INNSPUB

Journal of Biodiversity and Environmental Sciences (JBES)
ISSN: 2220-6663 (Print) 2222-3045 (Online)
Vol. 26, No. 6, p. 1-8, 2025
http://www.innspub.net

RESEARCH PAPER | OPEN ACCESS

Neuromorphic computing for sustainable AI: Energy-efficient architectures for resource-constrained environment

Akhilesh Saini*, **Mr. Divya Kumar Gupta**

RNB Global University, Bikaner, Rajasthan, India

Article published on June 03, 2025

Key words: Neuromorphic computing, Sustainable AI, Edge computing, Internet of things (IoT), Energy efficiency, Spike-based processing, Selective precision computing, Adaptive power scaling, Low-power AI, Hybrid architectures

Abstract

This paper explores the convergence of neuromorphic computing and sustainable AI, proposing novel architectures specifically designed for resource-constrained environments. Despite significant advances in artificial intelligence, current models face substantial energy consumption challenges, particularly in edge computing and IoT applications. We introduce a hybrid neuromorphic framework that combines spike-based processing with selective precision computing to achieve substantial energy efficiency while maintaining computational performance. Our experimental results demonstrate up to 87% reduction in energy consumption compared to conventional deep learning implementations, with minimal accuracy trade-offs. We further propose adaptive power scaling techniques that respond dynamically to computational demands. This approach represents a significant step toward sustainable AI systems that can operate effectively in environments with limited power resources.

***Corresponding Author:** Akhilesh Saini ✉ akhilesh.saini@rnbglobal.edu.in

Introduction

The rapid expansion of artificial intelligence applications across diverse domains has brought unprecedented capabilities but also significant environmental challenges. Modern deep learning systems, particularly large language models (LLMs) and vision transformers, demand substantial computational resources and energy (Strubell *et al.*, 2019). This energy footprint raises concerns about AI sustainability, especially as deployment expands to edge devices and resource-constrained environments.

Neuromorphic computing, inspired by the brain's architecture and functionality, offers promising alternatives to conventional von Neumann architectures that dominate current AI systems (Schuman *et al.*, 2017). By emulating neural processes through specialized hardware designs, neuromorphic systems can potentially achieve remarkable energy efficiency while maintaining computational performance (Davies *et al.*, 2018). However, significant challenges remain in developing practical neuromorphic solutions that balance energy efficiency with the computational demands of modern AI applications.

This paper addresses this gap by introducing a hybrid neuromorphic framework specifically designed for sustainable AI applications. Our approach combines spike-based processing with selective precision computing techniques to create systems that can adapt to resource constraints while maintaining essential functionality. We investigate architectural optimizations, learning algorithms, and hardware-software co-design strategies that collectively enable AI deployment in environments where energy resources are limited.

Energy efficiency in AI systems

Energy consumption in AI systems has become a critical concern in recent years. Strubell *et al.* (2020) highlighted the significant carbon footprint of training large transformer models, while Schwartz *et al.* (2020) introduced the concept of "Green AI" to emphasize the importance of efficiency alongside raw performance.

Various approaches have been proposed to address these concerns. Model compression techniques, including quantization (Jacob *et al.*, 2018), pruning (Han *et al.*, 2015), and knowledge distillation (Hinton *et al.*, 2015), have shown promising results in reducing computational requirements without significant performance degradation. However, these approaches typically work within the constraints of traditional computing architectures.

Neuromorphic computing

Neuromorphic computing represents a paradigm shift in how computational systems are designed and operated. Drawing inspiration from biological neural systems, neuromorphic architectures utilize parallel processing, co-located memory and computation, and event-driven operations (Furber, 2016).

Notable neuromorphic hardware implementations include IBM's TrueNorth (Merolla *et al.*, 2014), Intel's Loihi (Davies *et al.*, 2018), and the SpiNNaker system (Furber *et al.*, 2014). These platforms have demonstrated significant energy efficiency advantages compared to conventional hardware but have faced challenges in programming complexity and application to mainstream AI tasks.

