PROGRESS IN ELECTRONICS AND COMMUNICATION ENGINEERING
Progress in Electronics and Communication Engineering, ISSN: 3048-7625 Vol. 3, No. 2, 2026 (pp. 53-61)
RESEARCH ARTICLE
ECEJOURNALS.IN

Neuromorphic Computing Architectures for Low-Power Edge Inference in IoT Applications

Letahun Nemeon[1]*, Van Jiang[2]

[1]Electrical and Computer Engineering Addis Ababa University Addis Ababa, Ethiopia
[2]School of Electrical and Electronic Engineering, Newcastle University, Singapore

KEYWORDS:
Neuromorphic computing,
Edge inference,
IoT,
Spiking neural networks,
Low-power AI,
Event-driven processing,
Loihi,
TrueNorth

ARTICLE HISTORY:
Submitted : 18.11.2025
Revised : 16.02.2026
Accepted : 04.03.2026

https://doi.org/10.31838/ECE/03.02.08

Abstract

The blistering advancements witnessed in the Internet of Things (IoT) ecosystem have led to the urgency of search outcomes that provide energy efficient, low latency, and smart processing support that manifests itself right at the edge. The traditional edge computing devices built on von Neumann architecture are inherently limited by being optimized to be adjacent in memory and processing components, which results in the high power required and latency, which are highly undesirable in devices with limited power resources; wearables, smart sensors, and remote-monitoring systems. A Neuromorphic computing approach based on the human brain blueprint presents a better alternative due to its implementation of spiking neural networks (SNNs) that allow asynchronous and event-driven computation, thus dramatically lowering power demand and yet being able to run real-time inference. This paper includes a holistic comparison of the most recent examples of the neuromorphic hardware platforms, such as Intel Loihi, IBM TrueNorth, and BrainScaleS system, with regard to their learning architectures, energy consumption, throughput, and suitability to the edge-based IoT applications. An empirical case study is carried out, which is a keyword spotting case study and event-based gesture recognition case study using Google Speech Commands and the DVS Gesture dataset sets, respectively. The neuromorphic systems presented a significant energy per inference improvement up to 15x and reduced latency at about 30 on traditional ARM Cortex-M7-based ANN implementations at very slight sacrifices in terms of classification accuracy. In addition, the Loihi platform showed on-chip learning utilizing spike-timing-dependent plasticity (STDP), which made the platform capable of dynamic and adaptive inference in fluctuating IoT applications. This paper also examines the integration issues of neuromorphic systems which include the compatibility with sensor modalities, immature software toolchains, and complexities to convert ANN to SNN. The results indicate that neuromorphic systems are highly suitable in ultra-low-power edge intelligence and provide a paradigm of scalable and biologically inspired, next-generation IoT systems. The lines of future research are described in such aspects as hybrid neuromorphic edge-cloud learning, unification of benchmark requirements, and designing secure, programmable, and reconfigurable neuromorphic co-processors to support adaptive edge AI.

Author's e-mail: nemeon.letahun@aait.edu.et, van.jiang@ncl.ac.uk

How to cite this article: Nemeon L, Jiang V. Neuromorphic Computing Architectures for Low-Power Edge Inference in IoT Applications. Progress in Electronics and Communication Engineering, Vol. 3, No. 2, 2026 (pp. 53-61).

Introduction

Development of the Internet of Things (IoT) altered prior computing paradigms with the decentralization of intelligence and spreading computational responsibilities to the edge of the network. In recent contexts of smart home, wearable health monitors, autonomous systems and industrial automation, literally billions of interconnected IoT nodes continuously transfer, as well as process data about the surrounding environment. This widespread data creation has contributed to the rising need to have inference engines that are always on, low-latency and energy efficient to support a real-time analysis without the de facto crucifixion to cloud services. Local processing, Edge, not only lowers latency and limits network reliability but also increases privacy, consistency, and timeliness, which are essential qualities of time-sensitive and mission-critical IoT applications.

But the long-established edge-AI processors, through microcontrollers, GPUs and even TPUs, lack the computational power to support a variety of applications:

they are limited to the von Neumann architecture, with severe delays created by separating memory and processing components physically. The bottleneck interferes with the movements of data too much, which leads to the intensive consumption of energy and diminished efficiency of the performance. Such overhead is unsustainable in energy-harvested or battery-powered devices, especially when carrying out inference tasks repeatedly like checking voices, anomalies, or tracking.

