



TECHNOLOGY FEATURE Industrial News / Technology Report / Research Communication

AI Energy Crisis: Why Neuromorphic Computing Could be the Missing Piece

As the world is becoming weaker each day at an unprecedented rate, the brain-inspired neuromorphic chips have a groundbreaking solution to performing advanced AI consuming a fraction of the energy.

By **MUHAMMAD USMAN** | University of Agriculture | Published: October 2026

✉ musmanpikhi@gmail.com | Position: Student

Introduction

The artificial intelligence revolution in the world has a silent fire. Whenever you ask a generative chatbot a question, or it generates an image, or it is requested to create a complex code, an enormous data center somewhere is burning you a huge amount of energy. The very year in 2025 was a symbolic and practical turning point in the perception of AI; no longer a software-based technology, but a heavy-duty physical infrastructure that consumes gigantic quantities of electricity, water, and metals. According to the International Energy Agency, almost 4 percent of all electricity in the world is today used by data centers, of which an increasing portion are now devoted to AI. This worldwide AI energy crisis is brutally battered against a grim domestic backdrop: a long-running, crippling energy crisis. The Pakistani power sector has over a decade been in a state of affordability/reliability crisis with the circular debt skyrocketing past Rs. 2.4 trillion because of the consumption of imported fuels in the past. The frequent power outages and blackouts keep threatening the already fragile industrial sectors of the nation that already have been able to cut an estimated 3-4% off the national GDP.

“With only 20 W of power, a Go grandmaster’s brain matches capabilities that demand tens of thousands of watts from modern AI.”

With this unpredictability, the newly announced National Artificial Intelligence Policy 2025 is confidently demanding a 2,000 MW electricity quota to power local data centers and blockchain technologies, and hopes to educate one million professionals. Nevertheless, trying to power traditional, energy-intensive AI infrastructure on a grossly overstrained national grid is a financially risky plan. To be a truly digitally sovereign without adding any more to the energy crisis, the tech ecosystem in Pakistan must start considering something beyond traditional silicon. We will have to settle on a computing architecture that is founded on the end product low-power processor the human brain. The tens of thousands of watts that would be needed by the conventional AI systems of today would only be needed by the brain of a human Go grandmaster just a mere 20 watts.

Technology Overview: Getting out of the Bottleneck.

To see why neuromorphic computing is such a radical departure we must first see why conventional chips are so naturally inefficient to AI. The modern era of AI was born of the Graphics Processing Unit (GPU) unit. The GPU was first designed to provide graphics to video games, but has been repurposed to



serve AI due to its uncoded ability to do parallel calculation of mathematical functions. Nevertheless, the Von Neumann architecture, which separates a processor and a memory storage (RAM) physically, is used to build GPUs. All calculation is based on continually shifting data across this gap back and forth, which incurs wastage of gigabits of time and electricity. The cost of energy is astounding. Currently, the state-of-the-art NVIDIA H100 has a peak power draw of about 700 watts. The training run of LLaMA 3.1 (which needed more than 16,000 H100 GPUs) used an estimated 112,000 megawatt-hours of electricity. Moreover, AI is becoming a grid which constantly approaches its break-even point as a day-to-day inference on billions of user queries continues to push AI to and sometimes even beyond its break-even point. A paradigm shift in neuromorphic computing is that the biological neural systems that it models are inspirational to neuromorphic computing. Nonetheless, unlike the inflexible, continuous clock cycles, the neuromorphic chips literally integrate processing and memory in the same place. They adopt Spiking Neural Networks (SNNs), an event-based network. This implies that the system only uses dynamic power when a certain spike or data event causes a computation. When no active input is being applied then the artificial neurons will be in a more or less-sleep state and the artificial neurons will be nearly drawing no power. It has acquired a breakneck pace in 2026 the shift of lab experiments to commercial reality. In January 2026, Intel unveiled the **Loihi 3** chip. Screen shot Loihi 3, a 32-bit [12 bits in hardware terms], controlled by advanced software, is developed on a 4nm process with 8 million digital neurons and 64 billion synapses, uses advanced software to process complex data in a single pulse. When operated at full load it can only run at a peak load of 1.2 watts and is up to 1,000 times less efficient than standard GPUs at full load. In the meantime, IBM has already incorporated their North Pole chip into full production, and has completely done away with any external RAM, instead embedding all the memory on-chip. When performing vision tasks, NorthPole is shown to be 25x more energy-efficient than NVIDIA H100, and is capable of operating 3-billion parameter models with latency under a millisecond.

Key Findings & Impact

Advances in Energy-Efficient Spiking Neural Architectures The implications of ultra-efficient neuromorphic hardware are profoundly transformative, particularly for resource-constrained environments where the goal is to scale AI safely at the edge—directly on local hardware rather than relying on remote cloud servers. One of the earliest efforts in this direction involves adapting large-scale models into low-power spiking architectures. Researchers have demonstrated that it is possible to significantly reduce energy consumption without substantially compromising performance. A notable example is *SpikingMamba*, a 1.3 billion-parameter network developed in late 2025, which achieved a 4.76× improvement in energy efficiency while maintaining comparable performance, with only a 4.78% drop in zero-shot accuracy. Furthermore, complex language generation tasks can now be performed using directly trained spiking neural networks, requiring significantly fewer operations compared to traditional transformer-based models.

