Neuromorphic computing will require innovation. Spike-based or asynchronous communication requires these optical interconnects to support bursty and short pattern lengths. This requires a rethink of conventional link design and design of event-based equivalent of serializer and deserializer (SerDes) circuits at extremely low power. These will require new mixed-signal circuit primitives and their close co-integration with silicon photonic devices and compact chip-scale lasers.

**EPICs for Analog Neuromorphic Computing:** Recent demonstration of analog NN computing using PICs has spurred great interest in optical computing [5]. These architectures are based on optical interferometric arrays [6], crossbar arrays using WDM (***Fig. 2***), or diffractive optics. Thermo-optic devices commonly realize the optical-domain weights but incur thermal crosstalk between these components, which limits the scaling of these architectures (***Fig. 3***) [7]. Thermal crosstalk mitigation is the key challenge for realizing large-scale EPICs and is a topic of active research in novel materials, device fabrication, and circuit interfaces. The recent integration of low-loss phase change materials (**PCM**) has shown promise for multi-bit nonvolatile state retention but requires improvements in endurance [8].

Another major challenge, similar to analog NN computing arrays, is the size and energy of data converters (DACs and ADCs). These are particularly challenging for the Gbps rate at which the optical interfaces operate to amortize the energy cost of the laser and drivers [5]. Consequently, new optical architectures that circumvent the need for analog to digital conversion and vice versa and maintain NN activations in the analog domain can be transformative. Spiking optoelectronic neurons can alleviate this issue by precluding data converters and require robust circuits that can operate at the desired >100 TOPS equivalent throughout and near fJ-per synaptic event (fJ/SynOp) energy efficiency.

**Other Application Areas:** Optical neuromorphic primitives also apply to DOE's high-energy Physics (HEP) research, where a large multitude of detector readout circuits in a particle accelerator can be closely interfaced with fiber optics to improve clutter and performance in high-radiation environments. Asynchronous optical interconnects can play an important role in this space [9].

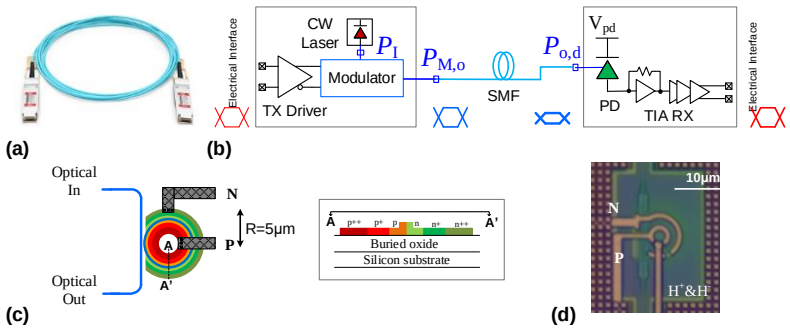



**Figure 1 Optical interconnects using silicon photonics and CMOS electronic ICs [4].**

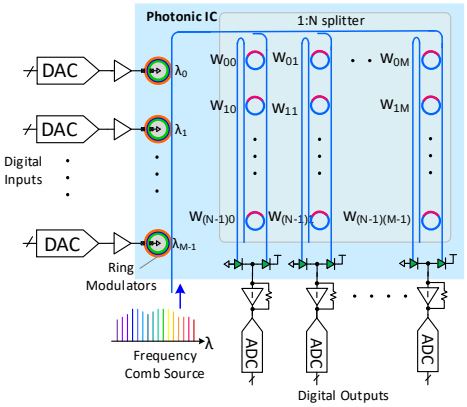



**Figure 2. Optical NNs using WDM and microring resonators.**

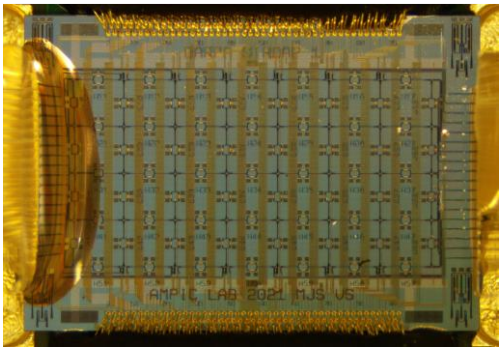


**Figure 3. A large-scale PIC for neural network computations [7].**

## References:


[1] "Taking Neuromorphic Computing to the Next Level with Loihi 2." [Online]. Available: https://download.intel.com/newsroom/2021/new-technologies/neuromorphic-computing-loihi-2-brief.pdf

[2] S.-C. Liu, T. Delbruck, G. Indiveri, A. Whatley, and R. Douglas, *Event-based neuromorphic systems*. John Wiley & Sons, 2014.

[3] V. Saxena, "Neuromorphic computing: From devices to integrated circuits," *Journal of Vacuum Science & Technology B*, vol. 39, no. 1, 2021.

[4] V. Saxena, A. Kumar, S. Mishra, S. Palermo, and K. R. Lakshmikumar, "Optical Interconnects Using Hybrid Integration of CMOS and Silicon-Photonic ICs," *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2023.

[5] C. Huang *et al.*, "Prospects and applications of photonic neural networks," *Advances in Physics: X*, vol. 7, no. 1, p. 1981155, 2022.

[6] Y. Shen *et al.*, "Deep learning with coherent nanophotonic circuits," *Nature photonics*, vol. 11, no. 7, pp. 441-446, 2017.

[7] M. J. Shawon and V. Saxena, "A 7× 4 Silicon Photonic Reconfigurable Optical Analog Processor with Algorithmic Calibration," in *Optical Fiber Communication Conference*, 2024: Optica Publishing Group, p. W2A. 10.

[8] Z. Fang, R. Chen, B. Tossoun, S. Cheung, D. Liang, and A. Majumdar, "Non-volatile materials for programmable photonics," *APL Materials*, vol. 11, no. 10, 2023.

[9] A. Kraxner *et al.*, "Radiation tolerance enhancement of silicon photonics for HEP applications," *Proc. Sci*, vol. 17, 2019.

# Neuromorphic Supercomputing with Superconducting Optoelectronic Networks

Jeffrey Shainline
National Institute of Standards and Technology

To enable neuromorphic supercomputing systems with billions of neurons and trillions of synapses capable of general intelligence and cognition, the guiding theme from neuroscience is to construct hardware capable of efficient movement of information across space and time. This capability is enabled by device, circuit, and network attributes, which all must be considered from first principles when designing neuromorphic hardware. Regarding network architecture, cognition derives from the ability of many interconnected processing modules to rapidly share the results of their computations across large networks. To accomplish this extraordinary feat of communication, each node must make many connections to other modules to maintain a short average path length across the network so that each module can communicate to any other with a small number of intermediate nodes. The connections must be fully dedicated (an independent communication channel for every connection), so latency is minimal and independent of network traffic. The physics of electrical wires makes these traits impossible to achieve with copper interconnects, so nearly all neuromorphic systems use address-event representation (AER), resulting in latency that depends on network activity. With AER, latency grows with network size. Systems of trillions of synapses are orders of magnitude slower than biological brains. However, if light is used for communication, direct connections can be realized in trillion-synapse systems, because light does not experience wiring parasitics. The low-energy limit of optical communication is one photon per synapse event, achievable with single-photon detectors at each synapse (Fig. 1). Superconducting detectors are ideal for this purpose. They must be cooled to 4K, but the energy saved from communicating with single photons more than offsets the energy required for cooling. Operating at 4K also brings the tremendous benefit of superconducting Josephson junctions, which naturally implement analog computational primitives that render neural systems powerful for efficient processing and movement of information. These operations include thresholding and saturating nonlinearities; coincidence and sequence detection (Fig. 2(a)); inhibition (Fig. 2(b)); and analog multiplication and addition. Superconducting circuits also enable a wide range of time constants, from a few picoseconds to indefinite signal retention through persistent supercurrents (as used in the memory cells of Fig. 3). Movement of information across space and time includes information storage, so myriad memory and plasticity mechanisms are required on time scales from the inter-spike interval to the operational lifetime of the system. Superconducting circuits with a few Josephson junctions can implement short-term plasticity, spike-timing-dependent plasticity, homeostatic mechanisms, and three-factor learning rules for metaplasticity. With superconducting optoelectronic circuits, many relevant neural cell types can be realized, including thalamocortical relay cells, pyramidal neurons with elaborate dendritic morphology, place cells, time cells, grid cells, interneurons, and numerous neuromodulatory cell types. We demonstrated many of these core operations in hardware, including single-photon synapses with weighting, a wide range of post-synaptic time constants, local synaptic-weight memory cells, and dendritic coincidence/sequence detection. We demonstrated multiplanar optical waveguides for axonal routing, single-photon communication links, and superconductor-semiconductor interfaces. Analysis of the scaling of this hardware to a supercomputer with 10 trillion synapses reveals that such a system would occupy a volume two meters on a side and would consume 10 kW of power if efficient light sources can be achieved. Such a system would incorporate the device, circuit, and network principles of the brain while operating a million times faster. We recently developed a compact, phenomenological modeling framework that enables us to efficiently simulate large systems of these circuits on GPUs. After demonstrating synapses, dendrites, transmitters, and waveguide interconnection networks, our immediate next step is to construct systems of 100 neurons implementing energy models to solve graph partitioning problems as well as a similar-scale system for video processing with a small-scale version of a mammalian visual and auditory cortex.

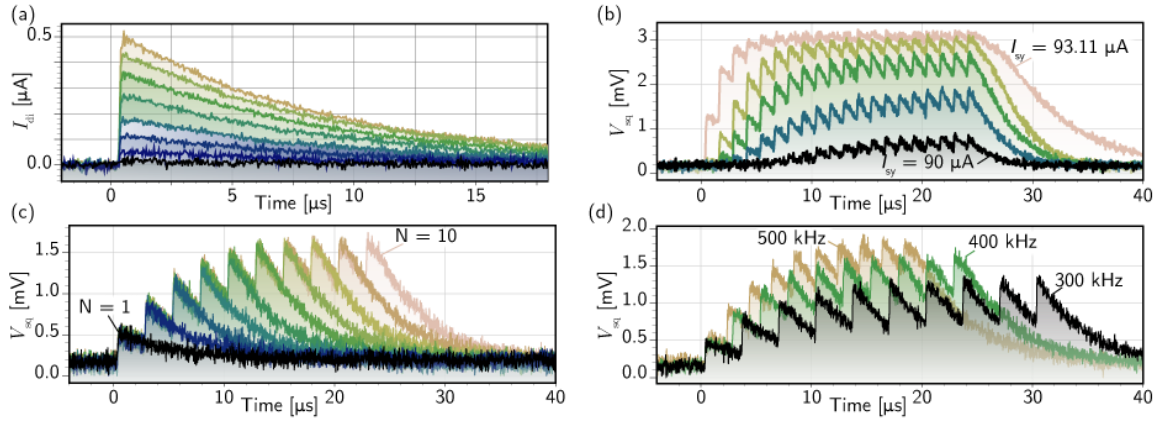



*Figure 1: Measured data from superconducting optoelectronic single-photon synapses. (a) Response to one synapse event. The various curves show different synaptic weights. (b) Response to a train of 20 input synapse events showing the ability of the synapse to integrate the post-synaptic signal. (c) Burst coding, where the integrated signal depends on the number of pulses in a train. (d) Frequency coding, where the integrated signal depends on the frequency at which synapse events arrive.*

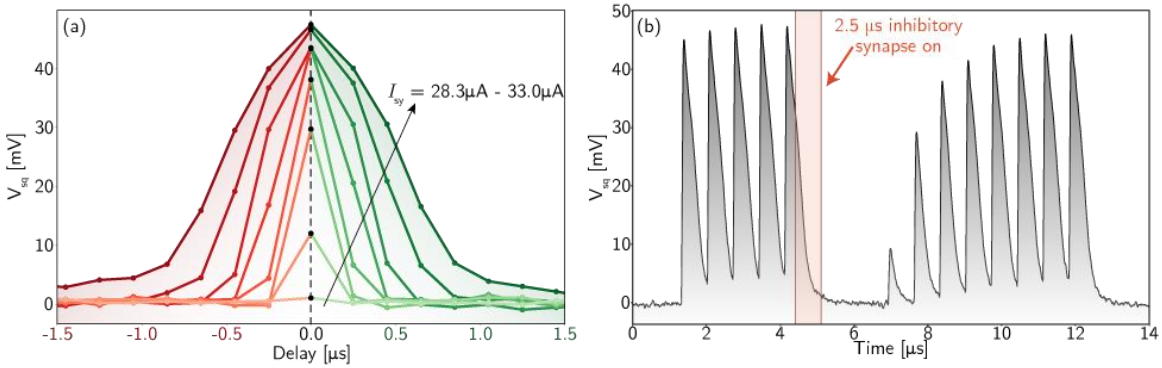



*Figure 2: Experimental demonstration of operations with two synapses connected to a common dendrite. (a) Coincidence detection. The integrated signal is only appreciable when the two synapses receive events within a microsecond time window. This operation is crucial for forming grid cells as well as for spike-timing-based learning rules. (b) Inhibition. A train of synapse events is incident upon an excitatory synapse. The dendritic signal is temporarily suppressed after the activation of an inhibitory synapse.*

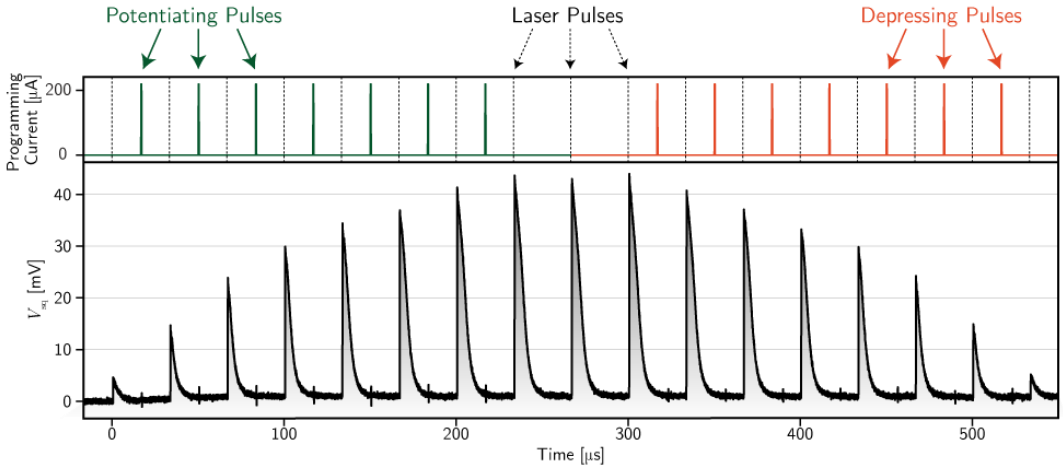



*Figure 3: Experimental demonstration of synaptic weight updates with local memory cell. A train of synapse events is interspersed with potentiating and depressing pulses to the memory cell, which increase or decrease the post-synaptic response. Here we show eight memory levels. We also demonstrated memory cells with 400 synaptic weight levels.*

# Tissue Vs Silicon – Holistic ML Systems Perspective To Unravel AGI, Scaling Laws, and Energy Efficiency

**PI**: Anshumali Shrivastava (Rice)
**Team:** Beidi Chen (CMU), Lu Mi (Georgia Tech), Saket Navlakha (CSHL), Ankit Patel (Baylor College of Medicine), Nir Shavit (MIT), Chenyan Xiong(CMU), Zhaozhuo Xu (Stevens)

The connections between computation in brains and machines have been topics of interest for scientists and philosophers for generations [1]. While a large body of work is studying the lofty goal of how these paradigms can influence each other conceptually in enabling learning and manifesting intelligence, we have only begun to mine the connections at the systems and implementation level [2,3]. Our expeditionary research will address this mostly untapped area, examining the relationship between biological neural circuits and machine learning hardware and software as efficient computing platforms [4].

Our goal will be a better understanding of how similar and different the computational infrastructure underlying these two paradigms is, and what can be deduced from one area about the other. Our work will not be just about neuromorphic computing, the mimicking of neural tissue in hardware, but rather how we can learn from brains in order to make machine learning algorithms, and computing systems in general, faster, more compact, and more energy efficient.

When one examines neural tissue, one cannot but be impressed by its efficiency. At a systems level, it has locality of reference (many neurons connect into their close-by neighbors), but also efficient long range communication between brain regions. It has synchronous behavior on the macro level (as in brain waves) and yet has asynchronous behavior at the micro level (neuron firing patterns). It is sparse in its connectivity patterns, and sparse in its firing patterns. It uses prediction of context to focus the computation (by way of inhibition) and deliver efficiency, ridiculously low power consumption, and better latency [5]. The list goes on. All of these are topics of great interest in the area of machine learning systems (MLSys), an area devoted to the design of better machine learning hardware and software. It is also of general interest in the design of parallel and distributed algorithms and data structures [6].

Two recent shifts make our research timely. On the neurobiological side, for the first time, we have access to large scale synaptic level connectivity maps of neural tissue together with neuron level recordings of its activity (the C. elegans and Fly brains, the Microns Mouse V1 dataset, etc). Ever larger datasets will be arriving in the coming years, including for the first time whole brain maps with recordings. On the MLSys side, the scale of deployment of ML has moved it from the small neural networks of the past that were easy to train and deploy, to brain scale computational systems with tens and hundreds of billions of weights and massive computing scale. Thus, MLSys direly needs breakthroughs in systems level design, and neurobiology is in the process of uncovering a platform that already has them.

The researchers setting out on this expeditionary project have already contributed to this budding area. They have designed better routing algorithms and data structures based on brain connectivity [1], better pruning algorithms, improved sparse transformer designs [7] and sparse execution algorithms both in weights and activations [8], better understanding of the computational effectiveness of neurons in both tissue and software, and the relations of neuronal computation to CS concepts such as hashing and cache efficiency.

Our goal is to bring research in neurobiology and ML systems together in order to push forward this interdisciplinary agenda: to design our future computing systems based on the principles that nature has already tested and deployed after millenia of evolutionary development.

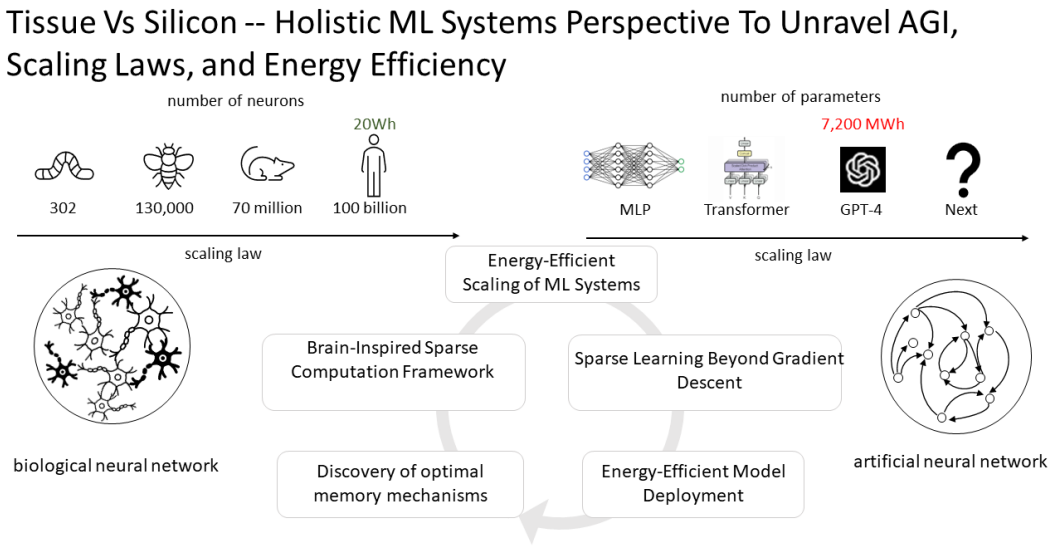



Figure 1: Summarization of our vision

## References


[1] Daniel Witvliet, Ben Mulcahy, James K Mitchell, Yaron Meirovitch, Daniel R Berger, Yuelong Wu, Yufang Liu, Wan Xian Koh, Rajeev Parvathala, Douglas Holmyard, et al. Connectomes across development reveal principles of brain maturation. *Nature*, 596(7871):257–261, 2021.

[2] Jeff W Lichtman, Hanspeter Pfister, and Nir Shavit. The big data challenges of connectomics. *Nature neuroscience*, 17(11):1448–1454, 2014.

[3] Yaron Meirovitch, Alexander Matveev, Hayk Saribekyan, David Budden, David Rolnick, Gergely Odor, Seymour Knowles-Barley, Thouis Raymond Jones, Hanspeter Pfister, Jeff William Lichtman, et al. A multi-pass approach to large-scale connectomics. *arXiv preprint arXiv:1612.02120*, 2016.

[4] Alexander Matveev, Yaron Meirovitch, Hayk Saribekyan, Wiktor Jakubiuk, Tim Kaler, Gergely Odor, David Budden, Aleksandar Zlateski, and Nir Shavit. A multicore path to connectomics-on-demand. In *Proceedings of the 22nd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, pages 267–281, 2017.

[5] Sanjoy Dasgupta, Charles F Stevens, and Saket Navlakha. A neural algorithm for a fundamental computing problem. *Science*, 358(6364):793–796, 2017.

[6] Saket Navlakha and Ziv Bar-Joseph. Distributed information processing in biological and computational systems. *Communications of the ACM*, 58(1):94–102, 2014.

[7] Beidi Chen, Zichang Liu, Binghui Peng, Zhaozhuo Xu, Jonathan Lingjie Li, Tri Dao, Zhao Song, Anshumali Shrivastava, and Christopher Re. Mongoose: A learnable lsh framework for efficient neural network training. In *International Conference on Learning Representations*, 2020.

[8] Zichang Liu, Jue Wang, Tri Dao, Tianyi Zhou, Binhang Yuan, Zhao Song, Anshumali Shrivastava, Ce Zhang, Yuandong Tian, Christopher Re, et al. Deja vu: Contextual sparsity for efficient llms at inference time. In *International Conference on Machine Learning*, pages 22137–22176. PMLR, 2023.