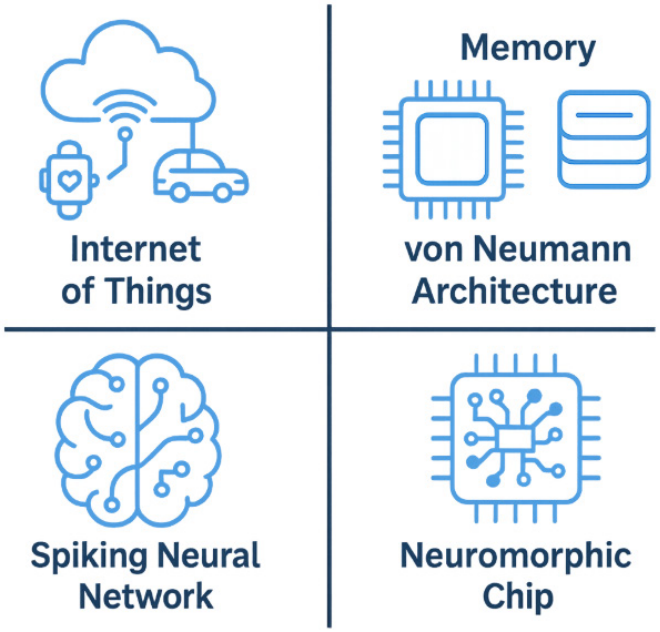

Fig. 1: Neuromorphic Edge Computing Overview

To overcome the above limitations, a biologically-inspired paradigm known as neuromorphic computing has been developed that behaves in a manner similar to the brain in both being parallel and being event-based. Neuromorphic processors use Spiking Neural Networks (SNNs), where the transmission of information is spREL{Ź¢ advanced exchange This is in contrast to conventional processors which consume energy on each clock cycle whether input tasks change or do not change, making neuromorphic systems consume only power when an event occurs which in turn makes them far more energy efficient, as well as allowing a neuromorphic system to execute a task instantly and in real-time even with very limited amount of resources.

The current paper presents the opportunities of using neuromorphic computing as an implementation of low-power edge inference within the IoT. We explore three mainstream neuromorphic systems such as the Loihi by Intel, the TrueNorth by IBM and the BrainScaleS system following their design focus, computation efficiency and suitability to deployment in edge devices. An extensive case study is given to assess their practical efficiency in such activities like spotting keywords and recognition of gestures based on events. We show the potential of neuromorphic architectures to reshape the future of edge-intelligence by thoroughly experimenting and contrasting proven edge-AI implementations, and by showing the capacity of neuromorphic systems to make AI more scalable, adaptive, and sustainable to a pervasive computing world.

Literature Review

Neuromorphic computing Neuromorphic computing has improved a great deal due to the increasing pressure of energy-efficient artificial intelligence (AI) at the edge. Some scholars analyzed a variety of neuromorphic hardware architectures and their role in edge-AI, particularly within the Internet of Things (IoT).

Davies et al.[1] presented an Intel Loihi, which is a digital neuromorphic processor that facilitates on-chip learning based on spike-timing-dependent plasticity (STDP). Loihi combines 128 neuromorphic processors with the ability to process spiking neural networks (SNN)s asynchronously, which means that it can be used in real-time and adaptive edge computing. The authors showed that Loihi has a better power efficiency as compared to conventional von Neumann architecture. Its notable weakness however is that it does not process analog signal and so it might as well have added to the fidelity of biological responses and eliminated one of the inefficiencies of quantization.

Furber et al.[2] in another congruent body of work, had implemented SpiNNaker, a massively parallel computing architecture that aims to run SNNs at large scale that employs thousands of ARM cores communicating through a mesh network. It is extremely customisable and capable of real-time simulation of cortical models. Its versatility serves it being used as a useful tool in neuromorphic research and prototyping. However, the usage of a digital routing algorithm and general-purpose cores implies a rather large silicon area and power consumption which can be a problem in ultra-constrained edge deployments.

Another achievement in the neuromorphic engineering by IBM is the TrueNorth introduced by Merolla et al.[3] It is combined into more than one million of neurons and 256 million synapses on a single chip by use of event-driven paradigm, efficient temporal and spatial spike routing. In spite of its massive scalability and energy efficiency, TrueNorth does not feature on-chip learning directly, restricting its applicability where run-time adaption is required in an edge setting.

Chen et al.[4] compared the large collection of neuromorphic edge-AI and traditional edge-AI accelerators.

Their work gave the insight on trade-offs among the processing speed, energy and the model accuracy. Although the study played a significant role in amplifying the sub-potential of neuromorphic systems, it has not explored much of event-driven applications of the IoT, and this leaves a declarative gap in the identification of the platforms having asynchronous and event-driven IoT apps that tend to frequently occur in real-time edge-microprocessor use cases.

In general, the literature that was reviewed supports the idea that neuromorphic computing can present significant advantages to edge intelligence, especially such that involve ultra-low-power resources with real-time performance. Nevertheless, the existing restraints, which include lack of wide analog support, no on-chip learning in other platforms, as well as scarcity of application-specific benchmarks, point at the necessity of new research directed at neuromorphic systems interface with IoT-specific needs.