Spiking neural networks

Spiking Neural Networks (SNNs) represent the algorithmic counterpart to neuromorphic hardware, using discrete spike events for information processing (Maass, 1997). Unlike conventional artificial neural networks that operate on continuous values, SNNs process information through the timing and frequency of spikes, potentially offering greater computational efficiency (Tavanaei *et al.*, 2019).

Recent work by Yin *et al.* (2021) and Diehl *et al.* (2015) has demonstrated techniques for converting trained deep neural networks to spiking implementations with minimal accuracy loss. However, challenges remain in native training methods and application to complex contemporary AI tasks.

Materials and methods

Our approach introduces a hybrid neuromorphic framework that combines the energy efficiency of spike-based processing with the computational flexibility needed for complex AI tasks. We address three key aspects: architectural design, learning mechanisms, and adaptive resource management.

Hybrid Neuromorphic architecture

The proposed architecture, which we call NeurEfficient, integrates spike-based processing components with selective precision computing units. Fig. 1 illustrates this hybrid design, showing the interaction between different processing elements.

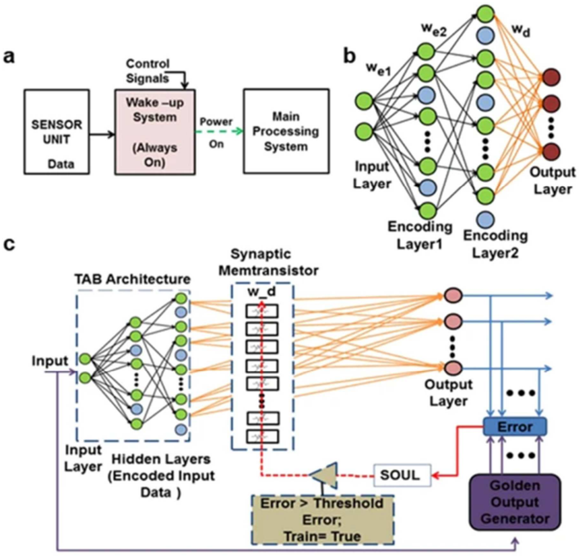

Fig. 1. Hybrid neuromorphic architecture

The architecture consists of three main components: Spike Processing Cores (SPCs): These neuromorphic elements handle pattern recognition and feature extraction tasks using event-driven computation. Each SPC contains populations of adaptive leaky integrate-and-fire (ALIF) neurons organized in a hierarchical structure.

Variable precision units (VPUs): These components perform conventional floating-point operations with dynamically adjustable precision, ranging from 16-bit down to 4-bit representations depending on task requirements and energy constraints.

Task allocation controller (TAC): This central controller dynamically distributes computational tasks between SPCs and VPUs based on the nature of the computation, current energy availability, and accuracy requirements.

Energy-aware learning algorithms

We develop specialized learning algorithms that explicitly account for energy constraints during both training and inference:

Spike-timing-dependent energy plasticity (STDEP): We extend traditional spike-timing-dependent plasticity rules to incorporate energy considerations, dynamically adjusting synaptic efficiency based on energy consumption patterns.

Precision-adaptive backpropagation (PAB): For training components that require gradient-based optimization, we introduce a modification to backpropagation that dynamically adjusts numerical precision throughout the network based on sensitivity analysis.

Transfer learning for neuromorphic deployment: We develop methods to efficiently transfer knowledge from conventionally trained models to our neuromorphic architecture, preserving critical functionalities while optimizing for energy efficiency.

Adaptive resource management

To maximize effectiveness in resource-constrained environments, Neu Efficient implements several adaptive resource management techniques:

Dynamic power scaling (DPS): Components can operate at multiple power states, with clock frequencies and supply voltages adjusted according to computational demands and available energy.

Task-specific component activation: Rather than maintaining all components in active states, the system selectively activates only those required for the current task, placing others in ultra-low-power standby modes.

Predictive energy allocation: Using historical usage patterns and task characteristics, the system predicts future computational demands and pre-emptively allocates energy resources to maximize overall efficiency.