# Spike Coincidence as a General Control Mechanism for Neuromorphic Processing

Andrew Sornborger
Los Alamos National Laboratory

Andrew Sornborger

# Spike Coincidence as a General Control Mechanism for Neuromorphic Processing

2024 Neuromorphic Computing for Science Workshop

Position Paper

Neuromorphic computing has roots reaching back to the infancy of digital computing [1]. Concepts from neuroscience gave birth to early efforts in neural networks [2], which have taken on a life of their own [3-7]. Spike-based neuromorphic computing has emerged from the field of neuroscience [8] and many modern spiking algorithms are neuroscience-inspired [9-10]. Due to their proven Turing completeness [11], it is known that spiking algorithms can, in principle, perform arbitrary computations, including anything that is computable with standard computer hardware. In this context, neuromorphic machine learning and AI algorithms are of particular importance due to their demonstrated low power consumption [12-13] relative to standard computer hardware.

The structure of neuromorphic computers is inherently non-von-Neumann in that neuromorphic computers are often analog and consist of many simple computational elements (neurons) connected via a massively parallel communication network. Due to this unique architecture, the set of mechanisms necessary for implementing neuromorphic algorithms of any sort (learning or otherwise) must be carefully considered and developed from the point of view of both computational efficiency and computational complexity (relative to standard computers).

Our position is that the focus of neuromorphic computing must move from the current largely implementational stage (as first named by David Marr [14]) to the algorithmic and computational stages of the information processing hierarchy. These second and third stages typically involve more abstract abilities for a neuromorphic circuit to control information flow, and, in particular, where and when (within a given circuit) processing, learning, and decision making occur. At these higher levels, entire modules may be invoked or suppressed at appropriate times depending on the processing context.

This change in focus not only allows for more complex neuromorphic algorithms and computation, but also for a modular approach to neuromorphic programming. Although new neuromorphic languages such as Lava [15] have been introduced to allow high-level neuromorphic programming, the understanding and implementation of underlying supporting neuronal mechanisms to allow this have struggled to keep up.

Neuroscience provides evidence of important mechanisms used at algorithmic and computational levels for the coordination of information processing in the brain. Of particular interest is the concept of Communication Through Coherence, first introduced by Pascal Fries in 2005 [16]. In his work, Fries gives evidence that the brain makes use of probabilistic information (spike) coincidence to make decisions and gate information appropriately within the brain. Based on this concept, we have developed a framework for controlling information processing, including complex, multi-layer learning systems on neuromorphic processors [17-19]. Our framework is generally applicable to both digital and analog neuromorphic systems. Additionally, we have demonstrated that it results in low power machine learning algorithms [18-19].

Our position is that, given our previous demonstrations that entire neuromorphic machine learning algorithms that include the ability to learn on-chip with low power may be realized [18-19], the concept of spike coincidence as a general control mechanism for neuromorphic algorithms should be more fully investigated. In particular, the gating use-case of spike coincidence above can be extended for the rapid multiplication of matrices for attention-based learning mechanisms. This, and the use of network-wide broadcast of spikes for coincidence controlled modular interactions should be targeted for efficient, hierarchical neuromorphic computation.

We consider the further extension of these general concepts to be crucial for the scaling of neuromorphic algorithms to the point that they can compete with standard processors and programming frameworks.

[1] McCulloch, Warren S., and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity." *The bulletin of mathematical biophysics* 5 (1943): 115-133.
[2] Hebb, Donald Olding. *The organization of behavior: A neuropsychological theory*. Psychology press, 2005.
[3] Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." *nature* 323, no. 6088 (1986): 533-536.
[4] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems* 25 (2012).
[5] Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert et al. "Mastering the game of go without human knowledge." *Nature* 550, no. 7676 (2017): 354-359.
[6] Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool et al. "Highly accurate protein structure prediction with AlphaFold." *nature* 596, no. 7873 (2021): 583-589.
[7] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
[8] Mead, Carver. "Neuromorphic electronic systems." *Proceedings of the IEEE* 78, no. 10 (1990): 1629-1636.
[9] James, Conrad D., James B. Aimone, Nadine E. Miner, Craig M. Vineyard, Fredrick H. Rothganger, Kristofor D. Carlson, Samuel A. Mulder et al. "A historical survey of algorithms and hardware architectures for neural-inspired and neuromorphic computing applications." *Biologically Inspired Cognitive Architectures* 19 (2017): 49-64.
[10] Strukov, Dmitri, Giacomo Indiveri, Julie Grollier, and Stefano Fusi. "Building brain-inspired computing." *Nature Communications* 10 (2019): 4838-2019.
[11] Date, Prasanna, Thomas Potok, Catherine Schuman, and Bill Kay. "Neuromorphic computing is Turing-complete." In *Proceedings of the International Conference on Neuromorphic Systems 2022*, pp. 1-10. 2022.
[12] Akopyan, Filipp, Jun Sawada, Andrew Cassidy, Rodrigo Alvarez-Icaza, John Arthur, Paul Merolla, Nabil Imam et al. "Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip." *IEEE transactions on computer-aided design of integrated circuits and systems* 34, no. 10 (2015): 1537-1557.
[13] Davies, Mike, Andreas Wild, Garrick Orchard, Yulia Sandamirskaya, Gabriel A. Fonseca Guerra, Prasad Joshi, Philipp Plank, and Sumedh R. Risbud. "Advancing neuromorphic computing with loihi: A survey of results and outlook." *Proceedings of the IEEE* 109, no. 5 (2021): 911-934.
[14] Marr, David, and Tomaso Poggio. "From understanding computation to understanding neural circuitry." AI Memos: AIM-357, (1976).
[15] Lava: A Software Framework for Neuromorphic Computing, Intel Labs. Lava Software Framework. https://lava-nc.org Accessed: 2024-7-10.
[16] Fries, Pascal. "Rhythms for cognition: communication through coherence." *Neuron* 88, no. 1 (2015): 220-235.
[17] Wang, Zhuo, Andrew T. Sornborger, and Louis Tao. "Graded, dynamically routable information processing with synfire-gated synfire chains." *PLoS computational biology* 12, no. 6 (2016): e1004979.
[18] Renner, Alpha, Forrest Sheldon, Anatoly Zlotnik, Louis Tao, and Andrew Sornborger. "The backpropagation algorithm implemented on spiking neuromorphic hardware." *arXiv preprint arXiv:2106.07030* (2021).
[19] On-Chip Neuromorphic Backpropagation Algorithm, Category: Software/Services: www.rdworldonline.com/rd-100-2022-winner/on-chip-neuromorphic-backpropagation-algorithm/, Accessed: 2024-7-10

# The Case for Waferscale Neuromorphic Architecture using Asynchronous Stream Computing

Mircea Stan
University of Virginia

**The Case for Waferscale Neuromorphic Architecture using Asynchronous Stream Computing**

The landscape of artificial intelligence has been dramatically transformed by scaling up the *depth and number of parameters/weights* in conventional AI models. This approach of deeper and deeper models with ever more number of parameters/weights has been pivotal in taking conventional Artificial Neural Networks (ANNs) from the old days of low-depth networks, such as the perceptron which showed interesting behavior but was so limited that it triggered an "AI winter," to the first truly deep models such as LeNet-5 Convolutional Neural Network (CNN) to the current generative models based on transformers with trillions of parameters but also with "superhuman" performance on many tasks. The only major drawback of the conventional ANN approach is their energy inefficiency which is unsustainable in the long run. These conventional ANNs are *inspired* by biological neural networks and the brain but only in a *high-level abstract way* implemented with digital circuits. In parallel with these conventional ANNs there has been an effort to create *spiking neural networks (SNNs)* which are trying to more *faithfully mimic features of biological neural networks* such as spiking behavior, spike-timing dependent plasticity (STDP), etc. Several commercial examples of this neuromorphic approach are TrueNorth from IBM and Loihi from Intel, with an extreme case in the Human Brain Project in the EU. While these neuromorphic approaches have showen promise it is quite cleat by now that the gap between conventional ANNs and neuromorphic ones is actually increasing, not decreasing as time goes by.

We believe that neuromorphic solutions don't need to try to *faithfully* mimic *all* biological features but only use *some* of the biologically-inspired aspects (such as the *asynchronous time-domain data* representation) while adopting the *scaled-up depth and large number of parameters/weights* of conventional ANNs. After all a plane does not flap its wings yet it arguably outperforms any biological bird.

This position paper advocates for the implementation of *waferscale ASC-based deep neuromorphic architectures based on Asynchronous Stream Computing (ASC)*, emphasizing energy efficiency, scalability, and performance, primarily addressing the theme of *translating neuroscience-inspired principles to analog microelectronic circuits* and secondarily focusing on *performance metrics and energy efficiency.*

Central to the ASC paradigm is the encoding of information into asynchronous streams of "ones" and "zeros", Fig. 1 [1]. This method captures the input signal characteristics, such as magnitude and rate of change, using the temporal domain for data representation. Unlike traditional digital systems, which often face limitations in parallelism and power consumption, ASC leverages the analog time domain to achieve significant energy efficiency. By employing continuous-time asynchronous sigma-delta modulators (ASDM), ASC generate and manipulate pulses akin to spikes that are naturally event-driven. The architecture features Compute-in-Memory (CiM) tiles, Fig. 2 [2], which bypass the limitations of traditional RRAM-based CIM solutions by eliminating the need for DAC/ADCs. This not only enhances energy efficiency but also reduces area and cost, making ASC a viable scalable solution for complex AI models.

One of the most compelling advantages of waferscale ASC is its potential for unprecedented energy efficiency. By mimicking the energy-efficient mechanisms of biological neurons, ASC reduces the power consumption typically associated with digital computations. The temporal encoding of data in ASC streams enables fine-grained control of power and performance trade-offs, crucial for handling large-scale scientific models and neural network ensembles. The architecture further enhances this efficiency through a hierarchical programming methodology, allowing directed configuration and reconfiguration of ASC cores, thus optimizing computational flow and minimizing energy expenditure, Fig. 3. The scalability of ASC-based architectures is another critical benefit. Traditional digital systems often struggle with scaling due to increased power and thermal issues. In contrast, waferscale ASC can accommodate extensive models on a single computational engine.

The transition to waferscale neuromorphic architectures using Asynchronous Stream Computing will represent a significant leap forward in AI neuromorphic hardware acceleration. By harnessing the principles of neuroscience and translating them into analog time-domain microelectronic circuits, ASC-based architectures offer a path to scalable, energy-efficient, and high-performance AI systems, overcoming the limitations of current SNNs and conventional digital solutions.

**References**


1. P. Gonzalez-Guerrero and M. R. Stan, “Asynchronous stochastic computing,” in 2019 53rd Asilomar Conference on Signals, Systems, and Computers, 2019, pp. 280–285 (Best paper award).
2. R. Sreekumar and M. R. Stan, “Easi-CIM: Event-driven asynchronous stream-based image classifier with compute-in-memory kernels.” ISQED 2024 (Best paper award)


**Figures**

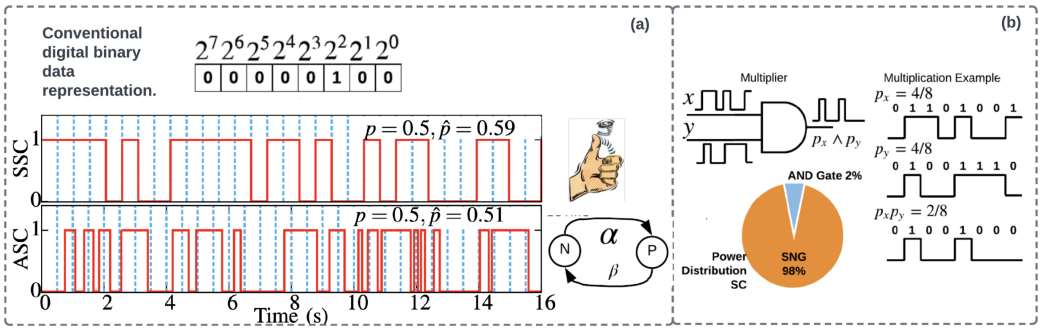



Fig. 1 Asynchronous Stream Computing (ASC) data representation

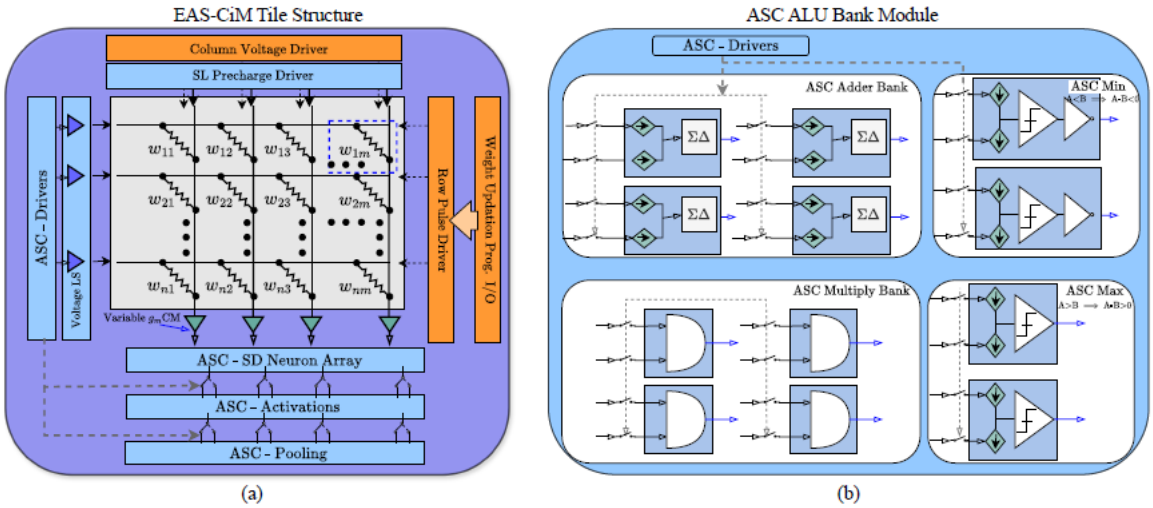



Fig. 2 Compute-in-Memory (CiM) tile using ASC

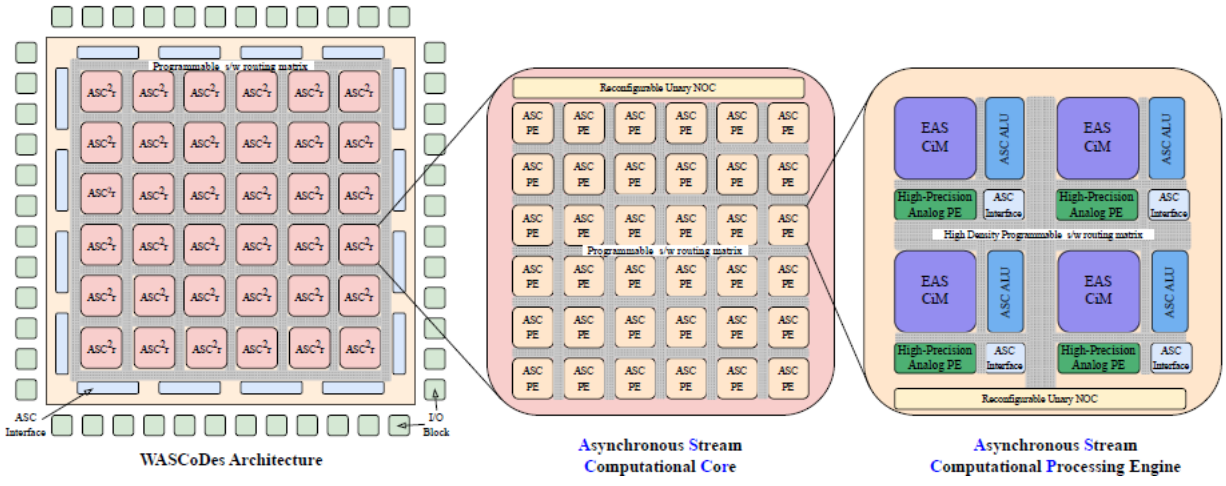



Fig. 3 Hierarchical reconfigurable waferscale neuromorphic ASC architecture

# Scaling Photonic Interconnects for High Neuromorphic Connectivity

Ishan G Thakkar, University of Kentucky, Lexington, KY
igthakkar@uky.edu

A typical mammalian neuron maintains parallel physical contact with ~10,000 other neurons. It is widely recognized that photonic interconnects could achieve such high (parallel) connectivity because of their massive wavelength-multiplexing-based parallelism and support for near-dissipation-free optical signal propagation [1]. Despite these advantages, however, no prototype of photonic interconnects has yet demonstrated even 5% of the required neuromorphic connectivity (~500 spatially parallel optical connections). Here, we summarize the challenges and opportunities the neuromorphic photonic interconnects research faces, and the future directions that could be taken, to meet this yet-elusive goal of realizing high neuromorphic connectivity.

**Shortcomings and Challenges.** Researchers have explored several different approaches for realizing photonic interconnects for neuromorphic connectivity. Among these, the most popular approach is to use integrated photonics. Using integrated photonics to realize interconnects offers compactness and energy efficiency. However, this approach faces the following shortcomings. *(1)* Lack of scalable, multi-wavelength, on-chip light sources. Recent advancements have enabled the realization of on-chip-integration-feasible optical comb sources that can provide up to 500 optical wavelength channels over the spectral span of ~ 200 nm across the O band or the C and L bands [2]. But 500 wavelength channels are insufficient, as they can barely cover 5% of the required connectivity of ~10,000 contacts per neuron. *(2)* Small free spectral range for wavelength multiplexing. The most common approach for realizing the required multiplexing and other optical signal manipulations (e.g., routing to realize flexible fan-in, and synaptic weighting) is to employ wavelength-selective active resonant devices. However, the most practical devices have a free spectral range of <35 nm, which is very small compared to the 200 nm spectral range available for exploitation through optical comb sources. *(3)* Low dependability of wavelength-selective devices. The wavelength-selective resonant devices typically employed in photonic interconnects show high device-to-device and chip-to-chip variability and high susceptibility to on-chip temperature gradients [3]. This keeps them from being useful for realizing scalable connectivity. *(4)* Unknown power handling limits of active devices. Most commonly used materials for realizing photonic interconnects (e.g., III-V, silicon, germanium, lithium niobate) exhibit multiple optical nonlinear effects. These nonlinear effects often compete in the active resonant devices made of these materials, inducing instability and/or metastability in the devices for specific values of resonance detuning, quality factors, and input optical power [4]. This effect is likely to limit the achievable connectivity to a few tens of optical signals per neuron [5]. *(5)* Optical losses versus electro-optic activity trade-off. Most electro-optically active devices, which are required in photonic interconnects for realizing essential optical signal manipulations (e.g., routing, synaptic weighting), exhibit moderate-to-high optical losses. The fundamental problem is the presence of a trade-off between electro-optic activity and optical signal losses in the materials employed for realizing these devices. For instance, silicon nitride-based devices exhibit low optical losses [6]. But they lack electro-optic activity [7], [8].

**Opportunities and Future Research Directions**. Prospects for integrated photonic interconnects are not all gloomy though. Opportunities exist for high-impact, transformative research *(1)* to design low-area, high-efficiency, and low-latency photonic resonant devices with very-large-FSR [9] or FSR-free [10], [11] operations, to consequently enable efficient and full use of the comb sources' available spectral range of 200 nm, *(2)* to innovate for realizing highly stable yet wavelength-selective resonant devices with massive fan-in, *(3)* to forge new designs of slow-light modulators [12], [13] to achieve dependable performance for low-overhead, on-chip light manipulation, and *(4)* to enable efficient use of optical mode division multiplexing [14], [15], [16] to multiply homodyne fan-in for realizing massive connectivity.

**References**


[1] P. L. McMahon, “The physics of optical computing,” *Nat Rev Phys*, vol. 5, no. 12, Art. no. 12, Dec. 2023, doi: 10.1038/s42254-023-00645-5.

[2] A. Rizzo *et al.*, “Massively scalable Kerr comb-driven silicon photonic link,” *Nat. Photon.*, vol. 17, no. 9, Art. no. 9, Sep. 2023, doi: 10.1038/s41566-023-01244-7.

[3] A. Mirza, F. Sunny, P. Walsh, K. Hassan, S. Pasricha, and M. Nikdast, “Silicon Photonic Microring Resonators: A Comprehensive Design-Space Exploration and Optimization Under Fabrication-Process Variations,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 41, no. 10, pp. 3359–3372, Oct. 2022, doi: 10.1109/TCAD.2021.3132555.

[4] M. de Cea, A. H. Atabaki, and R. J. Ram, “Power handling of silicon microring modulators,” *Opt. Express, OE*, vol. 27, no. 17, pp. 24274–24285, Aug. 2019, doi: 10.1364/OE.27.024274.

[5] V. S. P. Karempudi, J. Bashir, and I. G. Thakkar, “An Analysis of Various Design Pathways Towards Multi-Terabit Photonic On-Interposer Interconnects,” *J. Emerg. Technol. Comput. Syst.*, vol. 20, no. 2, p. 6:1-6:34, Feb. 2024, doi: 10.1145/3635031.

[6] D. B. \em et al, “Silicon nitride in silicon photonics,” *Proceedings of the IEEE*, 2018.

[7] Q. W. \em et al, “A versatile silicon-silicon nitride photonics platform for enhanced functionalities and applications,” *Applied Sciences*, 2019.

[8] R. Baets *et al.*, “Silicon Photonics: silicon nitride versus silicon-on-insulator,” in *Optical Fiber Communication Conference (2016), paper Th3J.1*, Optica Publishing Group, Mar. 2016, p. Th3J.1. doi: 10.1364/OFC.2016.Th3J.1.

[9] D. Liu, C. Zhang, D. Liang, and D. Dai, “Submicron-resonator-based add-drop optical filter with an ultra-large free spectral range,” *Opt. Express, OE*, vol. 27, no. 2, pp. 416–422, Jan. 2019, doi: 10.1364/OE.27.000416.