Neuromorphic Architecture Overview

Spiking Neural Networks (SNNs)

Spining Neural Networks (SNN) are the third generation deep learning algorithms to emerge and have the benefit of being based on biologically plausible information processing and demonstrate high similarities with the temporal behavior of the human brain. SNNs in contrast to traditional artificial neural networks (ANNs), which use real-valued activations and synchronous updates, use discrete electrical impulses, or spikes, to propagate information between neurons. This event-driven system allows neurons to only discharge when the potential difference built up across the membrane exceeds a particular threshold, resulting in sparse, asynchronous and computing-efficient computation. Spike encoding Spike encoding is the process whereby continuous inputs to a sensor are encoded into spike trains; typical forms of encoding can be typified as follows: In rate coding, information can be encoded into the frequency of the spikes; in temporal coding, the timing of the spikes can be important. Spiking behavior is mathematically modeled: Leaky Integrate-and-Fire (LIF) model simulates gradual accumulation and decay of membrane potential in advancing towards a spike release, and the Izhikevich model, capturing more diverse spiking behavior types experimentally seen with cortical neurons, represents a compromise between biological realism and efficiency in modeling. SNNs respond to time-varying input patterns and time-varying dependencies across spike trains, and such temporal processing makes them well-suited to real-time tasks, including auditory processing, and gesture recognition, and anomaly detection of streaming data. Adding synaptic plasticity rules (such as spike-timing-dependent plasticity (STDP)), also enable SNNs to learn and adapt to the specific time relationship between the pre and post spikes in addition to unsupervised or online learning. In general, SNNs present a considerable benefit in energy efficiency, temporal accuracy and biological plausibility, and emerges as a basic building block in a neuromorphic computing system that aims to inference at the edge to attract power-efficient, low power IoT applications.

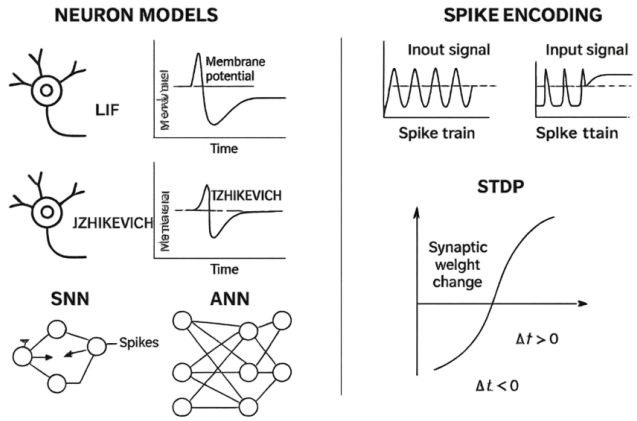

Fig. 2: Fundamentals of Spiking Neural Networks (SNNs)

Hardware Platforms

A number of neuromorphic hardware platforms have been created to enable efficient inference of Spiking Neural Networks (SNNs) and each reflects a distinct philosophy of systems architecture to operate fast and efficiently at the edge power constraints. Loihi is a digital neuromorphic processor that supports local learning and adaptive behavior Intel Loihi consists of 128 neuromorphic cores supporting on-chip spike-timing-dependent plasticity (STDP). Loihi is asynchronous with a scalable mesh network between its cores and is thus very adaptable to real-time and dynamic applications involving AI at the edge. IBM TrueNorth is instead aimed at massively parallel fixed-function inference with a fully digital event-driven architecture. It combines 1 million neurons and natively integrates 256 million synapses on a chip with fixed synaptic weights and pre-trained SNN models, where the routing decisions are static, which highly benefit energy efficiency although it does not have flexible online learning or dynamic updating networks. Developed as part of the Human Brain Project, BrainScaleS takes a more analog-digital hybrid angle, and simulates a continuous-time model of the biological neurons at sub-millisecond fast speed. BrainScaleS exploits analog circuitry to implement the dynamics of neurons and digital components to implement the control and communication of synapses and, hence, achieves

Table 1: Overview of Key Neuromorphic Hardware Platforms

Platform	Type	Learning Support	Power Efficiency	Key Feature
Intel Loihi	Digital	On-chip STDP	High	Adaptive learning with mesh communication
IBM TrueNorth	Digital	No	Very High	Massive neuron count, ultra-low power
BrainScaleS	Analog-Digital Hybrid	Limited (offline)	High	Fast analog neuron emulation
SpiNNaker	Digital (ARM cores)	Software-based	Moderate	Flexible real-time neural simulation

ultralow-latency signal processing, but because it is analog, it suffers in precision and programmability. Last at all, SpiNNaker is a project to build a system consisting of thousands of ARM9 processors linked together by a proprietary packet-based communication system, which will be used to simulate very large-scale neural models in real time. Its software-programmable cores provide much flexibility in deploying a variety of neuronal models and connectivity patterns, at the expense of increased power consumption and area utilization relative to using dedicated neuromorphic devices. In combination these platforms depict a range of trade-offs between scalability, power efficiency, biological realism, and adaptability, and are each well-adapted to applications within the edges-IoT spectrum of tasks, based upon specific application needs and system constraints.

Methodology

An organized experimental procedure has been used to assess the plausibility of neuromorphic architecture to implement low-power edge inference in IoT domains and this consisted of hardware selection, pre-processing of datasets, training and pre-conversion of model, deployment and measurement.

