TECHNOLOGY FEATURE Industrial News / Technology Report / Research Communication

Beyond the Bottleneck: Why the Future of AI is Human-Shaped

Conventional AI chips have an "energy problem" due to the Von Neumann bottleneck: data has to be moved between a separate processor and memory, consuming enormous energy.Neuromorphic computing breaks the bottleneck by bringing memory and processing together and using spiking neural networks (SNNs) that only fire when necessary Neuromorphic computing chips such as Intel's Loihi 2 use SNNs and in-memory processing to deliver up to two orders of magnitude energy efficiency and latency gains over GPUs.For instance, Loihi chips do complex inference with <1 W of power (compared to hundreds of watts for GPUs).The human brain consumes only ~20 W , so neuromorphic chips offer AI that "thinks" on human power. This allows Edge AI devices (e.g. drones, prosthetics) to locally compute information in real time, without relying on the cloud and giant batteries.

By **MURTAZA MAHBOOB** | University of Agriculture, FSD | **June,2026**
murtaza9653.1@email.com | Student

Introduction

Large AI workloads are computed on GPUs that are hundreds of watts, and tens of kilowatts for training.But the human brain consumes just ~20 W to conduct similarly complex operations.The efficiency imbalance results from the Von Neumann architecture (separate CPU and memory).All arithmetic operations must ship data back and forth across the separation - like commuters commuting between cities.Such data "commutes" are a major energy cost and scaling bottleneck for silicon AI.By contrast, neurons in the brain merge memory and processing, and use event-driven (spiking) signals.These are the principles of the brain that neuromorphic computing adopts: it collocates memory and processing, and communicates only when necessary using events (spikes).In this article, we will see how this eliminates the bottleneck and enables a new low-energy AI revolution.

> ***“While today’s AI systems need tens of thousands of watts to rival human intelligence, a human Go grandmaster’s brain consumes only about 20 W.”***

This pull-quote (from NIST) motivates.What follows is how our brain-efficient neuromorphic chips (e.g. Intel Loihi 2, IBM NorthPole) mimic the brain.Innovations include in-memory computing (no data shuttling)g) and spiking neural networks that do not compute unless triggered by spikes (bemocking neurons).These advances hold out the promise of AI running with an order of magnitude lower power.

Technology Overview

Neuromorphic hardware rewrites the rules.Neuromorphic chips are not clocked and not continuous like traditional chips.They integrate synaptic memory and neuron-like processing elements (analog or digital) on chip.According to IBM Research, the co-location of compute and memory "not only overcomes the von Neumann bottleneck, but these devices also store intermediate values beyond just the ones and zeros of typical transistors" .This translates to local lookups, not from off-chip DRAM. Spiking neural networks (SNNs) also improve efficiency: calculations are only done on spikes.Neurons in the brain only talk occasionally, so neuromorphic circuits are asleep until a spikes .This produces a “spike-driven” data flow.An example is Intel’s Loihi 2.A Loihi 2 chip has ~1 million spiking neurons and ~120 million synapsess.It has an asynchronous on-chip communication network: neuron cores only communicate sparse packets when they fire.Demonstrations of Loihi did inference with <1 W of power, compared to tens or hundreds of watts for CPUs/GPUs.So, Loihi had “several orders of magnitude” greater energy efficiency than a GPU for these tasks.The chip also had low latency: spikes are processed in parallel with little or no queuing, so the latency was as fast as or faster than conventional accelerators. Likewise, IBM’s NorthPole chip is a digital SNN.Weights are stored in local memory but, like Loihi, NorthPole overcomes the Von Neumann bottleneckk.Preliminary measurements by IBM have shown that NorthPole is 46× faster and 73× more energy efficient than the best GPU for a large modell.This highlights the fact that brain-inspired chips can greatly outperform GPUs on edge AI applications.

Key Findings & Impact

In-Memory Computing: Neuromorphic chips combine memory and computation.For instance, phase-change-memory or resistive RAM devices integrate multiplication and memorys.This avoids the data shuffling of conventional cores.This can save 70% or more energy per operation than Von Neumann designs.In reality, systems such as Intel's Loihi connect synapses to local SRAM, eliminating the need to access DRAM, and reducing the memory cost.

Spiking Neural Networks: SNNs activate only on events.They take continuous neural activity and transcode it to sparse spikes.This is like having thousands of microprocessors that are asleep until a neuron fires, and then sleep again..IBM points out that "our nerve cells are communicating sparsely, which is why we're so efficientt ”.The practical result is dramatic: Less than 1 W is consumed by Loihi's demos for tasks that consume hundreds of watts on GPUs

Energy Efficiency (GPU vs. Neuromorphic vs. Brain): The latest GPUs use about 300-1200 W under full load. Neuromorphic processors (such as Loihi), on the other hand, consume a few watts.The table below compares these with the human brain:

System	Power Consumption	Latency	Use-Case Suitability
GPU	~500–1000 W (per device)	Moderate (~10–100 ms)	Cloud/data-center AI, large-scale training
Loihi 2	< 1 W	Very low (µs–ms)	On-device/edge inference (always-on sensors)
Human Brain	~20 W	Moderate (10–100 ms)	General intelligence, sensorimotor tasks

Edge AI Applications: The power savings enable new edge AI. Take a micro drone: conventional AI chips limited its 46-min flight to 4.A neuromorphic vision-motor controller can enable full flight time by processing just the spikes. In robotics, small spiking chips can process sensor inputs and control in real time, at millisecond speed. Likewise, new prosthetics employs neuromorphic chips to translate tactile or motor events to spike events, with a delay of <100 ms. In sum, robots, drones and even wearable arms can "think" locally without needing the cloud, due to the hardware efficiency of the brain

What's Next?

Neuromorphic is a radical yet practical vision The feasibility of the approach is demonstrated by Loihi and NorthPole, but there's much engineering to do. Scaling up new types of memristive materials is difficult, and new neuromorphic chips primarily support inference (training still occurs on digital computers). Software is challenging too: current AI software doesn't natively translate to spikes. But the stakes are high. As NIST notes, there's lots to learn from the brain's 20 W efficiency . Research is helping to fill the tools gap (e.g. Intel's Lava). To conclude, neuromorphic engineering is a change of mindset to "think like a brain". It overcomes the current energy crisis by changing the data flow and computation primitives. If hardware and software challenges can be addressed then we are on track to have AI systems that draw energy similar to humans - making "AI that thinks like us" real.

About the Author

Murtaza Mahboob is a Computer Science student at the University of Agriculture Faisalabad who has a keen interest in Artificial Intelligence, Machine Learning and Data Science. He has a keen interest in leveraging new developments in AI and incorporating them into practical applications. Murtaza Mahboob can be contacted at murtaza9653.1@gmail.com

Keywords: Neuromorphic Computing; Spiking Neural Networks; In-Memory Computing; Edge AI; AI Energy Efficiency

References

[1] IBM Research (2023). How neuromorphic computing takes inspiration from our brains. research.ibm.com.
[2] Intel Labs (2022). Loihi 2 Neuromorphic Processor Whitepaper. intel.com.
[3] Wright, J., & Davis, T. (2023). Taking Measure: Brain-Inspired Efficiency. National Institute of Standards and Technology (NIST) blog