[10] N. Eid, R. Boeck, H. Jayatilleka, L. Chrostowski, W. Shi, and N. A. F. Jaeger, “FSR-free silicon-on-insulator microring resonator based filter with bent contra-directional couplers,” *Opt. Express, OE*, vol. 24, no. 25, pp. 29009–29021, Dec. 2016, doi: 10.1364/OE.24.029009.

[11] F. Morichetti *et al.*, “Polarization-transparent silicon photonic add-drop multiplexer with wideband hitless tuneability,” *Nat Commun*, vol. 12, no. 1, p. 4324, Jul. 2021, doi: 10.1038/s41467-021-24640-5.

[12] A. Opheij, N. Rotenberg, D. M. Beggs, I. H. Rey, T. F. Krauss, and L. Kuipers, “Ultracompact (3 μm) silicon slow-light optical modulator,” *Sci Rep*, vol. 3, no. 1, p. 3546, Dec. 2013, doi: 10.1038/srep03546.

[13] G. Chen *et al.*, “Compact slow-light waveguide and modulator on thin-film lithium niobate platform,” *Nanophotonics*, vol. 12, no. 18, pp. 3603–3611, Sep. 2023, doi: 10.1515/nanoph-2023-0306.

[14] K. R. Mojaver, S. M. R. Safaee, S. S. Morrison, and O. Liboiron-Ladouceur, “Recent Advancements in Mode Division Multiplexing for Communication and Computation in Silicon Photonics,” Apr. 04, 2024, *arXiv*: arXiv:2404.03582. Accessed: Jul. 23, 2024. [Online]. Available: http://arxiv.org/abs/2404.03582

[15] P. Guo, N. Zhou, W. Hou, and L. Guo, “StarLight: a photonic neural network accelerator featuring a hybrid mode-wavelength division multiplexing and photonic nonvolatile memory,” *Opt. Express, OE*, vol. 30, no. 20, pp. 37051–37065, Sep. 2022, doi: 10.1364/OE.468456.

[16] R. Yin *et al.*, “Integrated WDM-compatible optical mode division multiplexing neural network accelerator,” *Optica, OPTICA*, vol. 10, no. 12, pp. 1709–1718, Dec. 2023, doi: 10.1364/OPTICA.500523.

# Bio-inspired Adaptive and Self-Organized Learning

Dmitrii Torbunov <dtorbunov@bnl.gov>, CSI, Brookhaven National Laboratory

July 19, 2024

**Overview.** Biological brains exhibit remarkable properties that artificial neural networks have yet to fully emulate. These include adaptive learning rates, self-organized learning, efficient inhibition mechanisms, and the ability to learn continuously from a stream of data without catastrophic forgetting (cf. Figure 1). Understanding and implementing these bio-inspired features in artificial neural networks could potentially lead to more robust, efficient, and adaptable AI systems.

This whitepaper highlights several remarkable differences between learning mechanisms in the traditional artificial neural networks and biological neural networks. It proposes several potential avenues for bringing the bio-inspired learning to artificial networks. The proposed learning approaches have potential to synergize with the neuromorphic hardware, due to its similarity with biological networks. However, their application can be also explored with the traditional Deep Learning (DL) architectures to workaround their limitations and further advance the field of artificial intelligence in general.

**Adaptive Learning Rate Adjustment.** Biological brains are able to adjust their learning rates based on either the current prediction error or based on received rewards/punishments [2]. These learning rate adjustments are mediated by neurotransmitters and allow animals to quickly incorporate new information into the world model and reinforce rewarding behaviors. The release of neurotransmitters is controlled by the brain itself. That is, the brain can adjust its learning rate depending on the sensory inputs and its internal state. Similar bio-inspired learning rate adjustment algorithms can be explored for Artificial Neural Networks (ANNs) as means of more efficient learning, or as alternative strategies for Reinforcement Learning.

**Self-Organized Learning.** Unlike the traditional Deep Learning models trained by backpropagation of errors, biological brains present a completely different, self-organized learning mode [3, 1]. There are no computational graphs built in animal brains. There is no central agent that synchronizes computations layer by layer. There is no global optimizer that updates neuron weights. In biological brains, each neuron is an independent unit. Each neuron chooses when to propagate information. Each neuron learns independently by locally interacting with its neighbors. There are no explicit training/inference modes in the brain. Instead, a brain is able to learn continuously just by observing a stream of data. Animal brains have a remarkable ability to learn new modalities of data, and apparently do not suffer from catastrophic forgetting. The high adaptability of self-organized learning indicates that such learning has enormous potential for ANNs.

**Adaptive Self-Inhibition.** Biological brains are capable of selectively inhibiting a large fraction of neurons with the help of a special class of GABAergic neurons [1]. This inhibition mechanism allows the brain to save energy by avoiding useless computations. It also regulates self-organizing learning, distributing the computations evenly across the neighboring neurons, allowing them to learn different features from the data. The use of adaptive inhibition mechanisms holds potential to improve the energy efficiency of ANNs and sparsify their activations.

**Impact.** To illustrate potential applications, a conventionally pre-trained large language model (LLM) can be endowed with a self-organized learning mode. This will allow the LLM to continue to learn by simply interacting with users and environment, similar to a biological brain. Next, to make this model multi-modal, one can just inject image tokens (from a pre-trained image encoder) into the stream of text. The high adaptability of self-organized learning will allow the LLM to learn to align image representations to the LLM's world model and use new image data efficiently. Moreover, since the computational requirements of self-organized learning match requirements of a model's forward pass, the whole training could be done on a single GPU.

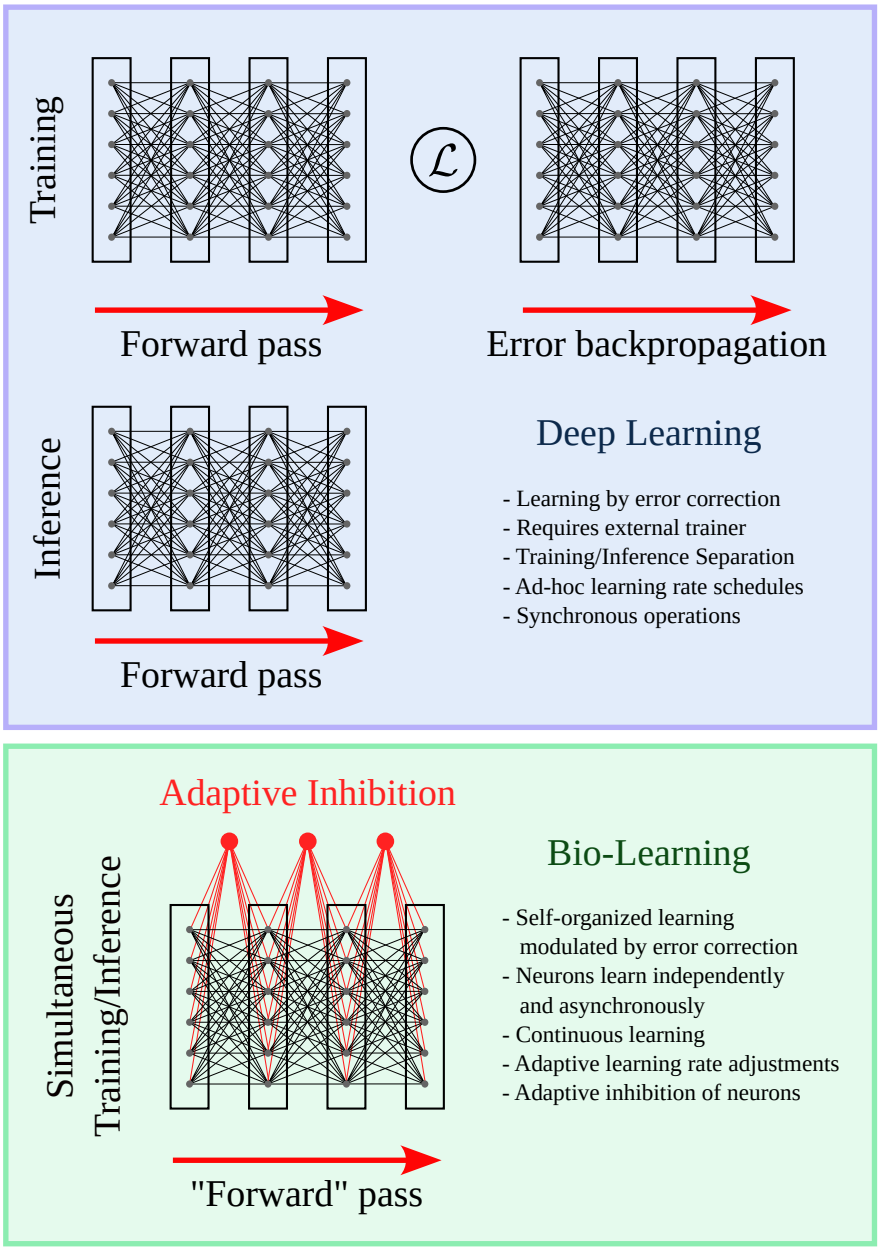



Figure 1: Some of the differences between Deep Learning and learning in Biological Brains.

## References

[1] *Computational Cognitive Neuroscience*. Online Book, 4th Edition, URL: `https://CompCogNeuro.org`, 2012.

[2] Jessica A Mollick, Thomas E Hazy, Kai A Krueger, Ananta Nair, Prescott Mackie, Seth A Herd, and Randall C O'Reilly. A systems-neuroscience model of phasic dopamine. *Psychological review*, 127(6):972, 2020.

[3] Yuko Munakata and Jason Pfaffly. Hebbian learning and development. *Developmental science*, 7(2):141–148, 2004.

Neuromorphic compressed temporal representation using spiking autoencoder

Inton Tsang, W. Van Leekwijck, M. Hartmann, S. Latré, J.M. Oramas, and I.R. Tsang
IMEC

# Neuromorphic compressed temporal representation using spiking autoencoder

Tsang IJ, Van Leekwijck W, Hartmann M, Latré S, Oramas JM, Tsang IR

One of the most significant challenges in processing streaming data generated from sensors, such as radar, multi-hyperspectral, or RGB frame-based cameras, is the sheer amount of data to be analyzed. Even dynamic vision sensor (DVS) can generate a stream of data that is overwhelming for machine learning systems. Data sampling or techniques in data reduction, which can incur information loss, are applied to mitigate and accommodate the hardware resources available for the tasks in terms of memory, processing, latency, and power consumption. Biological neural systems naturally ingest and process massive amounts of data efficiently. In modeling and simulating neuromorphic systems, spiking autoencoder architectures can compress temporal data, enabling a system that can process large amounts of sensor signals and seamlessly integrate with other computing frameworks, considering all the signal data produced by the sensors.

For this purpose, we propose using spike autoencoders (SAE) to encode and represent data. Figure1 shows the results of the spike autoencoder on the MNIST and CIFAR-10 datasets. In contrast, Figure 2 shows that the results obtained by the proposed method on the CIFAR-10 dataset achieve better Frechet Inception Distance (FID) and Inception Score (IS) scores than in [1]. Moreover, Figure 3 shows SAE applied to Frequency-Modulated Continuous Wave (FMCW) radar data, demonstrating the capability to learn representation, irrespective of sensor signal type. Autoencoders intrinsically compress input signals into the latent space dimension, representing a spatial compression space used for further downstream applications such as recognition or detection. To achieve a compressed temporal representation, we explore the temporal nature of SNNs. Figure 4 shows the basic concept. At training, we repeat the input $n$ times for the SAE to learn to reproduce the signal, while at inference, the system will ingest the signal as it comes, whereas the output of the SAE depends on the temporal sequence of the stream of data. For conceptual purposes, Figure 5 depicts this effect on the MNIST samples, showing a sequence of 10 inputs from left to right, the average image $< i(t) >$ of these inputs, and then the sequence of 10 outputs of the SAE $AE(i(t))$. There are two opposing sequences for each digit, showing the temporal dependency of the output signal. This can be attributed to a memory effect due to the stateful nature of the integrate and fire neural model. Figure 6 shows an example of applying the scheme on a sequence of radar data cubes. It shows the last range-Doppler radar cube from a sequence that contained 192 cubes and compressed to 24 cubes, thus an 8x factor.

To evaluate the quality of the compressed temporal representation, we applied the scheme to the Soli radar hand gesture dataset [2]. We compressed each gesture sequence of range-Doppler data and applied a convolution spiking neural network, achieving SOTA accuracy rates >99% using all available channels of the dataset. While the compressed latent space of the SAE can be used for the downstream application tasks, using the compressed temporal representation output reconstructs the physical dimensionality of the signal, i.e., range-Doppler for radar, spectrum values for multi-hyperspectral data or accumulated events from DVS frames. This scheme has several advantages, such as compatibility, as it allows for seamless integration with previous or legacy systems. For example, any post-processing or previous machine learning systems can continue to be compatible with the data stream. Second, modularity, as any changes or updates done on the SAE module, will not impact the other modules. Finally, flexibility, as the system does not require end-to-end training, an update on the SAE does not imply a need to re-train any other machine learning algorithms used after this module and the other way around.

This position paper addresses theme 3, modeling and simulation approaches. We hope to contribute to the scalable integration thrust. We demonstrate a scheme in which neuromorphic computing processes massive amounts of streaming data generated by sensor devices while proposing an architecture that allows seamless integration with other modules and computing frameworks.

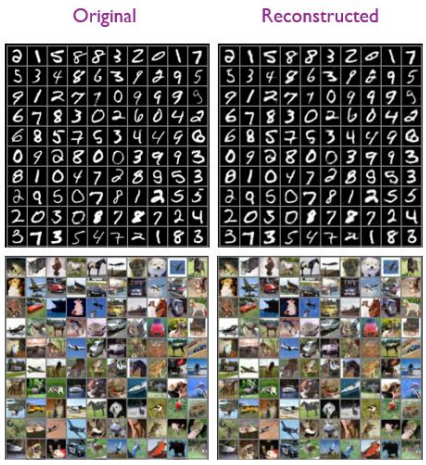



Figure1 Spiking Autoencoder on MNIST and CIFAR-10

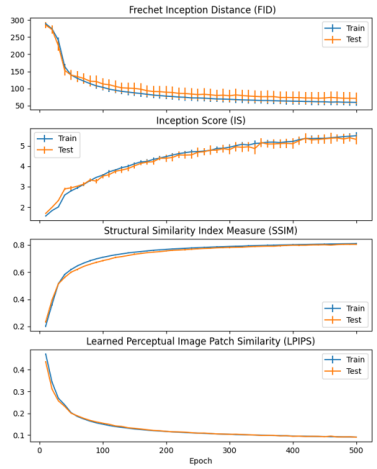



Figure 2 Reconstruction quality metrics for CIFAR-10

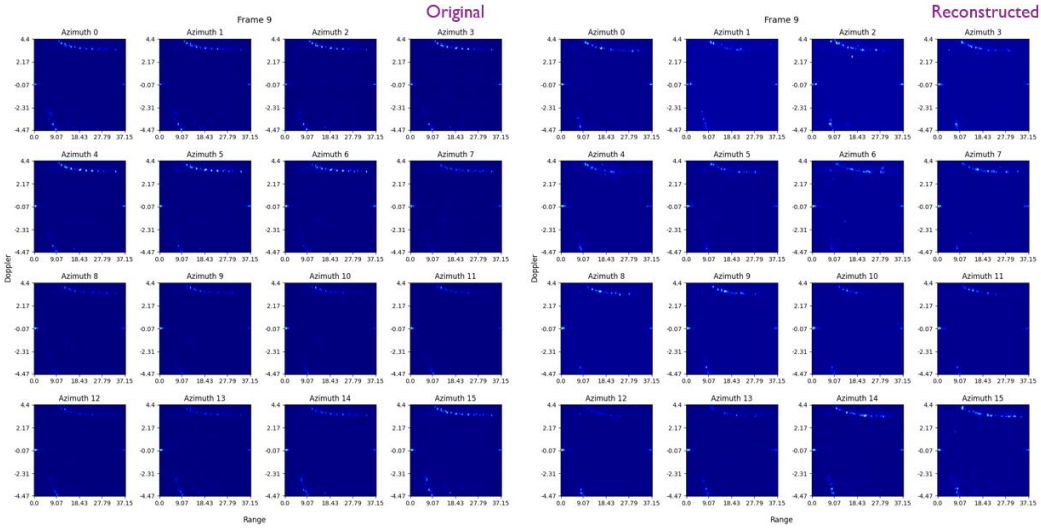



Figure 3 Reconstruction of FMCW range-Doppler data on different azimuths

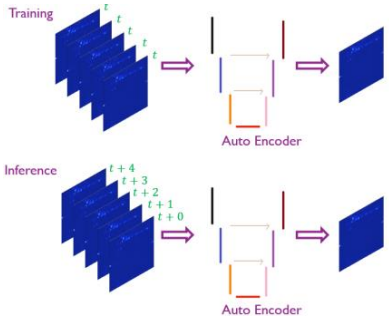



Figure 4 Exploring SNN temporal nature for compression.

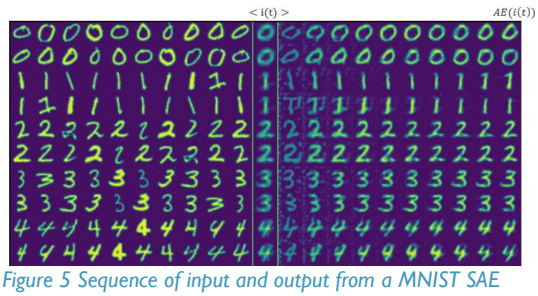



Figure 5 Sequence of input and output from a MNIST SAE

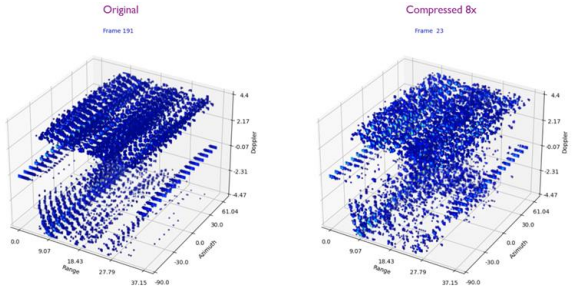



Figure 6 Compressed temporal representation on radar signal

[1] H. Kamata, Y. Mukuta, and T. Harada, "Fully Spiking Variational Autoencoder", AAAI, vol. 36, no. 6, pp. 7059-7067, Jun. 2022.
[2] Tsang IJ, Corradi F, Sifalakis M, Van Leekwijck W, Latré S. Radar-Based Hand Gesture Recognition Using Spiking Neural Networks. *Electronics*. 2021; 10(12):1405.

Co-Design Methodologies for Integrating Small Organism-Inspired Chiplets

Lav Varshney, José Schutt-Ainé, Shaloo Rakheja, and Saugata Ghose
University of Illinois Urbana-Champaign

**Title:** Co-Design Methodologies for Integrating Small Organism-Inspired Chiplets
**Authors:** Lav R. Varshney (with José Schutt-Ainé, Shaloo Rakheja, and Saugata Ghose)
**Topic:** algorithms, modeling and simulation, architectures, emerging technologies

**Challenges:** The rapid growth of AI and ML for science and the increasing demand for computing power will only exacerbate in the next five years. Memory walls and limitations of current AI chip packaging remain serious challenges. Novel technologies such as neuromorphic computing and chiplets are poised to become adopted solutions to address these barriers, but they have remained separated, rather than considered together as a unified solution approach. Indeed, neuromorphic systems have largely remained homogeneous monoliths that are only inspired by mammalian cortex, rather than having heterogeneous modularity that is inspired by the full armamentarium of neurobiology that includes small organisms. Moreover, design tools and methodologies are seriously lagging to leverage these technologies. Further, current neuromorphic systems have yet to demonstrate the cognitive functionality of mainstream AI methods (e.g., deep nets) or the energy efficiency of the brain. Existing approaches do not fully embrace the notion of co-design and co-optimization across the different layers of the system design hierarchy, which is needed to achieve neuromorphic systems that can tackle real-world problems, especially AI for science. Function disaggregation which allows architecting an integrated circuit (IC) from a system on chip (SOC) to a chiplet-based system in package (SiP) is currently performed ad hoc. There is no established methodology for optimization to help determine how many separate chiplets should be used to meet specifications. Once the chiplet design landscape becomes more democratized, there will be more choices from different vendors which will make disaggregation more chaotic.

**Opportunity:** To implement future computing systems with advanced nodes, the semiconductor community is shifting from traditional monolithic approaches due to inevitable physical constraints of Dennard scaling in semiconductor devices. Instead, the multi-chip ("chiplets") approach [1–3] is gaining momentum and support from chip foundries, IP vendors, assemblers, and testers due to its multi-fold advantages such as high yield, low wafer cost, decreased time-to-market, and more. On top of that, advanced packaging technologies for 2.5D, 3D (stacking) ICs such as Intel's Foveros [4] or TSMC's CoWoS [5] enable more complicated design methodologies at the chip level, give the designers more flexibility but also blur the boundary between chip and package design. The highly-efficient anatomical and functional neurobiologies of small organisms such as *Caenorhabditis elegans*, *Megaphragma mymaripenne*, *Ciona intestinalis*, and *Heterodera glycines* are becoming better understood [6-9], so as to inspire basic worm, microwasp, and tunicate chiplets that can be assembled together. There are maturing nanoscale neuro-primitives based on CMOS-integrable emerging materials, like magnetics, ferroelectrics, and 2D materials, with advanced functionalities such as non-volatility, fusion of logic and memory, reconfigurability, and ability to replicate brain dynamics.