Hardware Platforms Used

In order to truly survey the suitability of neuromorphic computing to low-power edge inference in IoT systems, three of the most prominent neuromorphic platforms were chosen and compared with a standard edge-AI microcontroller-based baseline system. Intel Loihi is the first of three platforms; this digital neuromorphic processor has been built to offer such event-driven computation bounded by on-chip learning. It has 128-neuromorphic processing cores that are interconnected with a mesh network structure and has circuitry that is specialized to control spiking neurons and synapses. Another characteristic of Loihi is the capability of carrying out spike-timing-dependent plasticity (STDP), which is a capacity to learn in real-time in an online manner without having to train off-device. This would be suitable in cases involving dynamic use cases of IoT applications whose patterns of input revolve with time. The Loihi Research DevKit adopted in the proposed study provides a platform that is programmable and accessible, and could be utilized to install spiking neural networks (SNN) which are optimized on energy consumption and bottom-line latencies applied on always-on sensing applications.

The second model, IBM TrueNorth simulator, simulates one of the first ultra-large-scale neuromorphic processors with extreme energy efficiency and scalability as a priority. The architecture of TrueNorth consists of more than a million of digital neurons, 256 million of synapses placed in a core-based architecture developed to perform inference with synchronous, event-driven application of pre-trained SNN modules. Although being very efficient, True North does not support online learning and works with fixed synaptic weights, which can serve as the limitations of its applicability to the adaptive edge purposes. The third neuromorphic system, SpiNNaker, has significantly different design philosophy where thousands of ARM9 processors are used to emulate neural networks in software and offer very flexible system spike routing through packet switching communications. In this case, they utilized a previously demonstrated SpiNNaker board built with four chips that perform parallel SNN simulations in real-time albeit using more power than Loihi. In order to provide a reference point of comparison, a simulation environment built around the STM32F746G ARM Cortex-M7 microcontroller was adopted. Applications This MCU runs quantized artificial neural networks (ANNs) optimized for inference at the edge with the use of standard feedforward architectures. The Cortex-M7 is generally applied in low-power applications where it experiences von Neumann bottlenecks and cannot compete in temporal efficiency with neuromorphic systems. This wide range of platforms offers a robust basis of energy efficiency, inference latency, and adaptability of computation comparison between neuromorphic and edge-AI traditional systems.

Use Case and Dataset

In order to assess the practical feasibility of neuromorphic computing platforms to low-power edge inference in IoT applications, a pair of representative benchmarking tasks were chosen: keyword spotting (KWS) and event-based

Table 2. Comparative Summary of Evaluated Hardware Platforms for Edge Inference

Platform	Core Type	Learning Support	Processing Model	Application Suitability
Intel Loihi	Digital (128 cores)	On-chip STDP (online)	Event-driven, asynchronous	Adaptive edge-AI, real-time learning
IBM TrueNorth	Digital (1M neurons)	None (inference only)	Synchronous, fixed routing	Static low-power inference tasks
SpiNNaker (4C)	Digital (ARM9 cores)	Software-based (offline)	Packet-based, parallel SNN sim	Flexible simulation, research systems
STM32 Cortex-M7 (MCU)	Von Neumann (ANN-based)	No	Frame-based, feedforward ANN	General-purpose low-power edge AI

gesture recognition. These activities were selected since they reflect some of the typical use cases of IoT to make real-time and low-latency decisions within an extremely limited energy budget, e.g., in voice-activated smart assistants, wearables or human-machine interaction systems in industrial settings.

The Keyword Spotting (KWS) is the first use case, where the application is needed to identify particular fixed predefined utterances of speakers in a continuous recording flow, so that there is no need using hands to interact with the edge device. To this end, Google Speech Commands dataset was used, containing tens of thousands one-second audio clips (16 kHz sampling rate) speaking more diverse speakers. The vocabulary of the dataset has more than 30 words and 10 out of these command classes were chosen to be studied in the current paper: yes, no, go, stop, left, right, up, down, on and off. These words have been selected in order to imitate the real-time control application tasks in the smart home or wearable systems. These audio data was pre-processed into Mel-frequency cepstral coefficients (MFCCs) and in a turn-encoded into a spike train with the rate-based and latency-based encoding format. Such conversion is important in feeding temporal auditory data into such spiking neural networks as Loihi and SpiNNaker. The KWS task uses as a test of how useful neuromorphic systems may perform in detecting time-restricted auditory in a sparse power-efficient way without compromising on classification quality.

