МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ХАРЧОВИХ ТЕХНОЛОГІЙ

ВСП «ПОЛТАВСЬКИЙ ФАХОВИЙ КОЛЕДЖ НУХТ»

KHARKIV IT CLUSTER

Перша всеукраїнська

науково-практична конференція

«Комп’ютерна інженерія: сучасний стан та особливості підготовки фахівців»

15 – 16 квітня 2026 р.

Полтава ВСП ПФК НУХТ 2026

УДК 004

Наукові праці Першої всеукр. наук.-практ. конф. «Комп’ютерна інженерія: сучасний стан та особливості підготовки фахівців» (CE-2026), 15-16 квітня 2026 р. (Полтава, Україна). П. : ВСП ПФК НУХТ, 2026. 163 с.

У працях конференції наведено доповіді за напрямами:

- сучасні тенденції, технології та архітектурні рішення у сфері комп'ютерної інженерії та інформаційних технологій
- інноваційні педагогічні технології та моделі формування професійних компетентностей фахівців з комп’ютерної інженерії та інформаційних технологій
- екосистема розвитку комп’ютерної інженерії: індустріальні практики, інновації та інтеграція науки, освіти і бізнесу

Праці конференції будуть корисні науковим та інженерно-технічним працівникам, студентам ЗПФО, ЗВО та всім, хто цікавиться сучасними інформаційними системами та телекомунікаційними технологіями.

Подано в авторській редакції.

Автори матеріалів несуть повну відповідальність за достовірність наведеної інформації та відповідність матеріалів нормам законодавства, моралі й етики.

ISBN

УДК 004

OPTIMIZATION OF LARGE LANGUAGE MODEL ENERGY CONSUMPTION THROUGH NEUROMORPHIC HARDWARE INTEGRATION

Hrama M.
National University of Food Technologies, Kyiv, Ukraine
E-mail: gramamp@nuft.edu.ua

Optimization of Large Language Model Energy Consumption through Neuromorphic Hardware Integration

The rapid advancement of Large Language Models (LLMs) has ushered in a new era of artificial intelligence characterized by unprecedented natural language understanding and generation capabilities. However, this progress is tethered to a significant environmental and economic cost due to the astronomical energy requirements of training and deploying these models on traditional Von Neumann architectures. This thesis explores the transition from standard graphical processing units to neuromorphic hardware as a primary strategy for energy optimization. By leveraging event-driven processing, spiking neural networks, and non-colocated memory-logic structures, neuromorphic engineering offers a path toward sustainable AI. The discussion centers on the architectural synergy between the attention mechanisms of transformers and the temporal dynamics of neuromorphic systems, proposing a framework for hybrid integration that maintains computational precision while drastically reducing the carbon footprint of industrial-scale machine learning.

The current trajectory of artificial intelligence development is increasingly defined by a critical tension between the scaling laws of deep learning and the thermodynamic limits of silicon-based hardware. Large Language Models, which now encompass hundreds of billions of parameters, require specialized data centers that consume megawatts of power, rivaling the energy demands of small metropolitan areas. The fundamental bottleneck resides in the Von Neumann architecture where the physical separation of the central processing unit and the memory subsystem necessitates constant data shuttling. This movement of data, rather than the computation itself, accounts for the majority of energy dissipation in modern LLM inference. As we move toward the horizon of trillion-parameter models, the reliance on traditional accelerators like GPUs and TPUs appears increasingly unsustainable, necessitating a paradigm shift toward hardware that mimics the biological efficiency of the human brain.

Neuromorphic engineering represents the most promising alternative to this energy crisis by fundamentally reimagining how information is processed. Unlike traditional digital systems that operate on high-frequency clock cycles and continuous data streams, neuromorphic chips utilize event-driven communication. In this architecture, processing only occurs when a specific threshold is met, resulting in "spikes" of activity. This sparsity is the key to energy efficiency. When applied to the dense matrix multiplications characteristic of LLMs, neuromorphic hardware allows for the suppression of redundant computations where input values are zero or negligible. Since the vast majority of activations in a fine-tuned transformer model are sparse, a neuromorphic implementation

can theoretically reduce energy consumption by several orders of magnitude without sacrificing the structural integrity of the model's knowledge base.

The transition of LLMs to neuromorphic substrates involves a complex process of converting standard artificial neural networks into Spiking Neural Networks or utilizing hybrid analog-digital systems. The attention mechanism, which is the core of the Transformer architecture, relies on calculating global dependencies across a sequence. In a neuromorphic context, this can be re-engineered through temporal coding, where the importance of a word or token is represented by the timing of a spike rather than its absolute magnitude. By utilizing the intrinsic dynamics of silicon neurons to perform these calculations, the energy cost per synaptic operation is reduced from picojoules to femtojoules. Furthermore, the massive parallelism inherent in neuromorphic chips allows for the execution of these operations in a decentralized manner, effectively eliminating the memory wall that plagues current hardware generations.

However, the path to neuromorphic optimization is not without significant technical hurdles. Most contemporary LLMs are trained using backpropagation and gradient descent, which require high-precision floating-point arithmetic and a differentiable loss function. Spiking systems are inherently non-differentiable in their raw form, complicating the direct training of LLMs on neuromorphic hardware. Current research focuses on two primary solutions: either training the model on traditional hardware and then converting the weights for neuromorphic deployment through sophisticated normalization techniques, or developing new learning rules such as surrogate gradients and biologically inspired plasticity. Achieving parity in terms of linguistic nuance and reasoning capabilities remains a challenge, as the quantization of weights into a spike-based format can lead to information loss if not managed with extreme precision.

The future of sustainable AI lies in the hardware-software co-design where the architecture of the LLM is optimized specifically for the strengths of the neuromorphic substrate. This includes the development of weight-stationary architectures where the model parameters remain stored within the processing elements themselves, a concept known as computing-in-memory.

In conclusion, the optimization of Large Language Models through neuromorphic hardware is not merely a technical improvement but a necessary evolution for the field of artificial intelligence. As the global demand for cognitive computing grows, the environmental impact of AI can no longer be ignored. Neuromorphic engineering provides a sophisticated solution that aligns the computational requirements of LLMs with the biological principles of efficiency. By embracing event-driven processing and in-memory computation, the industry can move toward a future where the power of language models is accessible, sustainable, and integrated into the fabric of everyday technology without compromising the ecological health of the planet. The shift from power-hungry digital accelerators to brain-inspired neuromorphic chips marks the beginning of the green AI revolution.