**Timeliness or maturity:** There is a massive energy-capability performance gap between neurobiology and current information processing systems. We therefore argue that it is imperative that new computer architecture and system-level integration be implemented soon. We believe there is need for a multidisciplinarity that will enable artificial general intelligence (AGI)-capable chips that comprise chiplets inspired by not just mammalian neocortex but also small-organism-neurobiology, and use novel physics in beyond-silicon materials that can be harnessed for computing purposes. These must be coupled with heterogeneous integration. Strong collaboration will be needed, using co-design methods that encompass cross-layer (materials, devices, circuits, systems), multiple domain (analog, mixed-signal, digital), multi-physics (electrical, thermal, mechanical, optical), and diverse neuroscience. A viable approach to research would consist of: (1) Exploring neurobiology-driven algorithms and hardware software co-design methodologies; (2) Modeling intelligent materials and devices; (3) Designing scalable circuit macros and architectures; (4) Integrating heterogeneous components and chiplets at the system level. In summary, the goal is to use neural design principles with rigorous information-theoretic foundations and materials-to-systems approach to overcoming the longstanding speed, energy, and cognitive capability limitations of today's AI hardware for scientific discovery.

**References**

[1] L. T. Su, S. Nafziger, and M. Papermaster, “Multi-chip technologies to unleash computing performance gains over the next decade,” in 2017 IEEE International Electron Devices Meeting (IEDM), 2017, pp. 1.1.1–1.1.8.

[2] G. H. Loh, S. Nafziger, and K. Lepak, “Understanding Chiplet Today to Anticipate Future Integration Opportunities and Limits,” in 2021 Design, Automation and Test in Europe Conference and Exhibition (DATE), 2021, pp. 142–145.

[3] W. T. Beyene, “Chiplet technology and heterogeneous integration,” IEEE EPS eNews, Apr. 2022. [Online]. Available: https://eps.ieee.org/publications/enews/april-2022/ 866-chiplet-technology-and-heterogeneous-integration-2.html.

[4] D. B. Ingerly, S. Amin, L. Aryasomayajula, A. Balankutty, D. Borst, A. Chandra, K. Cheemalapati, C. S. Cook, R. Criss, K. Enamul, W. Gomes, D. Jones, K. C. Kolluru, A. Kandas, G.-S. Kim, H. Ma, D. Pantuso, C. Petersburg, M. Phen-givoni, A. M. Pillai, A. Sairam, P. Shekhar, P. Sinha, P. Stover, A. Telang, and Z. Zell, “Foveros: 3D Integration and the use of Face-to-Face Chip Stacking for Logic Devices,” in 2019 *IEEE International Electron Devices Meeting (IEDM),* 2019, pp. 19.6.1–19.6.4.

[5] S.-P. Jeng, “CoWoS technologies,” in *Proceedings of Technical Program - 2014 International Symposium on VLSI Technology, Systems and Application (*VLSI-TSA), 2014.

[6] L. R. Varshney, B. L. Chen, E. Paniagua, D. H. Hall, and D. B. Chklovskii, “Structural Properties of the *Caenorhabditis elegans* Neuronal Network,” *PLoS Computational Biology*, vol. 7, no. 2, e1001066, February 2011.

[7] A. A. Makarova, A. A. Polilov, and D. B. Chklovskii, “Small brains for big science,” *Current Opinion in Neurobiology*, vol. 71, pp. 77-83, 2021.

[8] C. Bhatia and L. R. Varshney, “Structural Properties of the *Ciona intestinalis (L.)* Connectome,” presented at *SfN Global Connectome*, 11-13 January 2021.

[9] “New research looks to combat SCN through neuroscience,” https://csl.illinois.edu/news-and-media/32847

Enabling scalable neuromorphic systems with error-aware simulation frameworks

Cheng Wang
Iowa State University

Enabling scalable neuromorphic systems with error-aware simulation frameworks

Biological neural systems that orchestrate diverse modalities of neurons and synapses demonstrate high efficiency of cognitive processing. However, emulating the underlying computation of biological neural systems using analog device/circuit primitives and non-von Neumann architectures encounters the challenge of accuracy loss because of analog errors from hardware non-idealities. As of today, while neuromorphic hardware primitives have shown promising results at small-scale tasks such as MNIST hand-written digit classifications [6, 5, 4], implementing neuromorphic systems for large-scale complex applications remains challenging. We envision that one of the critical characteristics for effective large-scale simulation of neuromorphic systems is **understanding the implication of hardware errors and the related design space** that comprises error-aware hardware heterogeneity and optimizing the trade-off between functional accuracy and hardware efficiency.

In this position paper, we focus on **Theme-3:Modeling and simulation approaches**, and identify crucial topics on developing error-aware neuromorphic circuits and systems for enabling scalable neuromorphic computing. **The scalability metrics** include model sizes (number of parameters) and input sizes (ranging from MNIST to CIFAR-10 and ImageNet). As summarized in Fig.1, such developments inherently embody hardware-software co-design, calling for research efforts across the stack from customizing analog device/circuit designs to developing robust and efficient neuro-inspired algorithms.

First, it is imperative to develop a methodology with suitable metrics to analyze error sensitivity for given algorithm-level workloads. For example, **circuit-level metrics** can be the normalized difference between ideal and non-ideal voltage/current, while **system-level metrics** can be loss of inference accuracy. We will then use the metrics to profile the impacts of error on different parts of a workload. Generic perturbation-based sensitivity or saliency-based evaluation can be developed to quantify the impacts of errors on various portions of an algorithm [7, 10, 3]. Such workload analysis will be instrumental in partitioning high-level workloads at different granularity when implementing a system with inhomogeneous error resiliency [8]. For example, synaptic crossbar arrays with higher bits per cell achieve high storage density but suffer more from device/circuit variations. Hence, the most (least) sensitive portion of the workload may choose single- (multi-) bit representations of weight storage.

Next, to characterize the realistic behavior of neuro-inspired computing primitives, we will develop technology-aware simulations of devices/circuits based on measurement and simulation data (**related to Theme-2 on analog circuits)**. Both analytical and data-driven machine learning (ML) based models can be employed to emulate the device/circuit behaviors. It is important to note that while analytical solvers are the standard approach, ML-based emulators informed with underlying physics can potentially speed up the overall process while achieving satisfactory accuracy, circumventing excessive device-level or SPICE simulations [1, 2].

Moreover, the error-efficiency trade-off of neuromorphic architecture will be investigated as a part of the co-design framework. Leveraging the algorithmic robustness in neuro-inspired computing models, architecture-dependent simulations need to quantify the hardware efficiency and accuracy of the non-ideal circuits and devices at varying magnitudes of analog errors[9]. Conceptually, the analog error becomes a tunable knob to achieve a balanced performance in a heterogeneous system with reconfigurable building blocks. Quantifying the response of the neuromorphic system to errors helps bridge the gap between the device/circuit error characterization and the design of robust and efficient systems.

Finally, we envision that the cross-layer co-design of large-scale neuromorphic computing systems will also incorporate optimizing the algorithm-level neuro-inspired models with the awareness of analog errors. A likely scenario is that the erroneous hardware primitives will prefer a different algorithm design point than the ideal hardware. Exemplary differences include deep versus wide neural network topologies, or sparse versus dense weight matrices. Since the device/circuit errors impact the functional accuracy and are possibly linked with hardware efficiency, reinforcement learning-based automated network architecture search (NAS) will be implemented to cope with the ample design space.

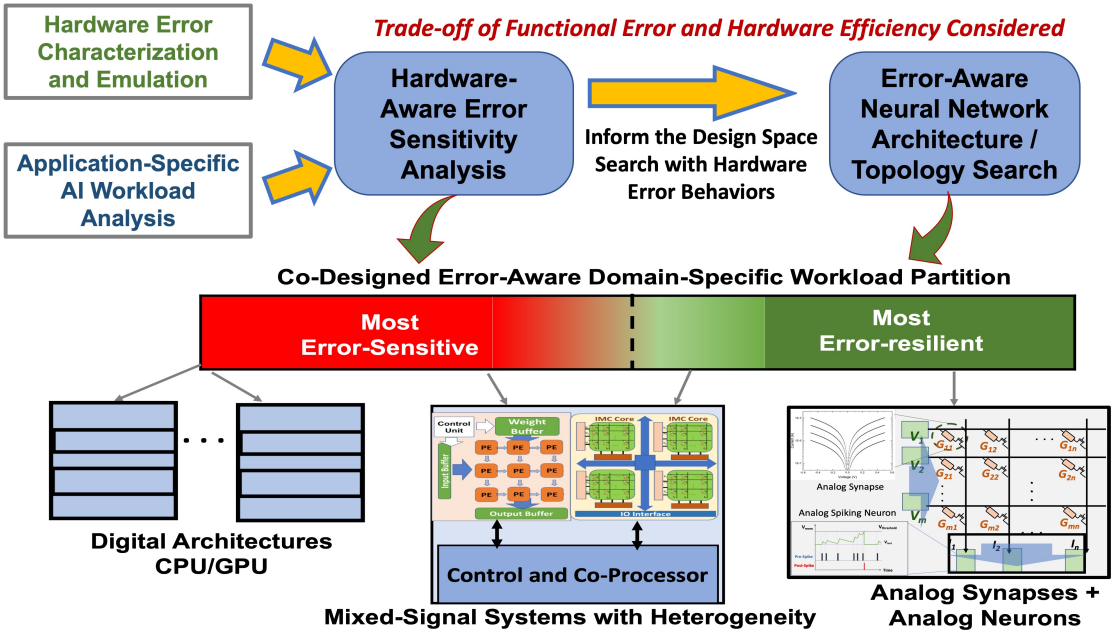



Figure 1: Overview of the error-aware framework for developing scalable neuro-computing

## References


[1] Indranil Chakraborty et al. "Geniex: A generalized approach to emulating non-ideality in memristive xbars using neural networks". In: *ACM/IEEE DAC*. 2020.

[2] Xing Chen et al. "Forecasting the outcome of spintronic experiments with neural ordinary differential equations". In: *Nature communications* 13.1 (2022), p. 1016.

[3] Christian Etmann, Sebastian Lunz, Peter Maass, and Carola-Bibiane Schönlieb. "On the connection between adversarial robustness and saliency map interpretability". In: *arXiv:1905.04172* (2019).

[4] Miao Hu et al. "Dot-product engine for neuromorphic computing: programming 1T1M crossbar to accelerate matrix-vector multiplication". In: *2016 ACM/IEEE DAC*.

[5] Seungchul Jung et al. "A crossbar array of magnetoresistive memory devices for in-memory computing". In: *Nature* 601.7892 (2022), pp. 211–216.

[6] Hyungjin Kim, MR Mahmoodi, H Nili, and Dmitri B Strukov. "4K-memristor analog-grade passive crossbar circuit". In: *Nature communications* 12.1 (2021), p. 5198.

[7] Priyadarshini Panda. "QUANOS: adversarial noise sensitivity driven hybrid quantization of neural networks". In: *Proceedings of the ACM/IEEE ISLPED*. ACM, 2020.

[8] Sourjya Roy, Cheng Wang, and Anand Raghunathan. *Evaluation of STT-MRAM as a Scratchpad for Training in ML Accelerators*. 2023. arXiv: `2308.02024 [cs.AR]`.

[9] T. Sharma, C. Wang, A. Agrawal, and K. Roy. "Enabling robust SOT-MTJ crossbars for machine learning using sparsity-aware device-circuit co-design". In: *2021 ACM ISLPED*.

[10] Chongzhi Zhang et al. "Interpreting and Improving Adversarial Robustness of Deep Neural Networks With Neuron Sensitivity". In: *IEEE Trans. on Image Processing* (2021).

# Towards a Scalable Neuromorphic Domain Specific Language

Felix Wang (felwang@sandia.gov), Sandia National Laboratories

More is different, and large-scale simulation brings along many challenges that are easily overlooked at the smaller scales [1]. For scalable integration, there is a critical need to develop simulation tools to be parallelized and distributed from the ground up, rather than attempting to modify existing serial or single-node solutions directly. Considering the distributed nature of neuromorphic computing, supporting data structures and algorithms should incorporate lessons on data locality, memory efficiency, and minimization of data movement during runtime, but also remain flexible to incorporating the levels of biological detail required for algorithm exploration and co-design.

Although many frameworks, libraries, and tools currently exist in the neuromorphic computing ecosystem [2, 3, 4, 5, 6], there has not been widespread adoption of any particular approach analogous to libraries like PyTorch for deep learning [7]. This position paper encourages the development of an "industry standard" specification for defining, constructing, and simulating neural networks at scale. Ideally, this would be split between a relatively intuitive set of graph-like abstractions (coming close to being a domain specific language) for a user to frictionlessly convert ideas to code, and a relatively performant and scalable implementation (i.e. parallel, distributed, potentially streaming).

An illustrative example of this need from a recent hippocampal simulation paper [8], was that although the network (5.3 million neurons and 40 billion synapses) used the popular NEST simulator [9], the authors reported that network construction nevertheless took 240 hours and required the custom parallelization of the generation script. We envision that an effort around scalable tooling will do more than just accelerate the construction of networks, but ideally also support model sharing, modification, and debugging, bringing neuromorphic computing modeling closer in maturity to integrated SoC design. Here, usability metrics such as the ease and efficiency of programmability also become important.

Of course, this will require the consideration of many competing design specifications and trade-offs. Taking inspiration from existing tools, we offer the following desiderata: The language (or data object) used to define a network should be compressed where possible (such as in the case of procedural connections, or duplicated structures), but also allow for uncompressed components (such as in the case of user-defined arbitrary connections). One of the key programming abstractions that should be incorporated is a method to virtualize or group (sub)sets of neurons w.r.t. to their concrete instantiations, essentially providing a level of indirection that would allow for more flexible ways of constructing the connections between neurons. Another useful abstraction is to allow for hierarchical structure. Although this is often implemented at a neural population level, it should also be possible to define connections hierarchically (e.g. connecting two higher-level populations should resolve to connecting their lower-level neurons). At an implementation level, we expect these abstractions to involve multiple methods for indexing (e.g. partitioned global, local, virtual).

Beyond just the network definitions, other issues are how to handle or configure I/O, recording spikes and state information, and how to embed the network within potential experimental setups (e.g. training, testing, parameter sweeps, batch, interactive, real-time, etc.). Here, another key consideration is the ease of mapping networks onto hardware (in addition to providing a reference simulation). A major challenge of hardware is that oftentimes there will need to be modifications to the canonical or reference network, for example, when certain nodes exceed hardware limitations in fan-in/out or certain connections exceed delay limitations [10, 11]. These steps will necessarily modify both the structure and states of the reference network, potentially resulting in divergent computation. In addition to a partition-based implementation (which benefits both HPC and neuromorphic platforms), we also expect advantages to having an intermediate representation that enables reversible modification to the network as it undergoes hardware-specific repartitioning and mapping.

```
import snn                                  # import the snn dsl as a library, e.g. in Python
net1 = snn.load('net1_snapshot.snn')        # path to snapshot header (data files separate)
net1.remove(pop1, keep_conn=False)          # remove pop1 (and any related connections)
net2 = snn.network()                        # initialize an empty network
pop7 = snn.population('lif', 20)            # instantiate new populations (standalone), with
pop8 = snn.population('izhikevich', 20)     #   potentially more biologically realistic models
net2.add(pop7, pop8)                        # add the populations into the empty network
p7_out = snn.virtual([net2.pop7[-10:]])     # when setting up virtual populations, we can use
p8_in = snn.virtual([pop8[0,1,3,5,6]])      #   either the standalone or nested references
p7__p8 = snn.connect(p7_out, p8_in, snn.conn.uniform(0.1, 1)) # parameterized connection types
net2.add(p7__p8)  # add the connection to the network (this should also add p7_out and p8_in)
n2_out1 = snn.virtual([net2.pop7[:5], pop8[-5:]])  # some more virtual populations to connect,
n2_out2 = snn.virtual([net1.net2.pop8[10:15]])     #   with some more examples of indirection
n2__p2 = snn.connect(n2_out1, p2_in, snn.conn.one_to_one())  # connection sources/targets need
n2__n3 = snn.connect(n2_out2, n3_in, snn.conn.all_to_all())  #   to resolve to vertices only
net1.add(net2, n2__p2, n2__n3)              # add net2 and some connections hierarchically to net1
```

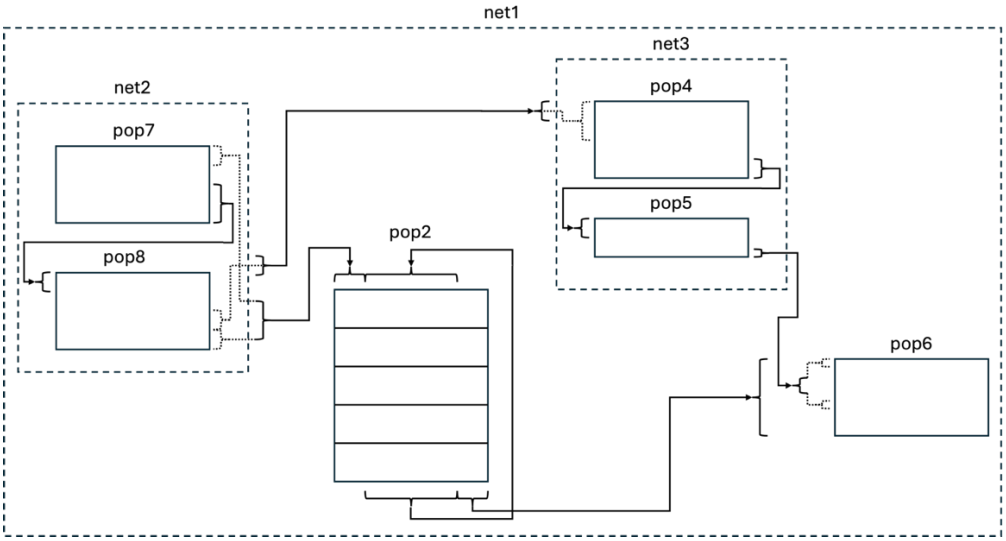



*Figure: Conceptualized example of using a neuromorphic modeling library to develop a network with hierarchically nested structure, virtual connections (dotted lines), duplicated structure and recurrent connections (population 2), complex neuron types, multiple connection types, network snapshots, and swappable components (in code).*

**References:**


[1] F. Wang, S. Kulkarni, B. Theilman, F. Rothganger, C. Schuman, S.-H. Lim, and J. B. Aimone, "Scaling neural simulations in STACS," *Neuromorphic Computing and Engineering*, vol. 4, no. 2, p. 024002, 2024.

[2] J. B. Aimone, W. Severa, and C. M. Vineyard, "Composing neural algorithms with Fugu," *in Proceedings of the International Conference on Neuromorphic Systems (ICONS)*, p. 3, 2019.

[3] Intel Labs, "Lava: A software framework for neuromorphic computing," *https://lava-nc.org*, 2024.

[4] A. Davison, D. Brüderle, J. Eppler, J. Kremkow, E. Muller, D. Pecevski, L. Perrinet, and P. Yger, "PyNN: a common interface for neuronal network simulators," *Frontiers in Neuroinformatics*, vol. 2, 2009.

[5] E. Yavuz, J. Turner, and T. Nowotny, "GeNN: a code generation framework for accelerated brain simulations. Scientific Reports 6, p. 18854, 2016.

[6] M. Stimberg, R. Brette, and D.F.M. Goodman, "Brian 2, an intuitive and efficient neural simulator," *eLife* vol. 8, p. e47314, 2019.

[7] A. Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *arXiv*, p. 1912.01703, 2019.

[8] D. Gandolfi, J. Mapelli, S.M.G. Solinas, P. Triebkorn, E. D'Angelo, V. Jirsa, and M. Migliore, "Full-scale scaffold model of the human hippocampus CA1 area," *Nature Computational Science*, pp. 1–13, 2023.

[9] M.-O. Gewaltig and M. Diesmann, "NEST (Neural Simulation Tool)," *Scholarpedia*, vol. 2, no. 4, p.1430, 2007.

[10] M. Davies *et al*., "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," *IEEE Micro*, vol. 38, no. 1, pp. 82-99, 2018.

[11] S.B. Furber et al., "Overview of the spinnaker system architecture," *IEEE Transactions on Computers*, vol. 62, no. 12, pp. 2454–2467, 2013.

# Exploiting Brain-Scale Computing Through a Memristor Based Bio-Inspired Analog Architecture

Chris Yakopcic, Shahanur Alam, and Tarek M. Taha
University of Dayton

# Exploiting Brain-Scale Computing Through a Memristor Based Bio-Inspired Analog Architecture

Chris Yakopcic, Shahanur Alam, Tarek M. Taha

**Introduction:** The traditional digital approach to computer processing leads a path to looped, bit-wise, multiply-adds, and an inevitable memory bottleneck, well before brain-scale computing is achieved. Thus, we present our memristor-based SNN neuromorphic system, and how this design has the potential to yield extreme gains in energy and compute efficiency when compared to traditional CPUs. The following are the key aspects of our system that most significantly impact of our ability to complete extreme low SWaP (Size, Weight, and Power) neuromorphic processing.

**Large Scale Analog Computation:** Synaptic density is a key contributor to overall system efficiency. However, many operations in neuromorphic applications require large sparse matrices. Thus, there is a mismatch between the concept of an ultra-high-density crossbar, and a synaptic weight matrix that contains mostly zeros. An abundance of zero multiplies results in a lot of wasted computation elements, and unnecessary analog noise. Thus, in our system we employ multilayer 3D crossbars (Fig. 1), reducing the number of elements needed in a single x-y crossbar grid, also reducing wire resistance and detrimental parasitic effects. Cross points in our 3D crossbar are staggered between layers, so input data streams can be routed to different layers to achieve desired sparsity, without sacrificing chip area and SNR.

**Biologically Accurate Analog Data Propagation:** To accurately model biological processes, it is important to have spiking neuron circuits (Figs. 2,3) be the output of the 3D memristor crossbars. Key benefits of spiking neurons are the dramatically lower communications costs and the elimination of power-hungry analog to digital converter circuits. Our capacitive accumulation approach allows for the time integration and leakage characteristics needed at low area and energy overheads. Our analysis has shown that the Izhikevich [1] neuron model would be the best model to emulate in the neuron circuits, given its flexibility and biological accuracy. We have already shown that reworking traditional algorithms in spiking form leads to great gains in efficiency [2], and this flexible neuron allows for many different algorithm-to-neuron mapping options.