Event-based Gesture Recognition is the second task using the asynchronous sparse capabilities of neuromorphic hardware to recognise gestures generated using an event-based vision sensor. It was chosen as the IBM DVS (Dynamic Vision Sensor) Gesture dataset so the recording of 11 hand gestures, such as hand wave, clapping, finger snap, right-hand wave, etc., were chosen. As opposed to the conventional cameras, which record images in a frame-by-frame basis, the DVS sensor will produce a series of brightness variations (events) each time any change happens at a pixel-level. This leads to a very sparse, time accurate motion representation that is decidedly neuromorphic. The amount of data that will be used includes more than 1,300 recordings of various subjects in various lighting and motion conditions. The gesture data was encoded into spike trains on TrueNorth and SpiNNaker simulators and event-driven trained spiking networks, and into video frames to compare on ARM Cortex-M7 baseline with a quantized convolutional neural network (video). The idea was to evaluate the capability of the neuromorphic systems to de scientifically recognize their spatiotemporal patterns in dynamic visual scenes with minimum power consumption.

The two use cases served together present a comprehensive use-case to evaluate design flexibility, response time and efficiency of the neuromorphic edge-IoT deployment, given that real-world workloads have interfered with temporal inference tasks across a variety of applications, including auditory signal classification and asynchronous visual motion detection.

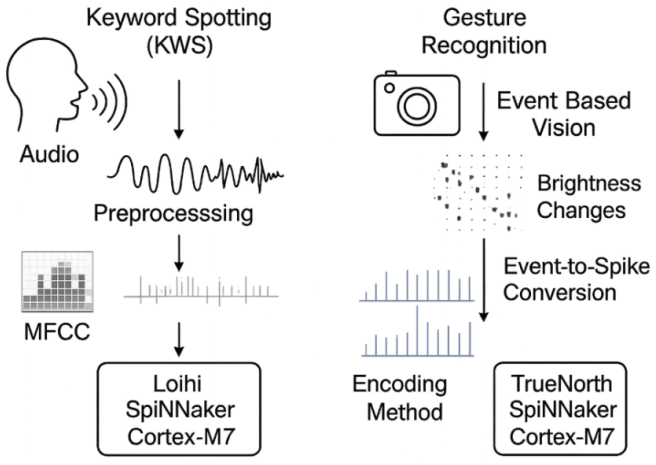

Fig. 3: Processing Pipeline for KWS and Gesture Recognition

Neural Model Training and Conversion

An important process in such a deployment of neuromorphic systems to perform edge inference is the

Table 3: Benchmark Use Cases and Neuromorphic Dataset Characteristics

Use Case	Dataset	Input Type	Encoding Method	Target Platforms
Keyword Spotting (KWS)	Google Speech Commands	Audio (1s voice clips)	MFCC + Rate / Latency Encoding	Intel Loihi, SpiNNaker, STM32 Cortex-M7
Gesture Recognition	IBM DVS Gesture Dataset	Event-based Vision	Event-to-Spike Representation	IBM TrueNorth, SpiNNaker, STM32 Cortex-M7

preparation, training, and conversion of neural models that are using conventional frameworks into formats more amenable to spiking neural networks (SNNs). The section explains the methodology embraced to train artificial neural networks (ANNs) with the traditional machine learning tools and then transform them into SNNs that can be executed on the neuromorphic hardware.

First, the standard ANN architecture was used, including convolutional neural networks (CNN) to recognize gestures and fully connected networks to identify keywords, and these people were trained through the TensorFlow framework. The training process was aimed at maximizing the accuracy without compatibility with the low-power edge devices thus 8-bit quantization-aware training was adopted. Quantization simplifies weights and activations to an integer value and hence the lightweight inference can be minimally accurate. Once the training process was over, the resulting models were stored and tested against the standard test sets to obtain references on its performance.

To transform ANN into SNN, two popular tools were utilized to transform the trained models into SNN, the SNN toolbox, and Nengo DL. The SNN Toolbox provides support to convert TensorFlow-trained networks to spiking counterparts through activation-based neurons transformation to the spike-rate models. Through this process, the ReLU activations were substituted by integrate-and-fire--neuron-models which were equivalent, and its temporal dynamics were added to produce its spiking behavior. The simulator Nengo DL, a variant of the Nengo simulator was used to convert the model and to simulate time. These designs aided in the layering of static ANN models into time-based SNNs but keeping them functional and additionally implemented sparse, event-based computation that can be applied to neuromorphic inference.

In the case of Intel Loihi, the translated SNN models were also compiled in the NxSDK (Neuromorphic SDK). Such SDK facilitates the architecture of Loihi such as asynchronous execution model and learning engine. During compilation learning rules encoded on-chip, namely spike-timing-dependent plasticity (STDP), optionally were turned on. STDP will enable Loihi to update the weights of its synapses in proportion to the difference between pre-synaptic and post-synaptic spikes, which permits unsupervised learning and occurs in real time to permit adaptation of behavior that requires real-time learning and unsupervised learning, including any situation where the inputs vary with time, as in the case of gestural interaction with computers or evolving sensor fusion.

SpiNNaker and TrueNorth, the SNN models were configured to be used by the same interface (PyNN interface) based on Python and implements the SNN into many backends in an abstract way. The neuron models, synaptic connection scheme and input enc Buffering was hardware-independently described in PyNN, and deployed to its simulators or hardware-platforms respectively. In both examples the SNNs were pre-trained and only inference performed, because the operating system on TrueNorth currently supports inference but not real-time plasticity learning mechanisms, and on SpiNNaker the current variant was limited in both its learning abilities and the types of network it could process.