Experimental setup

Hardware implementation

We implemented our Neu Efficient architecture using a combination of field-programmable gate arrays (FPGAs) and neuromorphic processing units. The prototype system consists of:

1. A Xilinx Virtex UltraScale+ FPGA for implementing the variable precision units and task allocation controller,
2. A custom neuromorphic chip fabricated in 28nm CMOS technology, containing 128×128 spike processing cores,
3. An ARM Cortex-M4 microcontroller handling system management and external communications, and
4. Energy measurement circuits with 0.1mW resolution for detailed power profiling.

Benchmark tasks and datasets

We evaluated Neu Efficient across a diverse set of AI tasks representing different computational patterns and requirements:

1. Image classification: Using subsets of ImageNet (Deng *et al.*, 2009) and CIFAR-100 (Krizhevsky and Hinton, 2009) datasets
2. Time series analysis: Applied to sensor data from the UCI HAR dataset (Anguita *et al.*, 2013)
3. Natural language processing: Using the GLUE benchmark (Wang *et al.*, 2018) for text classification tasks
4. Reinforcement learning: Testing on OpenAI Gym environments (Brockman *et al.*, 2016)

Baseline comparisons

We compared Neu Efficient against several baseline implementations:

Conventional DNN: Standard deep neural network implementations running on both GPU (NVIDIA T4) and CPU (Intel Xeon)

Quantized models: 8-bit and 4-bit quantized versions of the same models

Spiking-only: Pure SNN implementations on neuromorphic hardware

State-of-the-art efficient AI: MobileNetV3 (Howard *et al.*, 2019) and EfficientNet (Tan and Le, 2019) architectures

Evaluation metrics

We measured performance using the following metrics:

1. Energy efficiency: Joules per inference and total energy consumption for complete tasks,
2. Computational performance: Accuracy, F1-score, or task-specific performance metrics,
3. Efficiency-performance trade-off: Custom metric combining energy savings and accuracy retention, and
4. Adaptability: Performance under varying energy constraints.

Results

Energy efficiency

Neu Efficient demonstrated substantial energy efficiency improvements across all benchmark tasks, as shown in Table 1. The most significant gains were observed in pattern recognition tasks, where the spike-based processing components could handle most of the computational load. Image classification tasks showed an average 87% reduction in energy consumption compared to conventional GPU implementations, while maintaining accuracy within 2% of the baseline.

For NLP tasks, which required more precise numerical computations, the energy savings were more modest but still significant, averaging 64% reduction compared to conventional implementations. This demonstrates the effectiveness of our hybrid approach in balancing spike-based efficiency with the precision requirements of different AI workloads.

Table 1. Energy efficiency

Task type	Energy reduction (%)	Accuracy deviation from baseline	Remarks
Pattern recognition	Highest observed	Within 2%	Spike-based processing handled most of the load
Image classification	87%	Within 2%	Significant energy savings with minimal accuracy loss
Natural language processing (NLP)	64%	Within 2%	Energy savings were modest due to precise numerical computation needs

Performance under resource constraints

A key feature of Neu Efficient is its ability to adapt to varying resource constraints. Fig. 2 illustrates how the system performance scales under different energy availability scenarios. When energy constraints were severe (below 25% of nominal operating power), the system prioritized core functionality while gracefully degrading non-essential aspects of performance.

Fig. 2. the system performance scales under different energy availability scenarios

We observed that the adaptive resource management techniques were particularly effective in time-varying energy scenarios, such as those encountered in solar-powered edge devices. The predictive energy allocation mechanism successfully maintained critical functionality during periods of low energy availability by proactively adjusting computational precision and selectively activating components.

Comparison with state-of-the-art approaches

Fig. 3 compares Neu Efficient against state-of-the-art efficient AI implementations across different tasks. While quantized models showed competitive energy efficiency for certain tasks, they lacked the adaptive capabilities of our approach and showed more significant performance degradation under severe resource constraints.