**Spike Time Driven On-Chip Learning:** On-chip learning is an inevitable necessity of neuromorphic hardware. Based on our previous work in this area [3], we use analog processes to transmit weight update amounts to the correct layers. Non-binary weight updates are achieved by scaling both pulse width and voltage magnitude for high resolution trainability. Given that system uses spiking neurons exclusively, STDP is used to generate weight update amounts. The use of STDP provides us with simplicity and efficiency, compared to traditional back propagation, where we would be required to store high resolution values for both input and outputs of the activation function to determine gradients for weight updates [3].

**Physically Accurate Large-Scale Circuit Simulation:** Based on our previous work in physical modeling of memristor based neuromorphic systems [4], we developed a multilevel approach for simulating this neuromorphic system at a large scale. This involves physical modeling at the device level for the memristor arrays (Fig. 4). We use SPICE to evaluate bio-inspired neurons (Fig. 3), and MATLAB scripts that capture wire resistance and multilayer data propagation. We also use a software-in-the-loop approach [3] where MATLAB initiates SPICE calls so that intricacies of spiking neurons can be captured over the course of larger input data streams. Fast simulations of larger memristor arrays, we developed a fast simulation approach that can reduce the time to simulate a 256×256 crossbar circuit in SPICE from over 10 hours down to less than 2s [5] with 99.9% accuracy (Fig 5).

**System Simulation Using Loihi Array:** While we have future plans to translate our neuromorphic system to silicon, analog chip design is extremely costly and time consuming. Prior to designing the full analog circuits, we are working on simulating a large collection of neuronal circuits modeling brain activity. Our extensive experience [2,6] with the Intel Loihi processor has convinced us that this is a great platform for such modeling. The Loihi can model a large collection of neuron types, supports STDP learning, and can easily scale to a biological level through a large collection of Loihi chips.

**Conclusion:** In the future we plan to demonstrate through simulation, that our proposed neuromorphic system is capable of extreme low SWaP processing with real time input data streams. This will allow us to perform large scale benchmarking and system comparison, quantifiably demonstrating the necessity of a brain-scale system.

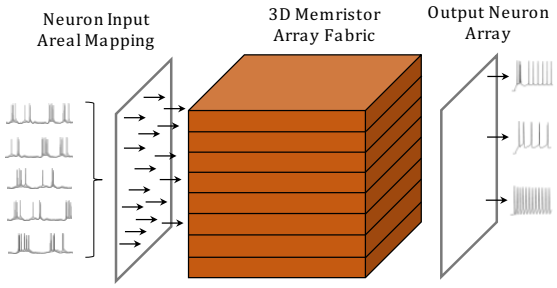



Fig. 1. 3D memristor array structure provides more input mapping flexibility, and is a better match for sparse weight matrices. Memristor array volume allows for a reduction of cross points per layer without compromising synaptic density, and reduces sneak currents and parasitic effects.

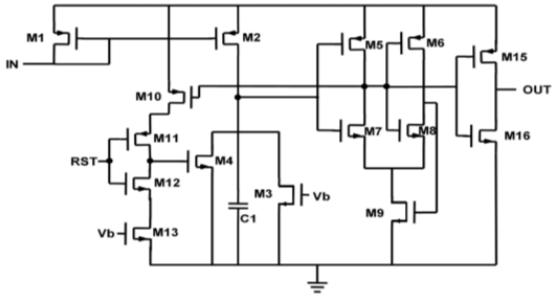



Fig. 2. Analog neuron analyzed in [7], proposed in [8].

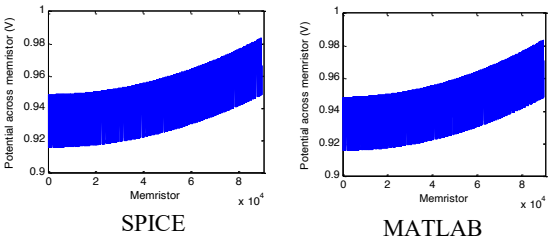



Fig. 5. Voltage potential across all memristors in a 300 by 300 crossbar, obtained through SPICE simulations, and the proposed MATLAB framework based simulations, with virtually a perfect match.

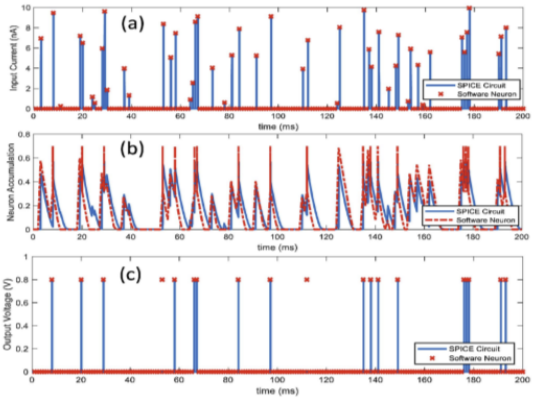



Fig. 3. Simulation of analog neuron in Fig. 2, small scale SPICE simulation, and large scale MATLAB simulation [7].

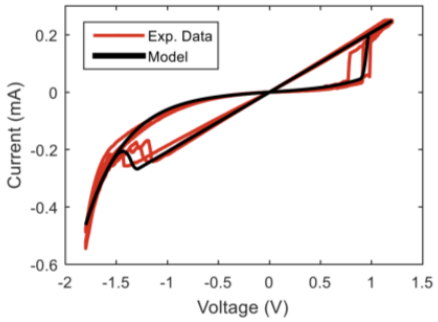



Fig. 4. Example of memristor model [4] capable of automatically deriving an IV curve based on input characterization data, where model is an average representation of multiple input IV sweeps.

## References


[1] E. M. Izhikevich, “Simple Model of Spiking Neurons,” IEEE Transactions on Neural Networks, 14:1569- 1572, 2003.

[2] C. Yakopcic, N. Rahman, T. Atahary, T. M. Taha and S. Douglass, “Leveraging the Manycore Architecture of the Loihi Spiking Processor to Perform Quasi-Complete Constraint Satisfaction,” International Joint Conference on Neural Networks, Glasgow, UK (Virtual), July, 2020.

[3] C. Yakopcic, T. M. Taha, and M. R. McLean, “Method for ex-situ training in a memristor-based neuromorphic circuit using a robust weight programming method,” Electronics Letters, vol. 51, no. 12, pp. 899-900, June, 2015.

[4] C. Yakopcic, T. M. Taha, D. J. Mountain, T. Salter, M. J. Marinella, M. McLean, “Memristor Model Optimization Based on Parameter Extraction from Device Characterization Data,” IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 39, no. 5, pp. 1804-1095, May, 2020.

[5] N. Rahman, C. Yakopcic, R. Lent, J. C. Briones, D. Chelmins, R. Dudukovitch, A. Smith, A. Gannon, M. Lowry, M. S. Murbach, A. J. Salas, T. M. Taha, “Neuromorphic Hardware in Outer Space: Software Defined Networking Executed on an In-Orbit Loihi Spiking Processor,” IEEE Cognitive Communications for Aerospace Applications Workshop (CCAAW), Cleveland, OH, June, 2023.

[6] Raqibul Hasan, Tarek M Taha, Chris Yakopcic , “A Fast Training Method For Memristor Crossbar Based Multi-Layer Neural Networks,” Analog Integrated Circuits and Signal Processing, Analog Integrated Circuits and Signal Processing, Volume 93, Issue 3, Pages 443 – 454, December 2017.

[7] A. Henderson, C. Yakopcic, C. Merkel, H. Hazan, S. Harbour, T. M. Taha, “Memristor Based Liquid State Machine with Method for In-Situ Training,” IEEE Transactions on Nanotechnology (Accepted, 2024).

[8] J. Shamsi, K. Mohammadi and S. B. Shokouhi, "A low power circuit of a leaky integrate and fire neuron with global reset", 2017 Iranian Conference on Electrical Engineering (ICEE), pp. 366-369, 2017.

# The forgotten chemical connectome: how to implement a neuromodulatory system and why should we care about it

Angel Yanguas-Gil (Email: ayg@anl.gov), Jeffrey W. Elam, Argonne National Laboratory

## Why neuromodulation and why it matters

The costs of computing and data movement in biological systems explain many of the attributes of the brain: 1) local computing is cheap, done primarily via chemistry: this leads to complex stateful neurons and gap junctions for local computing; 2) moving data far and fast is expensive: this leads to spiking neurons in larger brains; 3) broadcasting of signals is cheap, using passive diffusion or relying on the circulatory system: this leads to neuromodulators. The connectionist approach is based on constraints #1 and #2. In contrast, #3 has been comparatively ignored.

Neuromodulation is a key building block of the brain that enables unique functionality. The so called chemical connectome is pervasive across species. From a functional perspective, one of the key enabling features is providing a conceptual framework to build top-down mediated multifunctional networks, where the specific functionality is modulated based on a global context. Recently, we adapted neuromodulatory design principles found in the olfactive system of insects to design compact, multifunctional networks whose functionality is dictated by top-down inputs (Figure 1). Network size was further reduced by implementing deep task learning, where a larger number of tasks is projected into a smaller dimensional space of neuromodulators. This further provides a simple way of building complex recurrent systems on top of these networks to process complex data streams via coupling to state machines.

## The challenge of implementing modulatory systems in neuromorphic systems

There are essentially three approaches to implement modulatory systems in hardware:

The first option is expanding current **spike routing** approaches in digital implementations. While currently we are exploring AI testbeds at ALCF to test the mapping of neuromorphic architectures to highly distributed systems, this approach is likely to lead to slow systems performing at sub MHz range, similar to neuromorphic chips such as Loihi or TrueNorth.

A second option is to leverage existing **microfluidics** approaches to use molecules as neuromodulators. Setting aside the challenge of having reservoirs, the need to find orthogonal chemistries to achieve the desired selectivity and the right surface chemistry to control the residence time of adsorbed species, preliminary calculations of transport of molecules inside microfluidic channels lead to rates that are comparable to those of biological systems (i.e. kHz range). Some of the surface chemistry challenges can be addressed using techniques such as atomic layer deposition, which would allow to design surface with the desired reactivity, as shown in our prior research.

A third route is to use other physical channels, such as **EM fields**. In the case of photonic devices, two key differences between this and prior approaches is that modulatory systems do not require point-to-point connectivity nor require the ability to operate with multiple wavelengths. We can address sensing selectivity through the use of photonic structures tailored to act as selective reflectors or absorbers, as shown in Figure 2. As before, we can use high precision manufacturing techniques such as ALD to create patterned multilayered structures with the right optical properties or to functionalize already patterned structures to control selectivity. This is something we have done in the past in the context of selective solar absorbers.

## Conclusions

Modulatory systems can greatly expand the capabilities of existing neural networks. Regardless of the proposed approach, there are still fundamental challenges that need to be overcome from a computing, simulation, and fabrication perspectives.. Finally, we need a deeper understanding of the fundamentals of modulation and their role in the brain.

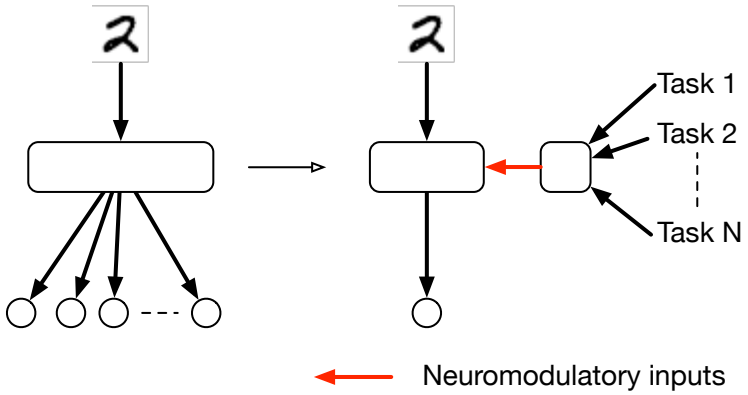



Figure 1: Expanding conventional neural networks with neuromodulatory interactions provides a straightforward approach to the design of compact networks capable of doing multiple tasks. Application of this method to classification tasks (MNIST in this example) leads to networks capable of identifying any digit depending on the top-down context that are 60% smaller than the original networks without any loss of accuracy. We have used these networks as building blocks of more complex tasks, mimicking how neuromodulators drive the antennal lobe of insects towards specific tasks based on top-down contextual information

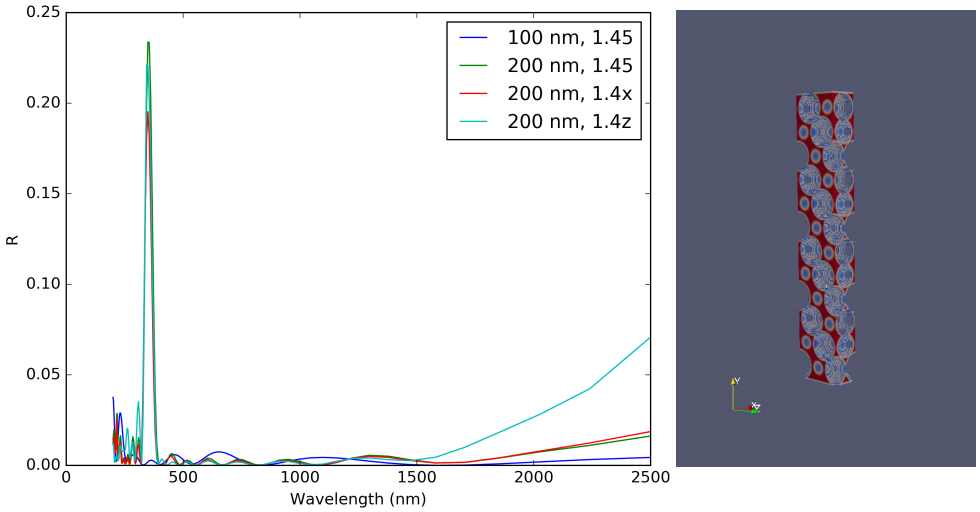



Figure 2: Using physics simulation tools such as meep and synthesis methods such as atomic layer deposition we can design photonic layers with highly tailored optical properties. These can be used to provide the selectivity required to fabricate modulatory systems based on photonic architectures. A key difference with neuromorphic photonic approaches is that information is broadcasted across large areas, circumventing limitations related to the large size of photonic elements, since a single photodetector can serve a large section of a chip.

# Brain-Derived Neuromorphic Computing with 3D Photonic-Electronic-Ionic Circuits

***Author:*** *S. J. Ben Yoo; email: sbyoo@ucdavis.edu*
***Organizations:*** *University of California, Davis*
***Position Paper:*** *ASCR Workshop on Neuromorphic Computing for Science 2024*
***Topics:*** *Translation to analog microelectronic circuits, Neuroscience-inspired computing principles, Modeling and simulation approaches, Performance metrics, data requirements, and energy efficiency*

## 1. Challenges

For many decades, there have been world-wide efforts to design and realize a brain-like flexible learning system of similar capability, comparable power consumption, and compact size as the human brain. While the software and algorithms for machine learning (ML) and artificial intelligence (AI) have advanced remarkably, the actual ML and AI hardware systems lagged significantly compared to the brain in terms of its flexible learning capability and its size, weight, and power. The human brain, in contrast, is capable of remarkably fast learning in a manner that is flexible and enables generalization to new situations and tasks, and it does so with a remarkably low level of energy consumption relative to traditional computational hardware. To address the failures of previous efforts in reverse-engineering the brain, we identify four fundamental scientific and technological "Gaps" summarized in **Table 1**[1], [2].

## 2. Four Gaps, Four Hypotheses, and Four Objectives in Neuromorphic Computing

**Table 1.** Possible four 'Gaps' commonly seen in the previous neuromorphic computing research activities, and the corresponding *Hypotheses* and the *Objectives* of the Brain-derived methods for neuromorphic computing .

| |
|---|
| ***Gap 1: Lack of understanding of the principles of learning, plasticity, and dynamics in the context of networked neurons with structured connectivity*** |
| ***Hypothesis 1:*** Human-level intelligence emerges through dynamic networked interactions between multiple specialized brain systems with structured and efficient large-scale hardware connectivity. |
| ***Objective 1:*** Develop a comprehensive simulator and build a novel neuromorphic computing prototype system that incorporates insights from cutting-edge modeling and experiments about synaptic plasticity, network dynamics, and learning in cortical circuits, and fundamental attributes of human learning and memory. In reverse, utilize the prototype system to understand the brain. |
| ***Gap 2: Lack of methods to realize neuromorphic dynamics in bio-derived materials*** |
| ***Hypothesis 2:*** Conventional electronic materials such as silicon are unable to faithfully replicate the dynamicity driven by ions, molecules, and structural changes in the dendrites, synapses, and somas. |
| ***Objective 2:*** Pursue new photonic, electronic, and ionic memristive materials that can closely resemble the dynamic mechanisms responsible in the biological neural systems. |
| ***Gap 3: Lack of methods to realize brain-derived neuromorphic devices*** |
| ***Hypothesis 3:*** Conventional electronic devices such as CMOS transistors and electrical wires are unable to faithfully replicate the dynamicity seen in the dendrites, synapses, and somas. |
| ***Objective 3:*** Pursue new photonic, electronic, and ionic memristive devices that can closely resemble the dynamic mechanisms seen in the biological neural systems. |
| ***Gap 4: Lack of scalable and energy-efficient interconnecting circuits for brain-like hierarchical learning*** |
| ***Hypothesis 4:*** Current (analog) electronic approaches are unable to achieve the connectivity (e.g. ~8000 synaptic connections per neuron) at scale (e.g. billions of neurons) limited by electronic wirings. |
| ***Objective 4:*** Pursue 3D photonic-electronic integrated circuits that offer high density and high connectivity with extreme efficiency at scale while supporting hierarchical learning in optical macro-circuits and electronic micro-circuits. We will conduct simulation and experimental testbed studies. |

## 3. Brain-Derived Neuromorphic Computing with 3D photonic-electronic-ionic integrated circuits

We propose to pursue Brain-Derived rather than Brain-Inspired neuromorphic computing that addresses the four Gaps of **Table 1** and exploiting the new direction depicted in Figure 1. The new 3D Nanoscale Photonic-Electronic-Ionic neuromorphic computing pursues new material, device, circuit, and system capabilities of co-designed 3D photonic and electronic integrated circuits (3D EPICs) designed for hierarchical learning. As Figure 2 illustrates, the proposed 3D nanoscale photonic-electronic integrated circuits for hierarchical neuromorphic computing consisting of photonic neuromorphic computing circuits and electronic/ionic neuromorphic integrated circuits [3], [4].

## 4. Timeliness

The popularity of the deep learning AI systems is currently driving AI-related energy-consumptions to double every 3.4 months [6] . Comprehensive research on Brain-Derived neuromorphic computing research covering all topics of the Workshop should start now.

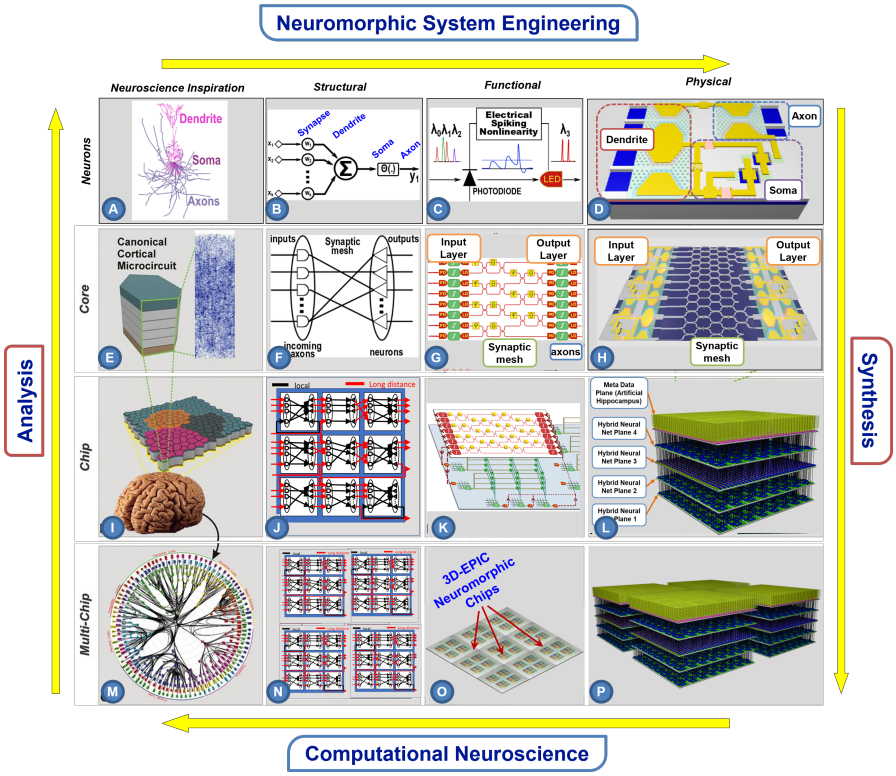



Figure 1. A conceptual 3D Hierarchical Neuromorphic Nanocomputing architecture extending on **the framework by** [7]**. (A)** A canonical neuron, **(B)** Neuron's minimal structure, **(C)** Neuron's simplified functional diagram (optoelectronic neuron example), **(D)** A physical schematic for a nanoscale optoelectronic neuron, **(E)** cortical microcircuit, **(F)** Structure of a neurosynaptic core with axons as signal carriers (inputs/output), synapses as directed connection strength, and neurons as nonlinearity. **(G)** Functional view of a photonic synaptic mesh between presynaptic and post synaptic neurons. **(H)** Physical layout of (G). **(I)** A two-dimensional map of cortical columns in a functional network. Multichip scales are both created by interconnecting **(J)** neuron microcircuits reconfigurable optical synaptic interconnects. **(K)** Hybrid optical (red) and electronic (green) neural network forming a hierarchical macrocircuit. **(L)** Schematic of 3D electronic photonic integrated circuit (EPIC) neural network consisting of multiple planes of (K). **(M)** Illustration of long-range connections between cortical regions in the macaque brain [8]. **(N)** Interconnections of many functionally specialized neural macro-circuits (J). **(O)** Multi-3D EPIC chip neural networks emulating functional specializations of interconnected human brain structures. **(P)** Schematic of (O).