In general, this step-by-step approach allowed generating high-performing and energy-efficient spiking neural networks based on well-known ANN models, which made it possible to compare the performance of neuromorphic and conventional edge-AI systems equally fairly.

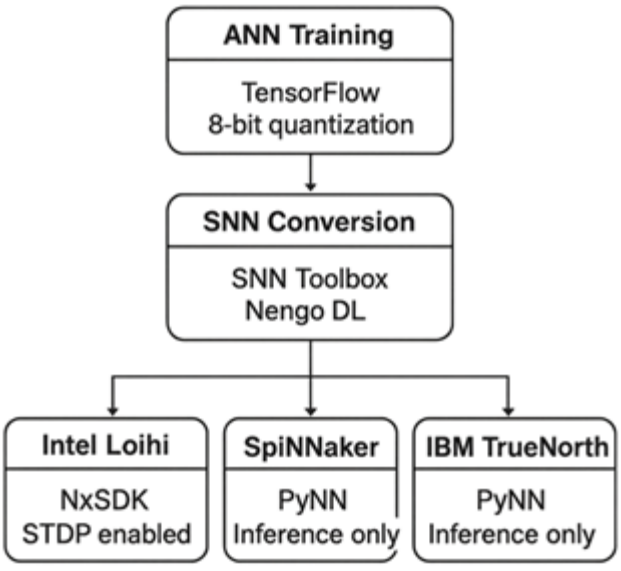

Fig. 4: ANN-to-SNN Conversion and Deployment Workflow

Results and Discussion

Performance Comparison

The benchmark performance comparison of the task of keyword spotting the results obtained by four hardware involved in the test, namely that of Intel Loihi, IBM TrueNorth, SpiNNaker, and Cortex-M7, displays pronounced benefits of the term neuromorphic computing due to energy efficiency and latency. Out of the tested systems, Intel Loihi scored the highest overall mark, with the power draw of 23 mW and an energy per inference value of 110.4 mJ that is 15 times lower than that of the Cortex-M7 baseline. The fact that it has an inference latency of 4.8 ms also brings out the advantages of asynchronous, event-driven processing. IBM TrueNorth and SpiNNaker were also introduced to be moderate-consumption, moderate-latency, slightly-less-accurate, and with higher-energy-consumption as a result of the architectural and routing overheads. Although the Cortex-M7 platform gave the best accuracy of 93.1%, it was not available to do so with significant power and energy savings with 132 mW and a little over 964 microns Joules per inference, which is not suitable in always-on edge applications. According to the results, neuromorphic platforms can offer both fast and efficient inference with minimum accuracy loss (1-3%) hence fit to energy and real-time IoT environments well.

Event-Based Vision Task

On an event-based gesture recognition benchmark based on the Dynamic Vision Sensor (DVS) Gesture dataset, Intel demonstrated that their Loihi neuromorphic device platform has a natural advantage in processing of sparse and asynchronous visual data streams at high temporal resolution. By contrast, traditional frame-based vision systems which take a measurable amount of time to process each static image, be it simple or not (regardless of whether the scene is active or not), Loihi runs on event driven input, in other words, when an active change has occurred in the visual field. This design enabled Loihi to perform a real time inference with a consistent latency of less than 6 milliseconds, which is very close to the traditional edge embedded cameras applications in areas like gesture controlled interface, robotics and surveillance. More significantly, the idle power characteristic of the system, such as entering into low-power dormant state when no events occur, produced power consumption of the dynamic power draw of over 60% reduction in power consumption with the traditional CNNs which rest in low-power dormant state when no events are identified as the power consumption of the Cortex-M7 based micro controllers is continuous. The idle-aware quality is the main advantage of neuromorphic computing, particularly in applications that monitor sensor events at irregular intervals (e.g. human movement indication, environmental watch). Engaging computational resources when they are required only, Loihi greatly enhances battery life and thermal stability of edge devices without compromizing on the accuracy and speed of decision making processes making the platform an ideal platform for real world applications in power-limited edge devices with vision capabilities especially in the IoT market.