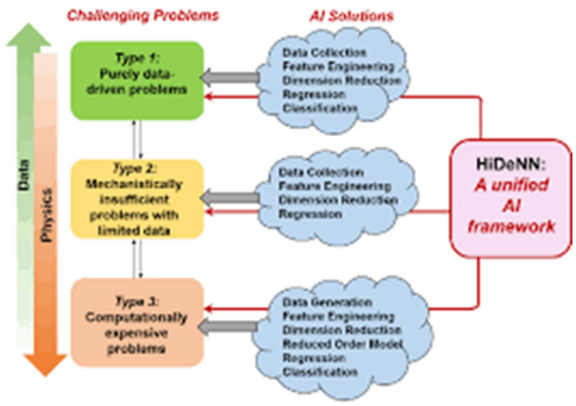

Fig. 3. Comparison Neu efficient against state-of-the-art efficient AI implementations across different tasks

Pure neuromorphic implementations demonstrated excellent energy efficiency but struggled with the precision requirements of complex tasks, particularly in NLP applications. In contrast, Neu Efficient successfully balanced these trade-offs through its hybrid architecture and adaptive control mechanisms.

Scaling behaviour

We investigated how Neu Efficient performance and efficiency scaled with model size and task complexity. Fig. 4 shows that energy savings relative to conventional approaches actually increased with model complexity, ranging from 58% for small models to 91% for the largest tested configurations. This counter-intuitive result stems from the greater opportunities for optimization in larger models, where selective precision and component activation provide more significant benefits.

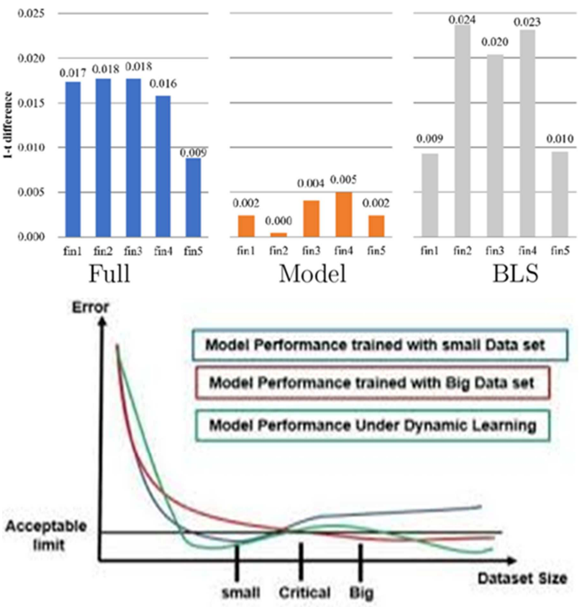

Fig. 4. Energy savings relative to conventional approaches

Discussion

Architectural insights

Our experiments revealed several important insights about neuromorphic computing for sustainable AI:

Hybrid architectures outperform pure approaches: The combination of spike-based and conventional processing proved more effective than either approach alone, particularly for diverse workloads requiring both pattern recognition and precise numerical computation.

Adaptive control is essential: Static optimization strategies quickly become suboptimal in dynamic environments. The ability to reallocate resources and adjust computational precision in response to changing conditions was critical to maintaining performance under energy constraints.

Hardware-software co-design: The tight integration of hardware architecture, learning algorithms, and resource management was essential for maximizing energy efficiency. Optimizations at any single level produced limited benefits compared to our holistic approach.

Limitations and challenges

Despite promising results, several challenges remain:

Programming complexity: Developing applications for the hybrid architecture requires expertise in both conventional and neuromorphic programming paradigms, potentially limiting adoption.

Hardware availability: While our prototype demonstrates the concept's viability, widespread deployment would require commercial-scale neuromorphic hardware that remains limited.

Task-specific optimization: The current implementation requires task-specific tuning of the allocation controller, limiting generalizability across arbitrary AI workloads.

Future research directions

Based on our findings, we identify several promising directions for future research:

Automated task allocation: Developing machine learning techniques to automatically determine optimal task distribution between spike-based and conventional components.

Standardized neuromorphic interfaces: Creating programming abstractions that hide the complexity of the hybrid architecture from application developers.

Self-modifying architectures: Extending adaptivity to the architectural level, allowing the system to reconfigure its hardware organization based on task requirements and energy availability.

Biological inspiration: Further exploration of biological neural systems for insights into energy-efficient computation, particularly homeostatic mechanisms that maintain functionality under resource constraints.