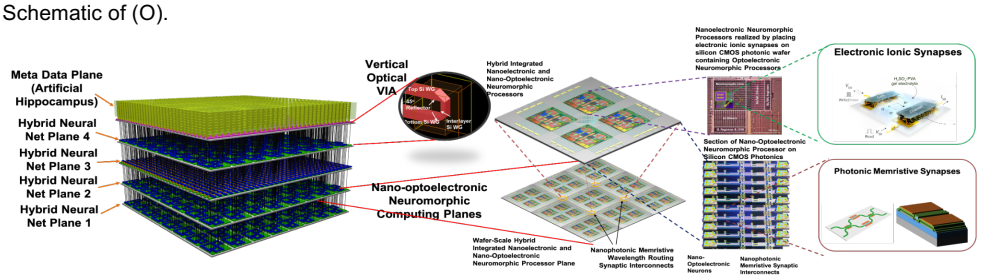



Figure 2. The proposed 3D nanoscale photonic-electronic integrated circuits for hierarchical neuromorphic computing consisting of photonic neuromorphic computing circuits and electronic/ionic neuromorphic integrated circuits.

## 5. References


[1] S. J. Ben Yoo *et al.*, "Towards Reverse-Engineering the Brain: Brain-Derived Neuromorphic Computing Approach with Photonic, Electronic, and Ionic Dynamicity in 3D integrated circuits", Available: arXiv:2403.19724

[2] S. J. Ben Yoo, "Brain-Derived 3D NanoPhotonic-NanoElectronic Neuromorphic Computing," in *2022 IEEE Photonics Conference (IPC)*, IEEE, Nov. 2022, pp. 1–2. doi: 10.1109/IPC53466.2022.9975516.

[3] W. Wan *et al.*, "A Voltage-Mode Sensing Scheme with Differential-Row Weight Mapping for Energy-Efficient RRAM-Based In-Memory Computing," in *2020 IEEE Symposium on VLSI Technology*, 2020, pp. 1–2.

[4] Y. van de Burgt *et al.*, "A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing," *Nat Mater*, vol. 16, no. 4, pp. 414–418, 2017, doi: 10.1038/nmat4856.

[5] D. Amodei, D. Hernandez, G. Sastry, J. Clark, G. Brockman, and I. Sutskever, "AI and Compute," Open AI.

[6] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for modern deep learning research," in *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, 2020. doi: 10.1609/aaai.v34i09.7123.

[7] P. A. Merolla *et al.*, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science (1979)*, vol. 345, no. 6197, pp. 668–673, Aug. 2014, doi: 10.1126/science.1254642.

[8] D. S. Modha and R. Singh, "Network architecture of the long-distance pathways in the macaque brain," *Proceedings of the National Academy of Sciences*, vol. 107, no. 30, pp. 13485 LP – 13490, Jul. 2010, doi: 10.1073/pnas.1008054107.

# Deep Co-design, with Cross-discipline Teams, Leveraging State of the Art Technology to Push Neuromorphic Computing Forward

Aaron R. Young
Oak Ridge National Laboratory, Oak Ridge, TN, USA
youngar@ornl.gov

Topic: Modeling and simulation approaches.

## 1 Challenge

Through rigorous study of biology, we know that it is possible to perform very complex human-capable tasks, which computers have traditionally struggled with, using very low power with a complex biological system, i.e. a brain. Estimates state that the brain uses around 10 W of energy and operates at around 10 Hz, with billions of neurons, each having thousands of connections resulting in trillions of synapses. This level of energy efficiency, performance, and scale of elements is currently out of reach; however, the target is in front of us.

Some of the key challenges we face when trying to replicate the success of the brain are as follows: 1) We do not fully understand the underlying biological features contributing to the brain's capabilities. Specifically, we do not know if a particular feature observed in biological systems is key to its capabilities or if it is a byproduct of the electrochemical/biological medium. Therefore, we don't know if the feature is needed to replicate the capability in microelectronics. 2) Researchers have developed novel devices that exhibit complex brain-like spiking behavior; however, There remains a large research gap between designing an interesting device and building a neuromorphic architecture capable of leveraging the device's characteristics in a complete system. 3) Scaling the simulations and models of complex microelectronics to biologically realistic scales and complexities is computationally prohibitive on existing HPC systems.

## 2 Opportunity

With these great challenges, there are also tremendous opportunities to work together to expand our technological capabilities and our understanding of the brain. 1) Deep co-design in the technology stack from devices, circuits, architectures, software, algorithms, and applications, along with close interdisciplinary collaboration with expert neuroscientists, is needed to develop a completely feasible solution and to understand the intricacies of the brain's operation [1]. 2) New heterogeneous and specialized computing architectures are needed both for the neuromorphic computing systems under development but also as evaluation and simulation systems capable of modeling the design space. The boon that AI/ML has enjoyed recently is powered by the widely available processing capabilities present with GPUs and the open-source development of software toolchains. Spiking-based neuromorphic systems will need a similarly well-suited hardware accelerator able to evaluate and model novel analog neuromorphic circuits and a more open collaboration effort to accelerate the development.

## 3 Timeliness

Already, the neuromorphic systems and algorithms currently being researched that leverage simplistic neuron models and traditional digital circuitry [2]–[4] are capable of matching or exceeding traditional approaches in terms of application performance, power, and area usage [5]–[7]. By integrating novel devices, a more modern understanding of the brain, emerging architecture designs [8], development platforms [9], improved spike-based communication [10]–[13], and packaging techniques [14], neuromorphic computing systems will continue to improve.

## References


[1] Jeffrey S. Vetter, Prasanna Date, Farah Fahim, et al. "Abisko: Deep codesign of an architecture for spiking neural networks using novel neuromorphic materials". In: *The International Journal of High Performance Computing Applications* 37.3-4 (2023), pp. 351–379. DOI: 10.1177/10943420231178537. eprint: https://doi.org/10.1177/10943420231178537. URL: https://doi.org/10.1177/10943420231178537.

[2] J. Parker Mitchell, Catherine D. Schuman, Robert M. Patton, et al. "Caspian: A Neuromorphic Development Platform". In: *Proceedings of the Neuro-Inspired Computational Elements Workshop*. NICE '20. Heidelberg, Germany: Association for Computing Machinery, 2020. ISBN: 9781450377188. DOI: 10.1145/3381755.3381764. URL: https://doi.org/10.1145/3381755.3381764.

[3] J. Parker Mitchell, Catherine D. Schuman, and Thomas E. Potok. "A Small, Low Cost Event-Driven Architecture for Spiking Neural Networks on FPGAs". In: *International Conference on Neuromorphic Systems 2020*. ICONS 2020. Oak Ridge, TN, USA: Association for Computing Machinery, 2020. ISBN: 9781450388511. DOI: 10.1145/3407197.3407216. URL: https://doi.org/10.1145/3407197.3407216.

[4] Narasinga Rao Miniskar, Aaron Young, Kazi Asifuzzaman, et al. "Neuro-Spark: A Submicrosecond Spiking Neural Networks Architecture for In-Sensor Filtering". In: *International Conference on Neuromorphic Systems 2024*.

[5] Shruti R. Kulkarni, Aaron Young, Prasanna Date, et al. "On-Sensor Data Filtering Using Neuromorphic Computing for High Energy Physics Experiments". In: *Proceedings of the 2023 International Conference on Neuromorphic Systems*. ICONS '23. Santa Fe, NM, USA: Association for Computing Machinery, 2023. ISBN: 9798400701757. DOI: 10.1145/3589737.3605976. URL: https://doi.org/10.1145/3589737.3605976.

[6] James Ghawaly, Aaron Young, Dan Archer, et al. "A Neuromorphic Algorithm for Radiation Anomaly Detection". In: *Proceedings of the International Conference on Neuromorphic Systems 2022*. ICONS '22. Knoxville, TN, USA: Association for Computing Machinery, 2022. ISBN: 9781450397896. DOI: 10.1145/3546790.3546815. URL: https://doi.org/10.1145/3546790.3546815.

[7] James Ghawaly, Aaron Young, Andrew Nicholson, et al. "Performance Optimization Study of the Neuromorphic Radiation Anomaly Detector". In: *Proceedings of the 2023 International Conference on Neuromorphic Systems*. ICONS '23. Santa Fe, NM, USA: Association for Computing Machinery, 2023. ISBN: 9798400701757. DOI: 10.1145/3589737.3605980. URL: https://doi.org/10.1145/3589737.3605980.

[8] Aaron R. Young, Jeffrey S. Vetter, Frank Liu, et al. "Emerging Heterogeneous Systems Provide Great Opportunities for Codesign". In: *ASCR Workshop on Reimagining Codesign*. 2021.

[9] Brett Witherspoon and Aaron Young. "Event-Driven Sensing and Embedded Neuromorphic Platforms for Gamma Radiation Monitoring". In: *Proceedings of the Great Lakes Symposium on VLSI 2024*. GLSVLSI '24. Clearwater, FL, USA: Association for Computing Machinery, 2024, pp. 779–784. ISBN: 9798400706059. DOI: 10.1145/3649476.3660363. URL: https://doi.org/10.1145/3649476.3660363.

[10] Aaron Reed Young. "SNACC: The Scaled-up Neuromorphic Array Communications Controller". PhD thesis. University of Tennessee, May 2020. URL: https://trace.tennessee.edu/utk_graddiss/5843/.

[11] Aaron Reed Young. "Scalable High-Speed Communications for Neuromorphic Systems". MA thesis. University of Tennessee, 2017.

[12] A. R. Young, A. Z. Foshie, M. E. Dean, et al. "Scaled-up Neuromorphic Array Communications Controller (SNACC) for Large-scale Neural Networks". In: *2020 International Joint Conference on Neural Networks (IJCNN)*. July 2020, pp. 1–8. DOI: 10.1109/IJCNN48605.2020.9206920.

[13] A. R. Young, M. E. Dean, J. S. Plank, et al. "Neuromorphic Array Communications Controller to Support Large-Scale Neural Networks". In: *IJCNN: The International Joint Conference on Neural Networks*. Rio de Janeiro, Brazil, July 2018.

[14] Narasinga Rao Miniskar, Pruek Vanna-Iampikul, Aaron Young, et al. "A 3D Implementation of Convolutional Neural Network for Fast Inference". In: *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*. 2023, pp. 1–5. DOI: 10.1109/ISCAS46773.2023.10181622.

# Livewired Neuromorphic Systems: Where Evolving SNNs Meet Evolvable Hardware

Ramtin Zand, Assistant Professor of Computer Science and Engineering, University of South Carolina

## Neuroscientific Basis of Proposed Research

The brain's remarkable ability to adapt and learn in response to experiences and environmental changes can be encapsulated in the concept of the "Livewired" brain, extensively studied by Stanford neuroscientist David Eagleman [1]. A key aspect of the livewired brain is its continuous rewiring and the constant competition for the brain's cortical territory among sensory systems sending information to the brain. Since the biological neurons in the brain typically have similar underlying dynamics, the cortical territories for various sensory inputs are interchangeable [1]. The brain allocates its resources based on importance and creates a competitive environment among its regions [2, 3]. This for example explains why, when a person's hand is amputated, the brain's cortical territory previously dedicated to the hand representation is taken over by neighboring face and upper arm territories (See Figure 1). The distribution of cortical real estate and continuous rewiring are key principles of brain functionality and capability that are less researched in the neuromorphic community. Therefore, translating these principles into neuromorphic systems could potentially lead to significant advancements in performance, robustness, and adaptability.

## From Livewired Brain to Livewired Neuromorphic Systems

This paper introduces *Livewired Neuromorphic Systems* by combining evolving spiking neural networks (eSNNs) with evolvable hardware. The key feature of eSNNs is the online learning of patterns by evolving their network structure [4]. For each new input, a new neuron is dynamically allocated and connected to the input neurons. The weights of the connections between the sensory input neurons and the new neuron are determined based on the rank-order rule [5], enabling fast, one-pass learning of the input patterns in both supervised and unsupervised modes [6-8]. In a unimodal perception mode, the system continues evolving until the entire hardware real estate is utilized. When all hardware resources are taken over, new input causes neurons with similar weighted connections to merge, freeing up space for new neurons. In the multi-modal perception mode, like a livewired brain, there is a do-or-die competition between different hardware territories processing different modalities, involving an ongoing cycle of neuron creation and pruning. Deploying eSNNs for real-world applications requires evolvable neuromorphic hardware with runtime reconfiguration and rerouting capabilities. To realize the livewired neuromorphic system, a novel reconfigurable and evolvable analog hardware is necessary. This hardware can leverage analog memristive synapses and neurons [9-14] interconnected through a highly flexible analog network-on-chip to support rewiring and evolution required in livewired evolving SNNs.

## Call for Action and Future Research Directions

The *Livewired Neuromorphic Systems* have the potential to achieve adaptability and robustness on an unprecedented scale. These systems necessitate innovative software-hardware co-design approaches and cross-layer collaborations to address critical research questions, including but not limited to the following:

1. What are the best strategies for dynamically allocating and connecting new neurons to ensure effective learning?
2. How do different input patterns affect the evolution and structure of the livewired neuromorphic systems?
3. How can the runtime reconfiguration capabilities of analog hardware be enhanced to match the dynamic needs of eSNNs?
4. What theoretical models can be developed to predict the behavior and performance of livewired neuromorphic systems?
5. How can the effectiveness of the proposed system in various real-world scenarios be empirically validated?

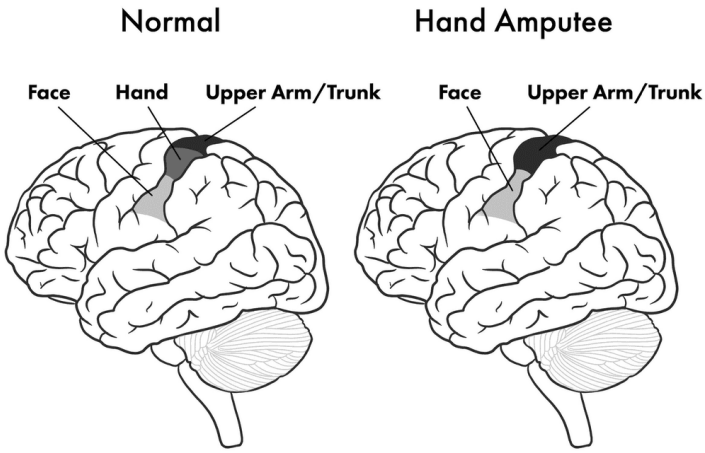



Figure 1: When a hand is amputated, the region of the brain dedicated to hand is taken over by the adjacent cortical territories. *Figure adapted from [1]*.

## References


[1] D. Eagleman, "Livewired: The inside story of the ever-changing brain," *Canongate Books*, 2020.

[2] N. Boddaert, et al., "Autism: functional brain mapping of exceptional calendar capacity," *The British Journal of Psychiatry*, 187.1 (2005): 83-86.

[3] J. LeBlanc, and M. Fagiolini, "Autism: a "critical period" disorder?," *Neural plasticity* 2011.1 (2011): 921680.

[4] S. Schliebs, and N. Kasabov, "Evolving spiking neural network—a survey," *Evolving Systems* 4 (2013): 87-98.

[5] S. Thorpe, and J. Gautrais, "Rank order coding," *Computational Neuroscience: Trends in Research*, 1998. Boston, MA: Springer US, 1998. 113-118.

[6] J. L. Lobo, et al., "Evolving spiking neural networks for online learning over drifting data streams," *Neural Networks* 108 (2018): 1-19.

[7] K. Dhoble, et al., "Online spatio-temporal pattern recognition with evolving spiking neural networks utilising address event representation, rank order, and temporal spike learning," *The 2012 international joint conference on Neural networks (IJCNN)*, 2012.

[8] N. Kasabov, et al., "Dynamic evolving spiking neural networks for on-line spatio-and spectro-temporal pattern recognition," *Neural Networks* 41 (2013): 188-201.

[9] R. Yang, H. M. Huang, and X. Guo, "Memristive synapses and neurons for bioinspired computing," *Advanced Electronic Materials* 5.9 (2019): 1900287.

[10] A. Serb, et al., "Memristive synapses connect brain and silicon spiking neurons," *Scientific reports* 10.1 (2020): 2590.

[11] S. Saïghi, et al., "Plasticity in memristive devices for spiking neural networks," *Frontiers in neuroscience* 9 (2015): 51.

[12] J. Li, et al., "Emerging memristive artificial neuron and synapse devices for the neuromorphic electronics era," Nanoscale horizons (2023).

[13] M. H. Amin, M. Elbtity, M. Mohammadi, and **R. Zand**, "MRAM-based analog sigmoid function for in-memory computing." In *Proceedings of the Great Lakes Symposium on VLSI*, 2022 (pp. 319-323).

[14] V. Ostwal, **R. Zand**, R. F. DeMara, and J. Appenzeller, "A novel compound synapse using probabilistic spin–orbit-torque switching for MTJ-based deep neural networks," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, 5(2), 182-187, 2019.

# Part 3: Pre-Workshop Report

# DOE Basic Research Needs for Neuromorphic Computing

# Pre-workshop report

## 1. Why the need for neuromorphic computing?

A major grand challenge for our time is to rethink computing in an efficient, intelligent, and extremely scalable way. Current Artificial Intelligence/Machine Learning (AI/ML) technologies suffer from extreme brittleness when confronted with situations outside their training distributions, reflecting an inability to learn causal world models that generalize the underlying cause and effect relationships to novel situations. Can we create a fundamentally new computing paradigm inspired by the brain's structure, learning capabilities, robustness and extreme energy efficiency? Engineering a neuromorphic computing architecture with similar learning abilities, robustness and extreme energy efficiency as the brain is one of the most exciting and difficult scientific endeavors of our time. Neuromorphic computing has the potential to transform everything from large scale scientific calculations, such as climate simulations, to revolutionizing how intelligent agents and sensors are integrated into our day to day lives. It could provide the capability for human level intelligence at minimal energy cost for instant ON systems that can be carried in the pocket, transforming robotics, healthcare, autonomous vehicles, security, environmental monitoring to name a few.

A neuromorphic system which can process and intelligently interpret data at extreme scales will require novel hardware fabrics that are efficient and scalable. We believe that analog hardware with capability for reconfigurability and modularity can provide support for parallelizing computing operations and resource allocations depending on the workloads, as well as supporting new features in security, fault tolerance, etc. Analog neuromorphic circuits organized in flexible hierarchical architectures inspired by brain organization could support this endeavor and deliver high temporal or/and high spatial processing resolution. Such analog circuits would have to operate the components under extremely low currents, small voltages, and use physical principles of their devices to directly perform complex computation efficiently in hardware. The metrics necessary to quantify performance at extreme scale for neuromorphic computing are also an open issue.

However, developing neuromorphic computing systems is rendered very difficult by our incomplete understanding of the computational mechanisms used by the brain. The challenge lies in the sheer number of neurons, neuron types, synapses, neuromodulators and connectivity patterns of biological neural systems, in addition to non-neural components, such as neuroglia. How much of this biological detail must be replicated to create the complex, diverse behaviors of the brain? The brain relies on biological resonant analog circuits that use electro-chemical stimuli to receive and interpret the world we live in. We believe that replicating the computational mechanisms employed by the brain will require fundamentally rethinking how to create efficient analog neuromorphic circuits. These neuromorphic circuits should capture the key aspects of the biological functionality, such as adaptive dynamics exhibited by biological neurons, synapses and circuits. Moreover, extensive simulation on high-performance computers will be required to understand how these analog circuits can be developed and improved at a fundamental level, and how they can be effectively combined into scalable systems. Such simulations are essential tools for advancing current research into biological processing mechanisms. Achieving a new computing paradigm based on the brain itself will require a massive co-design effort in which neuromorphic algorithms, simulations and hardware evolve in parallel. The final report will seek to lay out fundamental basic needs, research questions and a roadmap for neuromorphic computing.

The final report will be developed in connection with a workshop that aims to bring together interdisciplinary experts to discuss and draft a set of basic research challenges and next steps for rethinking the field of neuromorphic computing. The workshop will be centered around three key areas of exploration: first, exploration of key neuromorphic computing primitives drawing from recent insights from neuroscience regarding brain function; second, path-breaking research into novel technologies for prototyping neuromorphic hardware, and third, the investigation of new methodologies for simulating these neuromorphic circuits at massive scale in order to understand the orders-of-magnitude improvements in energy efficiency and cognitive abilities that are achievable with next-generation neuromorphic hardware.

The fundamental questions being address by this workshop are:

Q1. What are the key neuromorphic circuit primitives that are needed to capture critical biological computing mechanisms with the right level of abstraction?

Q2. What are the technologies needed to demonstrate and prototype these key neuromorphic circuit primitives?

Q3. What are the critical characteristics for effective large-scale simulation of neuromorphic circuits and systems?

Q4. What are the neuroscience-based benchmarks and datasets by which to effectively test and characterize neuromorphic computing circuitry and simulations?

The goal of the workshop is to understand the interdisciplinary research needs in identifying what the key neuromorphic primitives are, how they need to be implemented and how they can be scaled. A cross-cutting effort will be to discuss what testing data and benchmarking are needed for neuromorphic circuits to reflect biological-level performance and capabilities. This workshop will help build a research program to prototype the circuitry and simulation capabilities for a truly neuromorphic computer which will enhance the capabilities of AI by emulating biological neural systems while simultaneously achieving dramatic reductions in power consumption. Workshop participants will present relevant research and brainstorm to identify the key scientific research needs and challenges leading to the development of proof of principle neuromorphic circuits and reconfigurable ultra-low power systems for intelligent sensing. The impact of neuromorphic computing and the fundamental basic research needed in this area is essential to the success of scientific computing relevant to the Department of Energy (DOE).