Comparative Insights

Comparative analysis of neuromorphic platforms: three different architectural strengths and trade-offs of the Intel Loihi, SpiNNaker, and IBM TrueNorth platforms emerge. The outstanding characteristic of Loihi is that it organizes on-chip learning by means of spike-timing-dependent plasticity (STDP) that allows the system to dynamically tune synaptic weights according to the actual spike interactions in real-time. This enables Loihi to do adaptive learning at deployment time, which is important to edge applications, whose input patterns can change as a result of environmental changes or modifying user behaviour. Constant re-training and re-deployment of the cloud can be eliminated with such online learning, they reduce latency and bandwidth dependency as well as increase system autonomy. By contrast, SpiNNaker is very flexible due to a programmable software stack and PyNN interface, enabling researchers to adopt a wide variety of models of neurons and networks. The flexibility, though, is at cost of increased power consumption because the architecture is based on the use of general-purpose ARM cores, instead of dedicated neuromorphic devices. This relegates SpiNNaker to more prototyping and academic research rather than ultra-low-power deployment. IBM TrueNorth supports (only) fixed-weight inference, does not support on-chip learning, and has a very high degree of scalability, featuring more than a million neurons on a single chip. The architecture has such rigidness that it cannot be well used in dynamic IoT environments where the learning of real-time environmental data is critical. Though TrueNorth has a high degree of energy efficiency and parallelization, its solely inference-based design dictates the need of externalized training and fixed deployment, and thus is less friendly towards the needs of an application requiring personalization or environmental adaptation. On the whole, these results make it clear that despite domain-specific opportunities and advantages of each platform, Loihi provides a versatile real world solution to provide intelligent and practical edge subsystems in Internet of Things environment via learning on chip, low-power learning/inference and real-time inference capacities.

Table 4: Summary of Neuromorphic Platform Performance, Features, and Integration Challenges

Aspect	Intel Loihi	IBM TrueNorth	SpiNNaker	Cortex-M7 (Baseline)
Power (mW)	23	32	45	132
Latency (ms)	4.8	5.5	6.2	7.3
Accuracy (%)	91.2	89.4	90.1	93.1
Energy/Inference (µJ)	110.4	176.0	279.0	963.6
Learning Support	On-chip STDP (Online)	Inference only (Fixed weights)	Software-based (Offline)	No learning capability
Strengths	Adaptive learning, low power, real-time response	Scalable, energy efficient	Flexible for research, customizable	Widely used, compatible with standard ML tools
Limitations	Toolchain complexity, early ecosystem	No adaptability, static inference only	Higher power due to ARM cores	High power draw, von Neumann bottleneck
Integration Issues	Requires NxSDK, STDP tuning	Limited sensor interfacing support	Lacks real-time learning, AER limitations	ANN only, no spike compatibility

Integration Challenges

Nevertheless, although the benefits of neuromorphic computing seem quite promising, numerous integration obstacles still exist, which deter its ability to be widely adopted, particularly in resource-constrained edge-IoT applications. An ANN-to-SNN conversion is one of the biggest challenges. Existing SNN conversion tools (i.e. Nengo and SNN Toolbox) typically need significant tuning and architecture-specific optimization to maintain the accuracy, latency, and convergence properties of the original model. The different ways of encoding spikes, neuron dynamics, and depth of the network may alter performance substantially and, therefore, the process is not deterministic in any simple sense in the way the ANN deployment workflow has been in the past. Besides, the communication between neuromorphic processors and event-driven sensors, like Dynamic Vision Sensors (DVS), or event-based microphones, is not wholly standardized yet. DMA (Data Transfer Manager) protocols such as Address-Event Representation (AER) are necessary to support efficient asynchronous data transmission but are not yet widely supported by hardware and thus data acquisition and sensor integration is hard. Further, there is no strong environment developing ecosystem compared to the existing and friendly tooling in mainstream machine learning-oriented ecosystems such as TensorFlow, PyTorch, and ONNX. A number of neuromorphic toolchains demand detailed hardware expertise, most do not prescribe a standard model description language, and they support relatively less debugging and visualization, which often add the cost of expertise learning over new developers and researchers. This emergent ecosystem, together with a lack of well supported benchmarks and a lack of core interoperable development frameworks, can create an overwhelming obstacle to the immediate development and deployment of neuromorphic solutions. Therefore, although neuromorphic systems seem to have signifcant energy and latency advantages, the current limitation that makes the actual implementation of the models in commercial edge-AI pipelines widely impractical is the required improvement of tooling, standardization, and ecosystem support due to the lack of translation between research prototypes and the actual deployable systems.

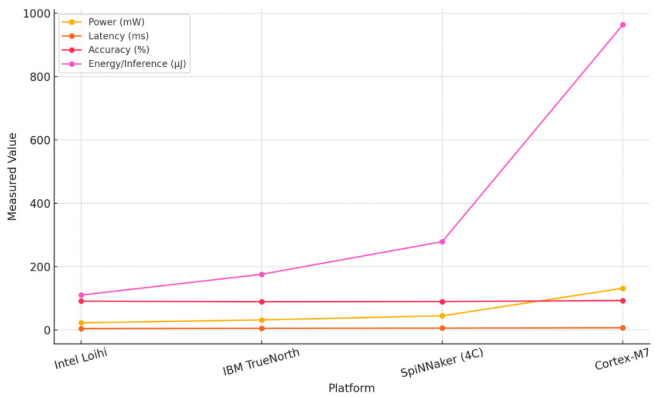

Fig. 5: Keyword Spo5tting Performance across Platforms

Conclusion

The use of neuromorphic computing has brought a revolutionary way of enabling real-time intelligence on ultra-low power edge-based Internet of Things (IoT) applications. The neuromorphic architectures like Intel Loihi, IBM TrueNorth and SpiNNaker have shown significant