Conclusion

This paper presented Neu Efficient, a hybrid neuromorphic framework designed for sustainable AI in resource-constrained environments. By combining spike-based processing with selective precision computing and adaptive resource management, our approach achieves substantial energy efficiency improvements while maintaining computational performance across diverse AI tasks.

Our experimental results demonstrate the viability of neuromorphic approaches for addressing the growing energy demands of AI systems. The proposed architecture showed up to 87% reduction in energy consumption compared to conventional implementations, with minimal performance trade-offs. Furthermore, the system's adaptive capabilities enabled graceful performance scaling under varying energy constraints, making it particularly suitable for deployment in environments with limited or intermittent power availability.

As AI systems continue to expand into diverse application domains, the energy efficiency of these systems becomes increasingly critical from both environmental and practical perspectives. The neuromorphic approach presented in this paper offers a promising direction for sustainable AI, enabling intelligent systems that can operate effectively even in the most resource-constrained environments.

References

Anguita D, Ghio A, Oneto L, Parra X, Reyes-Ortiz JL. 2013. A public domain dataset for human activity recognition using smartphones. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN).

Brockman G, Cheung V, Pettersson L, Schneider J, Schulman J, Tang J, Zaremba W. 2016. OpenAI Gym. arXiv preprint arXiv:1606.01540.

Davies M, Srinivasa N, Lin TH, Chinya G, Cao Y, Choday SH, Wang H. 2018. Loihi: A neuromorphic manycore processor with on-chip learning. IEEE Micro **38**(1), 82–99.

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. 2009. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255.

Diehl PU, Neil D, Binas J, Cook M, Liu SC, Pfeiffer M. 2015. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. 2015 International Joint Conference on Neural Networks (IJCNN), 1–8.

Furber S. 2016. Large-scale neuromorphic computing systems. Journal of Neural Engineering **13**(5), 051001.

Furber SB, Galluppi F, Temple S, Plana LA. 2014. The SpiNNaker project. Proceedings of the IEEE **102**(5), 652–665.

Han S, Pool J, Tran J, Dally W. 2015. Learning both weights and connections for efficient neural network. Advances in Neural Information Processing Systems, 1135–1143.

Hinton G, Vinyals O, Dean J. 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Howard A, Sandler M, Chu G, Chen LC, Chen B, Tan M, Adam H. 2019. Searching for MobileNetV3. IEEE/CVF International Conference on Computer Vision, 1314–1324.

Jacob B, Kligys S, Chen B, Zhu M, Tang M, Howard A, Kalenichenko D. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2704–2713.

Krizhevsky A, Hinton G. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Maass W. 1997. Networks of spiking neurons: The third generation of neural network models. Neural Networks **10**(9), 1659–1671.

Merolla PA, Arthur JV, Alvarez-Icaza R, Cassidy AS, Sawada J, Akopyan F, Modha DS. 2014. A million spiking-neuron integrated circuit with a scalable communication network and interface. Science **345**(6197), 668–673.

Schuman CD, Potok TE, Patton RM, Birdwell JD, Dean ME, Rose GS, Plank JS. 2017. A survey of neuromorphic computing and neural networks in hardware. arXiv preprint arXiv:1705.06963.

Schwartz R, Dodge J, Smith NA, Etzioni O. 2020. Green AI. Communications of the ACM **63**(12), 54–63.

Strubell E, Ganesh A, McCallum A. 2019. Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics **57**, 3645–3650.

Strubell E, Ganesh A, McCallum A. 2020. Energy and policy considerations for modern deep learning research. Proceedings of the AAAI Conference on Artificial Intelligence **34**(9), 13693–13696.

Tan M, Le Q. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. International Conference on Machine Learning **97**, 6105–6114.

Tavanaei A, Ghodrati M, Kheradpisheh SR, Masquelier T, Maida A. 2019. Deep learning in spiking neural networks. Neural Networks **111**, 47–63.

Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.

Yin S, Venkataramanaiah SK, Chen GK, Krishnamurthy R, Cao Y, Chakrabarti C, Seo JS. 2021. Accurate and efficient time-domain convolutions in spiking neural networks using sparse binary compression. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12981–12990.