## 2. Current state-of-the-art

The term 'Neuromorphic computing' was first introduced in the late 1980s. It has come to denote computing approaches, including both software and hardware, that are inspired by the structure and function of the brain. Unlike traditional computing architectures that are based on the von Neumann architecture in which memory and computation are physically separate, neuromorphic computing systems exploit biologically-inspired mechanisms for computation. Information is processed and stored locally where it is used, and signals are transmitted via efficient action potentials (or spikes). By using low-power analog elements and mitigating communication costs, neuromorphic circuits can achieve many orders-of-magnitude reduction in power requirements compared to conventional digital circuits. Moreover, because every component is continuously updating in parallel, neuromorphic circuits can also achieve orders-of-magnitude

improvements in speed compared to digital simulations of neuromorphic systems. In this workshop, we seek to develop a roadmap for realizing the next generation of neuromorphic systems that capture the rich structural and dynamical complexity of biological neural circuits with the ultimate goal of matching the cognitive capabilities of the brain.

Whereas numerous concepts intended for neuromorphic computing have been proposed, significant challenges remain for the development of emerging circuits that provide functional complexity comparable to biological neural circuits. The maximally-parallel, asynchronous, continuous-time analog dynamics intrinsic to biologically-realistic neuromorphic computing presents challenges for contemporary semiconductor circuits, which were primarily developed for executing clocked Boolean operations sequentially on physically distinct processing cores. On the other hand, analog and mixed-signal designs aiming to emulate the biological functions of neurons and synapses frequently incur substantial upfront hardware expenses and are difficult to reconfigure or scale to brain-sized systems. Much effort has focused on emerging nanodevices, such as memristive and spintronic technologies, which have gained popularity in research for their potential to closely mimic certain biological behaviors at exceedingly low power costs. Indeed, circuits based on these devices have shown potential to learn through both supervised and unsupervised approaches. Oscillatory and stochastic behaviors of various nanoscale devices have also attracted significant attention due to analogous biological behavior at the device level. Moreover, scaling challenges associated with electronic hardware systems, such as the plateau of Moore's law and the end of Dennard scaling, have motivated the search for alternative hardware systems beyond electronic platforms. In this regard, neuromorphic computing enhanced by photonic approaches presents an exciting avenue to potentially address some of these limitations and support high bandwidths and ultrafast switching capabilities while incurring low propagation losses. Nevertheless, bio-realistic circuits that leverage these behaviors remain elusive. Incorporating these emerging hardware principles into useful analog circuit building blocks often face significant obstacles related to device variation and robustness as well as obstacles related to achieving the rich, reconfigurable connectivity central to neurobiological learning. These obstacles have substantially impeded the construction of useful systems for practical applications. The primary goal of the workshop will be to identify novel emerging technologies and/or design strategies that can overcome the obstacles to the construction of biologically realistic neuromorphic circuits.

Despite significant efforts to explore the myriad design possibilities within neuromorphic computing, progress has been modest when compared to advancements in competing technologies like GPU driven generative AI, largely based on the simple perceptron model from the 1940s. Much effort has been expended on digital accelerators for deep feed-forward neural networks that rely on vector-matrix multiplication, but such uniform circuit structures have limited bio-realism and may therefore be unlikely to reproduce the learning capabilities, robustness, complexity, or energy efficiency of the brain. The true advantage of neuromorphic computing—in terms of ultra-high energy efficiency, resiliency, adaptability, and generalizability—lies in the development of *scalable* and tightly integrated biologically-realistic neuromorphic circuits, architectures, and algorithms. Indeed, much as the remarkable advances in deep learning have been driven by advances in digital acceleration, particularly as regards the essential role of GPUs, analogous advances in neuromorphic computing will be driven by concomitant advances in the development of biologically-realistic hardware. The complexity inherent in designing each of these neuromorphic components often leads to a disconnect between research at different levels. A major goal of

the workshop is to devise strategies for how research at different levels in the neuromorphic stack can coordinate most effectively in an integrated co-design effort.

Extensive hardware innovations have to come in tandem with simulation and algorithmic developments. There are several simulation packages that can accurately model biologically-realistic circuits but do not scale efficiently to large numbers of neurons. Large systems of simplified spiking neurons can be simulated using conventional digital accelerators, but such systems lack biological realism and performing supervised training on such networks presents challenges due to the need to maintain exactly reciprocal forward and backward copies of the network in order to support backprop. One approach is to train a non-spiking neural network model using back propagation and to then convert to an otherwise equivalent Spiking Neural Network (SNN) model. Surrogate-based methods and random feedback alignment are also approaches that are currently being studied. These approaches can produce SNN models at large scales that accurately approximate the performance of the original non-spiking model but fail to address the issue of online training and do not provide a method for incorporating more biologically realistic neurons and synapses. Other approaches involve the use of agent-based approaches for modeling different types of neurons and synapses. Again, a key limitation of these models is the ability to develop a scalable learning method. Spike-timing-dependent plasticity (STDP) is a learning rule that is observed ubiquitously in neurobiological preparations but its ability to explain the emergence of complex learned behaviors has yet to be demonstrated. Recently, deep attractor models with bidirectional connections have been shown to achieve classification performance similar to that obtained by conventional feed-forward neural networks possessing the same number of neurons, layers and independent connection weights. Because deep attractor models can be trained using only local Hebbian and anti-Hebbian learning rules, they are compatible with many existing neuromorphic hardware architectures, which like the brain, are restricted to employing only local learning rules. Whether deep attractor models can provide a basis for training more biologically realistic neuromorphic circuits is an open research question. A fundamental goal of the workshop is to focus on learning approaches that can efficiently scale to massive networks along with promising strategies for training biologically realistic neuromorphic circuits to perform useful, brain-like tasks.

Related to the question of simulation and algorithmic development is digital emulation. Specifically, a number of all-digital neuromorphic processors have been developed, many of which emulate important aspects of brain function, such as spiking dynamics, spike-based communication and STDP. While all-digital neuromorphic processors can in principle be scaled up to very large, brain-sized networks, the ability to emulate the cognitive performance or energy efficiency of the brain at scale has yet to be demonstrated. In part, this may reflect the lack of biological realism intrinsic to existing all-digital designs, which largely adhere to an integrate-and-fire paradigm. Nonetheless, an important goal of the workshop could be to identify how all-digital neuromorphic processors could be leveraged, if possible, to help discover the algorithmic possibilities and/or limitations of such systems.

The workshop seeks to integrate the above efforts into a cohesive vision. The urgency for this initiative stems from two critical challenges. At the hardware level, how can we emulate the functionality of biological circuits in the development of neuromorphic circuits? At the algorithmic level, advances in our understanding of biological computing mechanisms will require the co-design of neuromorphic circuits that can be assembled into complex, brain-like sub-systems. The workshop aims to convene a meeting of the nation's top experts to brainstorm and formulate a strategic basic research program for the co-design of

novel, biofunctionally-realistic neuromorphic circuits and neuroscience-based algorithms. The discussions will focus on how to best capture the functional complexity of their biological counterparts, ultimately providing a pathway for replicating the cognitive capabilities and computing requirements of the brain. This co-design effort will require innovative modeling approaches and advanced computing architectures for simulating interconnected neuromorphic circuits at the scale necessary to reveal complex emergent behaviors. In summary, our objective is to establish a foundation for future innovations in the realm of biologically-based artificial intelligence while addressing the pressing need for vastly lower energy consumption.

## 3. Long-term vision and near-term basic research goals

Unlocking the potential of neuromorphic computing hinges on discovering neuroscience-inspired algorithms that can efficiently process and model data in a manner inspired by the brain and mappable to modern micro- and nanofabrication techniques. Success also requires developing innovative architectures and hardware designs capable of emulating biological circuits with high efficiency. The goal is to decode the operations of biological neural networks, which can process diverse inputs and dynamically interact across a variety of spatiotemporal scales and modalities – including chemical, ionic, magnetic, optical, mechanical, and/or electrical. Our ultimate ambition is to explore a broad potential solution space that spans both software and hardware, aiming to create an integrated neuromorphic computing architecture. This emerging computing architecture could not only surpass the performance and scalability of current state-of-the-art deep learning systems but also achieve it with significantly lower hardware costs and computational demands.

To support this vision, this workshop research outlines a collaborative effort to design innovative computing architectures that prioritize neuro-realism while supporting the advancement of neuromorphic computing architectures and energy efficiency. The emphasis will be placed on developing and simulating novel analog computing circuits employing unconventional neuromorphic methodologies. Therefore, ideas that focus on single neuromorphic components in isolation without consideration of how functional neural circuits can be constructed, characterized, and vetted will be considered out of scope. The ideas proposed should focus on state-of-the-art prototyping, simulation mapping, and hardware development driven by biological realism with a focus on capturing the functionality and structural complexity of brain neural networks.

## 4. Open issues

The aim of the overall program in neuromorphic computing will be to support a vertically integrated vision, combining new computing insights from neuroscience with emerging neuromorphic devices to create a new type of computing architecture for both scientific and edge applications.

### 4.1. Neuroscience-based computing principles

This open basic research issue is driven by the fundamental question *"What are the key neuromorphic circuit primitives that are needed to capture critical biological computing mechanisms at the right level of abstraction?"* The goal of the activities in this space is to understand what principles of brain organization and dynamics underpin its functionality, robustness and cognitive capabilities and how these principles can

be translated into biofunctionally-realistic neuromorphic circuits and systems. Biological components are astounding in their diversity, energy efficiency, and complex behavior and the existing artificial neural network counterparts lack such sophistication, focusing primarily on simplistic neuronal and spiking behaviors. It is widely agreed that nervous systems can be aptly described as directionally interconnected networks of diverse neurons, each emitting complex spike analog signals when the combined inputs received from other neurons exceeds a threshold. The connection (synapse) between two neurons changes its weight continuously in an activity-dependent manner, but its sign remains constant and is uniquely determined by the excitatory or inhibitory nature of the sending neuron type. In this traditional "neural network" model, spatial-temporal neural patterns (which neurons spike and when) represent the content – memories, decisions, plans, etc. – whereas synaptic plasticity (weight changes) underlies learning. It is also broadly recognized that sparse activity is key to energy efficiency: fewer than 1% of the neurons spike at a time. In addition, neuronal adaptation seems to play a key role in computation. It is a slow process, and it builds up over several spikes. Most neurons, particularly the excitatory neurons, will respond to the input with a spike train where intervals between spikes increase successively due to adaptation until a steady state of periodic firing is reached. Bursting and stuttering neurons respond to constant current stimulation by sequences of spikes interspaced by long non-spiking intervals, periodically or aperiodically respectively. It is an open issue how to understand and model this neuronal diversity.

Another organizational principle of nervous systems is that brain connectivity is also extremely sparse: a typical mammalian neuron "only" contacts ~10,000 other neurons, i.e. less than 0.01% of the whole network even in small rodent brains. Moreover, connectivity is exquisitely specific and structurally plastic. Any given neuron has a limited pool of partners (typically ~100,000 or ~0.1% of all neurons) that it can possibly connect to, defining a crucial blueprint of that particular functional circuit (say, visual recognition vs. spatial navigation). Which 10% of its 'partner pool' a neuron actually contacts, varies over time based on their activity, providing a core substrate for memory storage and retrieval. Sparse, highly-specific connections and structural plasticity are absolutely fundamental to the working of the brain, yet are seldom captured in hardware and software models alike. While such an organization can in theory be emulated by all-to-all connectivity and a majority of quasi-zero weights, in practice, that solution is eminently non-scalable in physical space due to volume packing and heat dissipation. Therefore, that alternative artificial design prevents the assembly of the required minimum network size to tackle difficult computational tasks.

Long-range connection pathways typically remain computationally segregated when they converge on individual neurons based on non-linear dendritic processing and compartmentalization. Non-linear dendritic processing combinatorially increases neural information capacity by allowing every circuit to precisely gate and thus selectively control each independent processing stream.

Neuronal communication in the brain is mediated by synaptic transmission. For example, basket neuron cells in the hippocampus can receive input from more than 30 distinct neuron types. Synapses have been shown to change their synaptic strengths via Hebbian STDP. Different types of STDP responses have been observed. For example, symmetric STDP was observed in the hippocampus between pyramidal neurons. Asymmetric STDP was observed between cultured hippocampal neurons. Nevertheless, biological synaptic transmission in the brain is a stochastic process with considerable variability across spikes. Variability in synaptic transmission is present across different types of synapses in the brain, with different synaptic behavior in different regions. This variability arises because neurotransmitter release is probabilistic and the postsynaptic response to the neurotransmitter release has variable timing and amplitude. Nevertheless,

the brain seems to perform robust computation. How does the stochasticity of the synaptic transmission impact computation in the brain and what mechanisms are in place to support the robustness? These insights will provide support for the translation into new types of neuromorphic hardware.

Additional contemporary neuroscience insights that can help reimagine neuromorphic computing involve 1) incorporating models of a variety of brain cells, including glia, in addition to the more complex neuron and synapse models; 2) developing specialized circuits tailored to different brain regions such as the hippocampus, cortex, brainstem, and cerebellum; 3) innovating neuromorphic in-sensor computation by understanding the computational processes happening within biological sensors, like the retina; 4) advancing neuromorphic designs by leveraging insights from the study of biological neural networks and connectomic maps.

In addition, the locality of cortical plasticity, microcircuit diversity, self-organization capabilities, emergence of organized network spiking behaviors, etc. are additional features that could have a significant impact on the energy efficiency, adaptability and cognitive capabilities of neuromorphic systems. Although incorporating these neuroscience principles in next-gen neuromorphic designs undoubtedly constitute a formidable scientific and engineering challenge, the resulting technological breakthroughs promise to radically transform the power of machine computation, ushering in a new era for human society. The aim is to develop new computing principles that can support processing of zettascale data and complex spatio-temporal problems, which are currently not effectively addressed using traditional computing methods.

The neuroscience-based computing principles identified should:

- Offer a well-defined computation (computational primitive). The circuit/neuron must be well-studied enough by neuroscience that there are already agreed-upon (or at least widely accepted) descriptions of the functionality and evidence on how it influences macro-level behavior, such as learning or cognition.
- Be captured by a small neural circuit (the exemplar)
- Have well-characterized exemplars – at minimum the essential components and biological mechanisms for each exemplar should be identified and reasonably well described. The output of the circuit should be predictable given a defined set of inputs.
- Ideally the neuroscience challenge should offer multiple exemplars.
- Should include quantitative metrics for assessing how well a given neuromorphic circuit primitive, once fabricated, approximates the range of behaviors displayed by comparable biological circuits.

#### 4.1.1. Physical circuits for cortical, hippocampus, thalamus, sensing, motor control, etc.

There are numerous well-studied biological circuit motifs in which the relative contributions of subcellular and circuit level components to learned behaviors can be explored. An example is the dimensional expansion at excitatory feed-forward synapses into granule cells in the dentate gyrus of the hippocampus or in the cerebellar cortex. Other examples include the excitatory projections from the lateral geniculate nucleus to the spiny and non-spiny stellate cells in the primary visual cortex or from thalamic nuclei to barrel cortex in rodents. The early olfactory system, in both insects and vertebrates, provides another example system where the interplay between subcellular and adaptive circuit components can be explored

in fabricated neuromorphic materials. These few examples are a small subset of the circuit motifs found in biological systems that have been experimentally characterized.

### 4.1.2. Neuromorphic micro-brain approaches

Insects achieve complex movements, environmental adaptation, and reproduction despite their small nervous systems. Advances in imaging techniques in experimental neuroscience have supported the non-invasive determination of neuronal connectivity in a variety of brain types. Connectome data is now available for a range of species, including *Caenorhabditis elegans* and the fruit fly. These connectomes indicate that structural connectivity features non-optimal component placement which incurs higher energy costs for connection establishment and maintenance but enables a wider range of brain network dynamics. *C. elegans*, a type of roundworm, is the only animal for which a complete connectome is available, featuring 302 neurons for the hermaphrodite form and 381 neurons for the male form. The central brain of *Drosophila melanogaster* comprises approximately 135,000 neurons, significantly more so its connectome is only partially characterized. By comparison, the mouse brain has over 100 million neurons, the macaque brain has over 1.3 billion neurons and the human brain has tens of billions of neurons.

Neuronal circuits are often more completely characterized in smaller brains and offer insight into the necessary building blocks for brain function. For example, it was recently demonstrated that protocerebral bridge neurons in fruit flies (*Drosophila*) multiply their inputs encoding perceived heading and perceived velocity. What makes this a good exemplar is that the inputs to these neurons are also well-described in the neuroscience literature, so neuromorphic efforts to reverse engineer this circuit could be validated. Such engineered "artificial neuromorphic microbrains" could be integrated into autonomous robots, to efficiently analyze the sensor inputs, compute a suitable trajectory and actuate the robotic limbs. Biologically-inspired neuromorphic approaches capable of unsupervised learning could ensure a degree of error tolerance and adaptability unmatched by any other methods in such edge systems.

## 4.2. Neuromorphic circuit primitives

Understanding what the neuromorphic circuit primitives need to be comes hand in hand with figuring out their implementation. This effort is driven by the second fundamental question *"What are the technologies needed to demonstrate and prototype these key neuromorphic circuit primitives?"*

Novel neuroscience inspired circuits based on new devices and designs, and new principles, have to be developed to support the diversity of neuromorphic functionality needed. Such devices should enable high energy efficiency and the co-location of memory and processing capabilities in a compact way. These circuit technologies should cohesively span orders of magnitude in performance, in terms of energy efficiency, endurance, range of tunable resistance / capacitance / inductance values, temporal delays, etc. New circuitry metrics might have to be devised to capture non-ideal behaviors across large populations of non-ideal devices and allow device-architecture co-design. Prototyping platforms should support a variety of circuit and device technologies to experimentally demonstrate novel analog primitives and proof-of-concept neuromorphic computing principles.

The neuromorphic circuits primitives should be:

- Focus on emulating the functionality of the neuroscience-based computing principles to the best extent possible. The neuromorphic implementation should offer testable predictions in hardware and support the development of equivalent models for neuromorphic computing modeling and co-design simulation.
- Being willing to "fail" at accurately reproducing their biology counterparts is acceptable, provided the lessons learned identify addressable technological gaps or limitations.
- Be buildable from devices and materials that exist today or highlight new functionalities needed.
- Be capable to be imaged and probed
- Take advantage of nontraditional implementations/integration/uses of existing and novel devices for the development of neuromorphic circuit primitives and their interconnects.

The building blocks of neuromorphic circuits should be defined. Hardware-mappable neuronal primitives that reliably represent the functionality of their biological counterparts are needed. A co-design approach is needed, starting from neuroscience insights and experimental measurements of biological building blocks, e.g. neurons, synapses, glial cells, etc. Identifying all the different building blocks in a given region is an important step towards understanding in detail how this brain region works.

Neuronal behavior has been core to the modeling and mapping efforts, but limited progress has been made in defining comprehensive hardware-mappable models that realistically represent biological diversity. There are some approximate hardware-mappable hardware models in the literature, but each has advantages and disadvantages. Therefore, there are many open issues to tackle. For example, one of the basic neuronal models and the most widely mapped in hardware is the leaky-integrate-and-fire model. This model makes the assumption that the information is entirely encoded temporally, in the events that are happening at precise moments of time. No attempt is made to describe the shape of the actual action potential, which seems to be an important feature in the communication of stimulus history in the brain circuitry. Moreover, in this model, the input is integrated linearly and does not consider the state of the postsynaptic neuron. These are major limitations, since no memory of previous spikes is kept, resetting the state after spiking to the original state. This simplistic model aims to capture fast spiking neurons, e.g. certain subtypes of inhibitory neurons, but it is unclear if it is even sufficient for this modeling task. This simplistic model cannot capture the processes that lead to learning and adaptation in the brain and cannot capture the different spiking behaviors needed. Despite these severe limitations, the leaky-integrate-and-fire model has a widespread use in the modeling of spiking neural networks and for event-based hardware implementations due to their computational simplicity allowing for easy scalability. Compact transistor-based circuit equivalents have been developed to provide very fast digital pulses and address-event representations have been implemented to transmit spikes off-chip. While such models require only a few transistors per artificial neuron, they do not produce the diversity of behaviors necessary for investigating biofunctionally-realistic neural circuits.

More sophisticated models, e.g. Izhikevich models, and their hardware implementations have been developed to be able to capture a broad range of spiking behaviors seen in biological neurons. 4-parameter and 9-parameter variants exist for the Izhikevich models, which can represent a variety of regular spiking,

fast spiking and bursting behaviors with various degrees of complexity. Analog circuit implementations have been proposed, particularly for the 4-parameter Izhikevich model. While it can take more transistors than a leaky-integrate-and-fire model, it provides more capabilities for neuromorphic computing. Neuroscience inventories are also increasingly developing Izhikevich models based on experimental neuronal data. However, additional refinements of these models are needed, e.g. to limit the firing rates within biologically realistic ranges.

Electrophysiologically realistic models, such as Hodgkin-Huxley, can simulate in even more detail the biological properties of neurons. However, this biological realism comes at the expense of complexity. The model is based on differential equations that require a significant number of transistors for analog circuit implementations. Emerging device technologies should be explored to map biofunctionally-realistic neuronal models in an area and energy efficient way. Moreover, ways to incorporate behavioral variability should be incorporated, to mimic any diversity of spiking between neurons of the same type.

In addition, important consideration needs to be paid to the connectivity between neurons and the synaptic behavior, including the physical (e.g., electrical, magnetic, ionic, photonic) connectivity between the neurons and synapses. The richness of connectivity across the brain regions is highlighted in prior experimental neuroscience work and has to be considered for translation into the neuromorphic hardware. There is an open research question regarding how much of this complexity needs to be implemented in artificial computing circuits. This will likely require novel hardware to provide extensive computing resources for large-scale models. New hardware technologies, such as memristors have shown biological relevant STDP, as well as other behaviors such as short-term plasticity (STP) behavior in a compact and efficient way. However, such emerging hardware that can provide dense storage is notorious for having device-to-device and cycle-to-cycle variability. The goal is to investigate neuromorphic primitives using existing and emerging devices that show robustness to synaptic unreliability (e.g. variability, quantization and turnover) and can perform tasks of high accuracy of scientific relevance.