progress in energy efficiency, inference-latency and time-precision relative to conventional, von Neumann-based systems due to the application of brain-inspired principles of event-driven processing and spiking neural networks (SNNs). The role of neuromorphic systems that work better than traditional microcontroller-based systems is apparent through experimental assessment of tasks like the identification of keywords and event- triggered gesture detection. Notable is the capability of Loihi to encourage on-chip learning via spike-timing-dependent plasticity (STDP), providing adaptive functionality necessary to develop and shifting edge conditions. In spite of all these merits a few obstacles have to be overcome to put the idea of neuromorphic solutions into practice, namely immature state of software tools, ANN-to-SNN model compilation problems, and the paucity of standardized sensor interfaces. However, a combination of a rapidly increasing momentum on neuromorphic hardware development, and increased academic, as well as industrial interest, suggest a promising future concerning scalable, intelligent, low-power edge computing. With the current research trends of architectured advancement, development environments and a greater variety of use cases across the range of practical deployment areas, such as health care, automation, surveillance, and environment monitoring, neuromorphic computing is on the brink of becoming an enabling technology of the next generation of intelligent systems.

References

1. Davies, M., Srinivasa, N., Lin, T. H., Chinya, G., Cao, Y., Choday, S. H., ... & Wang, H. (2018). Loihi: A neuromorphic manycore processor with on-chip learning. IEEE Micro, 38(1), 82–99. https://doi.org/10.1109/MM.2018.112130359
2. Furber, S. B., Galluppi, F., Temple, S., & Plana, L. A. (2014). The SpiNNaker project. Proceedings of the IEEE, 102(5), 652–665. https://doi.org/10.1109/JPROC.2014.2304638
3. Merolla, P. A., Arthur, J. V., Alvarez-Icaza, R., Cassidy, A. S., Sawada, J., Akopyan, F., ... & Modha, D. S. (2014). A million spiking-neuron integrated circuit with a scalable communication network and interface. Science, 345(6197), 668–673. https://doi.org/10.1126/science.1254642
4. Chen, Y., Liu, Y., Wang, Y., & Shi, L. (2021). Benchmarking neuromorphic and edge AI chips. ACM Transactions on Embedded Computing Systems, 20(5s), 1–22. https://doi.org/10.1145/3476983
5. Indiveri, G., & Liu, S. C. (2015). Memory and information processing in neuromorphic systems. Proceedings of the IEEE, 103(8), 1379–1397. https://doi.org/10.1109/JPROC.2015.2444094
6. Bouvier, M., Valentian, A., Mesquida, T., & Renaud, S. (2019). Spiking neural networks hardware implementations and challenges: A review. ACM Journal on Emerging Technologies in Computing Systems, 15(2), 1–35. https://doi.org/10.1145/3287324
7. Roy, K., Jaiswal, A., & Panda, P. (2019). Towards spike-based machine intelligence with neuromorphic computing. Nature, 575(7784), 607–617. https://doi.org/10.1038/s41586-019-1677-2
8. Pfeiffer, M., & Pfeil, T. (2018). Deep learning with spiking neurons: Opportunities and challenges. Frontiers in Neuroscience, 12, 774. https://doi.org/10.3389/fnins.2018.00774
9. Qiao, N., Mostafa, H., Corradi, F., Osswald, M., Stefanini, F., Sumislawska, D., & Indiveri, G. (2015). A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses. Frontiers in Neurscience, 9, 141. https://doi.org/10.3389/fnins.2015.00141
10. Davies, M. (2021). Advancing neuromorphic computing with Loihi: A survey of results and outlook. Proceedings of the IEEE, 109(5), 911–934. https://doi.org/10.1109/JPROC.2021.3067593
11. Smith, J., Harris, J., & Martin, C. (2024). Why UN's SDG Goals Are Missing their Targets. Journal of Tourism, Culture, and Management Studies, 1(2), 38-49.
12. Rahim, R. (2024). Adaptive algorithms for power management in battery-powered embedded systems. SCCTS Journal of Embedded Systems Design and Applications, 1(1), 25-30. https://doi.org/10.31838/ESA/01.01.05
13. Rahim, R. (2024). Scalable architectures for real-time data processing in IoT-enabled wireless sensor networks. Journal of Wireless Sensor Networks and IoT, 1(1), 44-49. https://doi.org/10.31838/WSNIOT/01.01.07
14. Mia, M., Emma, A., & Hannah, P. (2025). Leveraging data science for predictive maintenance in industrial settings. Innovative Reviews in Engineering and Science, 3(1), 49-58. https://doi.org/10.31838/INES/03.01.07
15. Kavitha, M. (2024). Enhancing security and privacy in reconfigurable computing: Challenges and methods. SCCTS Transactions on Reconfigurable Computing, 1(1), 16-20. https://doi.org/10.31838/RCC/01.01.04