Edge AI Scalability in Resource-Constrained Environments The broader impact of these advancements lies in enabling scalable edge AI deployment in regions with limited infrastructure. Pakistan's National AI Policy 2025 emphasizes digital transformation across sectors; however, cloud-dependent AI systems often face challenges due to inconsistent internet connectivity and high energy costs. Neuromorphic computing offers a viable alternative by enabling intelligent processing directly on local devices, reducing reliance on centralized systems and making AI deployment more resilient and cost-effective.

Transformative Impact on Pakistan's Agriculture and IoT Ecosystem The agricultural sector in Pakistan, a key pillar of the economy, stands to benefit significantly from this technology. Research



involving Unmanned Aerial Vehicles (UAVs) equipped with AI has demonstrated strong potential for applications such as field monitoring, disease detection, and crop health management. However, traditional AI vision systems rapidly drain drone batteries, limiting operational time. By integrating low-power neuromorphic vision controllers—consuming as little as 1.2 watts—these drones can process complex agricultural data in real time using spike-based computation. This advancement enables longer flight durations, more efficient monitoring, and scalable deployment across large farming regions.

The greatest impact of this technology is that the democratization of Edge AI to Pakistan is achieved. The National AI Policy 2025 is correct in noting that the digital transformation across sectors needs to be achieved, but the problem with the cloud-dependent systems is that the hit and miss internet connectivity and the extreme cost of energy thwarts cloud-dependent systems. Consider the case of agricultural sector in Pakistan which is one of the pillars of the economy. Research made by various institutes using Unmanned Aerial Vehicles (UAVs) equipped with sophisticated AI shows the enormous potential of using Unmanned Aerial Vehicles (UAVs) to survey fields, disease detection, and control of crop health. The traditional vision processing of AI, however, boils the batteries of drones down to a few minutes. They are able to process complex farm-related data on their own through spikes in the real-time, by installing 1.2-watt neuromorphic vision controllers in the drones.

Metric/ System	Conventional GPU (e.g., NVIDIA H100)	Neuromorphic Chip (e.g., Intel Loihi 3)
Architecture	Von Neumann (Separate memory and compute)	Brain-inspired (Co-located memory and compute)
Data Processing	Continuous processing of all data points	Event-driven Spiking Neural Networks (SNNs)
Energy Profile	300–700 Watts (High dynamic/static power)	Peak load ~1.2W (Sleeps when inactive)

What's Next?

The trend is evident: the future of AI cannot be based on the brute-force scaling of power-hungry GPUs alone in case we are to achieve global sustainability goals. The industry is moving to a heterogeneous hardware ecosystem, where initial training is done with the assistance of huge GPU clusters, and ubiquitous inference at the edge is implemented with ultra-low-power neuromorphic chips. Nevertheless, the software has become an issue in changing to neuromorphic architectures. The technology industry must rethink decades of conventional software engineering, and drop continuous mathematics in favor of asynchronous, events-based programming paradigms. A number of open-source frameworks, such as the Lava of Intel, now attempt to fill this gap, but it will take focused education and time to become common practice among developers. The opportunity is enormous in the case of tech ecosystem in Pakistan. The existing export policy and financial limitations have rendered the availability of the best GPUs in the



Global South. Architecturally better suited to the realities of the electricity situation in Pakistan are neuromorphic edge systems, which are tolerant of intermittent power, and which do not require giant data centers. The government and the private sector must highly encourage the local development of edge AI solutions. As the next generation of engineers is trained to design and implement Spiking Neural Networks, Pakistan can accomplish the goals of its National AI Policy 2025 without disrupting its power grid. It will not be the number of megawatts that such a large data center will be consuming to run, but the beauty with which our engineering will be able to recreate the brilliant, low-power efficiency of the human mind.

About the Author

Muhammad Usman is a Computer Science student at the University of Agriculture Faisalabad who is a technology researcher focusing on advanced hardware architectures, decentralized infrastructure, and their socioeconomic impacts in emerging markets. He can be reached at musmanpikhi@gmail.com. Muhammad Usman can be contacted at musmanpikhi@gmail.com

Keywords: Artificial Intelligence; Neuromorphic Computing; Energy Crisis; Edge AI; Spiking Neural Networks; Pakistan Tech

References

- [1] Aivancity (2025). Behind AI: Energy, Water, and Carbon – The Environmental Footprint of 2025. Huang, Y. et al. (2025).
- [2] SpikingMamba: Towards Energy-Efficient Large Language Models via Knowledge Distillation from Mamba. arXiv:2510.04595. MEXC (2026).
- [3] IBM NorthPole vs NVIDIA H100 Energy Efficiency Research and Intel Loihi 3 Commercial Breakthrough. Paradigm Shift (2025).
- [4] Pakistan's National AI Policy 2025: Infrastructure Deficits and Energy Limits. Afghan Diaspora (2025). Pakistan's Power and Energy Crisis: Chronic Outages and Rising Costs.