#### 4.2.1. Emerging electronic analog circuits

Perhaps the greatest challenge facing neuromorphic engineers is the design of analog circuit primitives that capture the full dynamic range and adaptive capabilities of their biological counterparts. At the subcellular level, examples of physiological components that may be needed in order to capture full bio-mimetic functionality include spine morphology, dendritic processing, tripartite synaptic dynamics, somatic integration, spike generation, axonal propagation and stochastic synaptic release, the latter varying in accordance to STP rules. At the circuit level, examples of recurring motifs include feed-forward excitation/inhibition and both lateral and top-down excitatory and inhibitory feedback. All of these circuit motifs employ functional adaptation governed by local learning rules, particularly STDP, that underlies the acquisition of learned behaviors. This workshop will explore proposals for fabricating bio-mimetic circuit motifs that exhibit adaptive learned behaviors similar to those exhibited by equivalent circuits found in nature while also operating within similar size, weight and power (SWaP) constraints.

A key feature that should be considered as part of the neuron and synapse neuromorphic circuit design research approaches that has been a limitation of existing neuromorphic computing approaches is *scalability*. It is important that scalability be considered even in the design of small-scale circuits, as, eventually, it is critical to build large scale neuromorphic computing architectures (> 100 billion neurons).

Though it is important to start with a small number of neurons and synapses, planning for scalability should be a part of the basic research process. Moreover, compartmental representation of network connectivity has to capture the computationally distinct subunits that emerge from layer-specific axonal targeting, a key component of biological circuitry. This feature will likely complicate the hardware mappability and the scalability even further.

It is well known that there is a large diversity of different types of neurons and synapses in biological brains, even in simple organisms. To be able to produce different neuroscientifically accurate behaviors, as well as to allow for exploration of different capabilities, one path forward is to focus on the development of circuits for neurons and synapses that are *programmable* and *reconfigurable*. As such, it would be worthwhile to focus on developing this library of circuits that also emphasize plasticity, programmability, and reconfigurability. There is an open issue where and how these features can be incorporated into neuromorphic circuitry with maximal contribution to the computation while minimizing hardware costs.

A tremendous area of research in neuromorphic computing has focused on a wide array of different types of devices and materials for neuromorphic computing. New electronic devices, e.g., memristors, phase change memories, etc. have shown STDP and STP behavior and can be densely integrated with transistor circuits for a broad range of spiking functionalities. For example, oxide-based memristors have shown experimentally a diversity of STDP behaviors similar to those observed in biology for synapses in different layers of the neocortex and the hippocampus. These devices show complex internal behavior, have analog state programmability and state retention driven by ionic movement, with some resemblance to the ionic movement in biological neural systems. Moreover, these devices are suitable for back-end-of-line integration between different metal layers during the CMOS manufacturing, thus having potential for ultra-dense systems as required to map the large density of synapses in the brain structures. Other physical phenomena can also be considered. The rich physics of magnetism enables a wide range of opportunities to utilize spintronic phenomena in neuromorphic computing circuits. The hysteresis intrinsic to ferromagnetic non-volatility can emulate the memory capabilities of neurobiological systems by taking advantage of a variety of devices, each with unique features and relative strengths and weaknesses. Additionally, the magnetic field interactions between nearby magnets that are electrically isolated permit unique circuit structures leveraging the implications of such contact-free interactions on input/output isolation, and therefore cascading and fan-out. Magnetic technologies, e.g. STT-MTJ, are currently being developed, producing device characteristics relevant to neuromorphic circuits such as analog/multi-level resistance states, and neuron-like integration, leaking, and firing. Though many of these emerging devices have radically different behaviors, it would be of tremendous use to the research community to define and maintain a *library of common / canonical / primitive circuits building blocks*.

### 4.2.2. Potential photonics integration for neuromorphic computing

The brain is based on high connectivity, which can be difficult to achieve in purely electronic architectures. Light-based structures naturally offer higher bandwidths for data transmission compared to electronic connections. In the context of computing, this directly translates into faster communication between different components of a neuromorphic system, enhancing the overall processing speed. Higher bandwidths can also enable parallel processing, allowing multiple computations to occur simultaneously. This aligns well with the parallel nature of neural networks in the brain, enhancing the efficiency of

neuromorphic computing systems. In addition, photon propagation and manipulation can in principle occur with very high efficiencies and incur very low levels of loss. As a result, light can offer more energy-efficient computing compared to traditional electronic circuits. Lower energy losses also mean less heat generation and dissipation in photonic architectures, something that can minimize heat-related issues which pose major challenges in electronic processors.

Intense research efforts are devoted to advancing photonic neuromorphic computing. The potential success of this growing field depends upon addressing a number of key challenges. Firstly, developing high performance chip-scale photonic neuromorphic hardware requires photonic circuits capable of performing both linear and nonlinear operations on the same chip. This contrasts with the current-state-of-the-art which uses separate, often hybrid optoelectronic hardware, to accomplish linear processing and nonlinear activation functions. A key technology in this direction is the development of efficient, ultrafast all-optical switches that are compatible with common material platforms in integrated photonics. This is an active area of research in nonlinear optics and ultrafast photonics with valuable inputs from material science and technology. Secondly, efficient light-based processors need integrated light emitters and sources that are highly tunable while consuming low energy. In this respect, photonic neuromorphic computing could certainly benefit from recent advancements on chip-scale optical frequency comb sources that emit light in a wide range of spectra. Finally, achieving efficient and scalable integration of photonic and high-speed electronic components is yet another hurdle to be addressed. Along these lines, low-loss, ultra-high bandwidth electro-optic modulators and optical interconnects continue to evolve and will be indispensable parts of future photonic processors. Overcoming these challenges crucially depends on active collaborations among multiple fields ranging from ultrafast and nonlinear photonics and electronics to material and computer science and presents immense opportunities for future generations of neuromorphic computing hardware.

### 4.3. Neuromorphic computing modeling and co-design simulation

The efforts in algorithmic research and circuit development have to be scaled up in order for the full impact of novel neuromorphic technologies to become visible at the scientific and societal levels. Therefore, the third fundamental question to be discussed in the workshop is *"What are the critical characteristics for effective large scale simulation of neuromorphic circuits and systems?"*

There are a number of significant challenges simulating the behavior of biofunctionally-realistic spiking neural networks executing on analog neuromorphic circuits. Three key challenges relate to the necessity of approaching neuromorphic computing from the standpoint of biological functionality, the challenge of encoding spikes and the difficulty of reliably and efficiently training spiking neural networks.

In terms of rethinking the perceptron model, it is important to realize that this model has been the backbone of neural networks since the 1950s. This basic representation of neurons and synapses has proven remarkably powerful in modern AI. However, this model is focused on minimizing the computational complexity of neural networks, and less on the bio-plausibility of the model. The structure and function of the brain is quite different from that of the perceptron model. As powerful as the perceptron model is, there is great potential for significant advances in exploring a biofunctionally-realistic model as the base component for a neural circuit and SNN.

Encoding spikes in a bio-realistic way seems to be another key component required for neuromorphic computing, by comparison with traditional AI. One of the key strengths of generative AI rests in the ability to encode data into real numbers that can then be used to train ANNs, leveraging the sparsity of the data. An SNN works on binary signals that hold information based on the sequence and timing of their generation. These signals have a temporal sparsity that is difficult to encode without aggregating the spikes into a time window, thus losing an element of the temporal nature of the data. A significant research challenge is to explore how to effectively encode the sparsity of the temporal spikes such that a SNN can learn from them. There needs to be further research into how to train SNN models based on the latest neuroscience findings, and on how to optimize algorithms that simulate these training methods.

### 4.3.1. Design space exploration

As our brains all are different, the number of possible designs of an SNN as the size grows is limitless. Assessing what network architectures will produce reasonable results is a significant challenge. Evolutionary algorithms and Bayesian optimization have been successfully demonstrated on a small scale, but with strict constraints. A bio-plausible design will require massive numbers of components and connections. As our brains all are different, having different "designs" some aspects formed by nature, others by experience. We will need to explore what are the optimal design characteristics for a basic SNN, what it is born with. Then what are the optimal structures for the system to learn new information. This design process can take clues from neuroscience, and existing neural architecture discovery approaches, but novel and scalable methods will be required. These techniques can be utilized at the circuit design level as well, in a co-design approach. Given the complexity, these design space explorations will require significant interdisciplinary know-how and high-performance computing resources. Both human-driven design as well as automated AI-driven optimization will be needed to enable the fast scientific discovery of new neuromorphic computing primitives.

### 4.3.2. Scalable integration

There are several state-of-the-art SNN simulators, each supporting various neuro-inspired algorithms -such as plasticity, evolutionary algorithms, and backpropagation-based learning- with a variety of models and simulators for the underlying neuromorphic architectures. Despite these efforts, there has not yet been any success in translating these simulators at the biorealistic circuit level. There is an urgent need to determine the building blocks of the future neuromorphic simulators and study how to connect, program, and configure them, to unlock the full potential of neuromorphic computing at large-scale. This includes but not limited to, defining neuromorphic benchmarks that enable comparisons between different architectural and algorithmic simulators, defining *metrics* and evaluating criteria that capture the integrated performance of the neuromorphic computing, and developing *reconfigurable* architectures and algorithms that can be integrated to achieve scalability. Additionally**,** one promising avenue for basic research on scalable neuromorphic computing involves designing algorithms and architectures that seamlessly *integrate* with other computing frameworks. By enabling a seamless integration, researchers and engineers can leverage the unique strengths of each framework. This could lead to breakthroughs in computational efficiency, speed, and adaptability, significantly advancing fields such as artificial intelligence, complex modeling, and temporal and adaptive decision making, ultimately pushing the boundaries of what is computationally possible.

There is also a need to scale these simulation methods to hundreds of millions of neurons and synapses using high performance computing. This presents some key challenges, such as how to effectively use existing hardware (CPUs, GPUs, FPGAs and accelerators) to support the simulations, and how to effectively simulate the synapse interactions among the neurons. Simulating the communication among neurons typically requires a sophisticated message passing algorithm which creates a great deal of network activity. This input/output bound task is manageable for small simulations, but quickly becomes a limiting bottleneck as the model grows in size. New approaches are needed to create bio-plausible simulations of synapse connectivity and communication. The goal is to perform the necessary computational calculations quickly and in a massively parallel way. Research is needed in how to leverage the computational power of existing hardware and high-performance computing infrastructure for simulating biofunctionally-realistic computing primitives leading to neuromorphic architectures.

### 4.4. Data and testing towards analog neuromorphic computing architectures

A cross-cutting research question driving this effort is "*What are the neuroscience-based benchmarks and datasets by which to effectively test and characterize neuromorphic computing circuitry and simulations?"* This is a fundamental question that has to be addressed throughout the co-design effort is related to data and metrics, in order to ensure the cohesive research and development towards scalable neuromorphic computing architectures.

While a significant amount of data and neuroscience literature has been generated to characterize different types of neurons, synapses, etc. in different brain regions, the data is scattered across different sources, making it hard to compare and to use for modeling and hardware mapping. This data also often fails to consider the physical interconnections between synapses and neurons, neglecting the requisite interconnection circuitry and its associated hardware costs. Moreover, the data might be more abundant for particular neuronal types, for particular brain regions or for particular species. This lack of neuroscience data and insight might make it difficult to model biologically-realistic neuromorphic building blocks. Therefore, it is important to identify what neuroscience data is currently available and sufficient for mapping to software simulations and hardware emulations. For example, the rodent hippocampus is among the most intensively studied neural systems and there are inventories that curate the experimental evidence about neuronal types, synaptic variability, etc. together with models for these building blocks. However, additional work needs to be done to synthesize the experimental neuroscience data of different brain regions and aggregate it in inventories useful for developing neuromorphic building block-equivalents.

Testing and benchmarking these neuromorphic technologies for biologically realistic performance require access to neuroscience-relevant data and metrics at different levels of the stack. Discussions will center around what relevant data is available in the neuroscience literature that can support these efforts. Gaps will be identified to be able to guide future efforts. Potential metrics that bridge biologically-realistic neuromorphic circuits and neuroscience experimentation will also be discussed.

### 4.4.1. Metrics and requirements for circuit and architecture design

A comprehensive view of the metrics for neuromorphic circuits and computing architecture takes into consideration its neuroscience inspiration as well as engineering requirements. The circuit primitives would have to be considered along several interconnected dimensions.

The first dimension refers to whether the circuit primitive is representative of experimental data or is based on a theory of neuronal behavior. The data-driven circuit primitives would have high biological realism, with their behavior closely matching the biologically-equivalent behavior as measured by metrics across a broad range of inputs and conditions. For example, measures of spike train synchrony could be used to quantify the degree of similarity between spike patterns obtained from analog neuronal circuits and experimental measurements of biological neuron spiking. On the other hand, circuits could also be derived from first-principles dynamics, e.g. membrane channel behaviors, temporal integration, spike generation, etc. Such primitives could be easier to implement by mapping the respective set of equations to tunable circuit blocks. The advantage is that the resulting primitives could be used to produce predictions and potentially provide insights into what dynamics plays what role in supporting efficient and robust neuromorphic computation. Metrics such as the predictive power score, could provide insight into the predictive power of the circuit. At the architecture level, intermediate approaches seem possible and could be based on first-principles dynamics on neuronal models that are linked by data-based estimates of large-scale patterns of connectivity.

The other dimension is related to complexity. The fine-grained representations, e.g. mimicking molecular dynamics inside different neuronal structures in new types of devices, could provide significant capabilities in neuromorphic hardware to support emergent relationships between structure and function. However, measuring these emergent behaviors without disturbing the system could be challenging, so appropriate metrics are needed. Coarser-grained representations, at the level of neurons or ensembles of neurons might be useful in providing approximation of the properties of smaller units, in case they are not typically directly accessible for measurement or to avoid disturbances. Moreover, biological expressivity might also significantly increase complexity. Therefore, it will be critical to understand what are the neuroscience-based behaviors that are expected of the neuromorphic computing circuits and at what level of complexity, with any trade-offs that might impact functionality.

The third key dimension is related to implementation metrics. Metrics such as computing density, energy efficiency, computing accuracy, and on-chip learning capability have been proposed to quantify and benchmark the area, power consumption, performance, etc. These metrics might have to be expanded as new technologies and new insights are being adopted. At a minimum, the model circuit architecture should have the same activity sparseness and connectivity sparseness as the biological network it aims to emulate. Additional metrics, e.g. related to scalability and robustness to hardware imperfections, should also be carefully discussed and considered. Appropriate datasets that can provide biologically-realistic inputs and outputs are also needed to support the neuroscience-inspired metrics to be used and thus the end-to-end development of neuromorphic systems.

### 4.4.2. Potential for accelerating scientific discovery

The aspirational goal for this workshop is to set the vision that will enable the development of the necessary circuitry primitives for neuromorphic computing which could accelerate scientific discovery in 10-20 years.

Reverse engineering the smallest units of select brain structures into functionally-equivalent circuits would provide significant advancement in many complementary scientific fields. In the area of computing, this endeavor will provide insight into what foundational concepts have to support efficient and scalable brain-inspired computing. Algorithmic developments will be needed to support large-scale simulations, capable of connecting neuroscience learning paradigms and analog circuitry models in a scalable way on high performance computing systems. In the area of circuit design, this effort will contribute with methodologies to translate the functionality of key biological neuronal primitives into analog circuit primitives with similar behavior across a range of conditions. Device variability and process, voltage and temperature variations tend to significantly affect the behavior of analog circuits. Therefore, design developments inspired by the robustness of brain structures to non-idealities will be needed to support functional robustness of the resulting circuit primitives. New interconnection mechanisms should be developed to enable the integration of large numbers of these computational primitives to provide bio-mimetic behavior at the system level.

All these advances will advance the field of neuromorphic computing towards the development of hardware systems capable of performing at the energy efficiency and complexity of their biological counterparts. These hardware systems could be embedded into efficient high performance computing clusters beyond the exascale to support scientific advances in areas of interest to DOE for societal benefit, e.g. new materials, new technologies, better energy systems and better forecast capabilities. Long term, such neuromorphic hardware could also support the emulation of brain regions, support the advancement of neuroscience research and development of new therapies and prosthetics. The neuromorphic algorithmic and circuit insights could also support developments in other engineering fields in need of efficient and robust algorithms and circuits, e.g. communications and control systems.

## 5. Key open research questions:

1. What are the key neuromorphic circuit primitives that are needed to capture critical biological computing mechanisms with the right level of abstraction?

2. What are the technologies needed to demonstrate and prototype these key neuromorphic circuit primitives?

3. What are the critical characteristics for effective large-scale simulation of neuromorphic circuits and systems?

4. What are the neuroscience-based benchmarks and datasets by which to effectively test and characterize neuromorphic computing circuitry and simulations? -CROSS-CUTTING QUESTION

## 6. Conclusions

In conclusion, one of the grand challenges of our time is to rethink computing in a way that is both physically intelligent and scalable. A potential solution lies in creating a new computing paradigm inspired by the brain's cognitive functionalities, capabilities and energy efficiency. Engineering a neuromorphic computing architecture with similar attributes is one of the most exciting and challenging scientific endeavors, promising significant breakthroughs. However, this task is complex due to our incomplete understanding of the brain's computational mechanisms, including the intricate behaviors of neurons and their connectivity patterns. The potential impact of neuromorphic computing on the future of scientific

computing, particularly for the Department of Energy, is immense. It necessitates emerging computing architectures to efficiently process beyond exascale and autonomously analyze complex datasets. To achieve this, new circuit and device technologies must be developed and integrated to emulate the brain's complexity, and extensive research, simulation, and prototyping of neuromorphic computing circuit architectures are essential.

This workshop aims to gather interdisciplinary experts to draft a set of basic research needs and brainstorm the next steps for neuromorphic computing research and development. Our focus is on two key areas:

- Innovative research into neuromorphic circuits inspired by advances in neuroscience
- The development of simulation methodologies for neuromorphic circuits at increasing scales.

The primary goal is to create a basic research program that accelerates AI capabilities while significantly reducing power consumption, identifying key research needs and challenges to develop proof-of-concept neuromorphic circuits.

## 7. Call for white paper submissions

The 2024 Workshop on Basic Research Needs for Neuromorphic Computing will inform and draft a set of grand challenges for advancing the field of neuromorphic computing and developing proof of principle neuromorphic circuits applicable for High Performance Computer (HPC) acceleration for scientific discovery, and brainstorm ideas needed for a successful, robust, and world leading basic research program.

Engineering novel neuromorphic computing systems with functionalities, capabilities, and energy efficiency similar to biological brains is one of the most exciting and challenging scientific endeavors of our time. This workshop aims to identify key research needs, challenges, and next steps necessary to develop biofunctionally-realistic neuromorphic circuits primitives that capture the functionality of neural systems found in nature. Moreover, simulating neuromorphic computing primitives integrated into networks will be key to understanding their behavior at scale, particularly for those computing architectures where full-scale commercial fabrication is not yet readily accessible. Appropriate neuroscience datasets and metrics will have to be established to vet proposed neuromorphic circuits.

In the development of new circuits and methodologies for neuromorphic computing, it is critical that there is close collaboration among circuit designers, computer engineers, computational neuroscientists, and algorithms and simulation researchers. This workshop aims to bring together a diverse range of experts across three complementary technical areas.

Submit your position paper to the technical areas below:

1. **Neuroscience algorithms and translation to neuromorphic analog circuits**
   This technical area is driven by the fundamental question *"What are the key neuromorphic circuit primitives that are needed to capture the full functionality of critical biological computing mechanisms?"*. The goal of the activities in this space is to understand what principles and circuit structures of brain organization and dynamics underpin its functionality and robustness capabilities and how these principles can be translated into functionally-equivalent neuromorphic circuits and systems that could be practically implemented (with available technology?). Ideas related to neuromorphic computing principles inspired from brain regions/functions (cortical, hippocampus,

thalamus, sensing, motor control, etc.) are sought after. Topics related to neuromorphic approaches and emulations of small invertebrate brains are also of interest.

2. **Technologies and prototyping of neuromorphic analog primitives**
   This technical effort is driven by the fundamental question *"What are the technologies needed to demonstrate and prototype key neuromorphic circuit primitives?"* Ideas related to novel neuromorphic circuits based on new devices and designs, and new principles guided by neuroscience-inspired functionality are of interest. Ideas related to emerging analog technologies that provide orders of magnitude in performance, parallelism, energy efficiency, tunability range, temporal delays, etc., and that mimic the biological behavior and robustness of key primitives are welcomed. Also of interest are topics related to high neuromorphic connectivity capabilities, e.g. optoelectronic technologies and photonic interconnects.

3. **Scalable integration for neuromorphic computing modeling**
   The fundamental question driving this technical area is *"What are the critical characteristics for effective large-scale simulation of neuromorphic circuits and systems?"* New approaches are needed to create simulations of large-scale biofunctionally-realistic neural networks, diverse synapse connectivity, and sophisticated network activity. Of interest are ideas related to novel methods to integrate and to scale up the simulation of the neuromorphic circuit primitives using high-performance computing in order to understand their interactions in the context of hundreds of millions of neurons and synapses. Also welcomed are novel methodologies for the efficient exploration of the large co-design space between neuromorphic algorithms and circuit technologies.

When discussing the technical idea and how it fits in the technical area(s) and the overall vision of the workshop, include a discussion on the benchmarks, metrics, and/or datasets requirements for neuromorphic computing for your proposed implementation.

Submit your position paper in response to the call for position papers by **July 22, 2024**.

The structure for the ideal position paper may include several of the below themes:

1. Neuroscience-inspired computing principles
2. Translation to analog microelectronic circuits
3. Modeling and simulation approaches
4. Performance metrics, data requirements, and energy efficiency

The position paper should be an individual submission, one paper per investigator. The format is 1 page (+1 extra page for figures, captions and references only), 11pt font size submitted in a Word or PDF document. The primary theme should be mentioned during the submission, a secondary theme is optional.