The Dimensionality Gap: Three Scaling Barriers to AGI on Von Neumann Architectures, and a Path Forward Through Neuromorphic-Photonic Substrates

Jyotiprakash Mishra *

Posted Date: 27 March 2026

doi: 10.20944/preprints202603.2196.v1

Keywords: artificial general intelligence; von Neumann architecture; neuromorphic computing; photonic interconnects; brain-scale emulation; memory bandwidth; energy efficiency; 3D integration; bisection bandwidth; scaling barriers

 doi:10.20944/preprints202603.2196.v1

Article

The Dimensionality Gap: Three Scaling Barriers to AGI on Von Neumann Architectures, and a Path Forward Through Neuromorphic–Photonic Substrates

Jyotiprakash Mishra

Birla Institute of Technology and Science, K. K. Birla Goa Campus, Zuarinagar, Goa 403726, India; mail@jyotiprakash.org

Abstract

The prevailing assumption in artificial general intelligence (AGI) research is that scaling current architectures—more parameters, more GPUs, more data—will eventually yield human-level general intelligence. We challenge this assumption by identifying three formal scaling barriers that arise from the structural mismatch between biological neural computation and von Neumann silicon. The brain is a three-dimensional, asynchronous, massively parallel graph of ∼86 billion nodes with average fan-out ∼7,000, operating at ∼20 W. Contemporary processors are two-dimensional, clocked, and memory-bottlenecked. We derive: (1) a *communication complexity barrier* showing that emulating the brain's non-planar connectivity on a 2D substrate incurs $\Omega(N^{2/3} \cdot F)$ additional data movement per timestep; (2) a *serialisation barrier* showing that the required memory bandwidth exceeds 10^{15} bytes/s, roughly $10^3\times$ current GPU capacity; and (3) an *energy barrier* showing that silicon emulation at brain scale requires ∼6 MW under current technology, versus 20 W for biology. These are not engineering inconveniences—they are architectural incompatibilities that improve only polynomially with process scaling. We propose a constructive path forward: a hybrid architecture combining 3D-stacked neuromorphic silicon for local synaptic computation with integrated photonic interconnects for high-fan-out, low-energy long-range communication. We specify quantitative targets and identify the critical technology gaps. The paper does not claim that brain emulation is sufficient for AGI, nor that current deep learning is without merit—only that the dominant hardware substrate is structurally mismatched to brain-scale neural computation in ways that scaling alone cannot overcome.

Keywords: artificial general intelligence; von Neumann architecture; neuromorphic computing; photonic interconnects; brain-scale emulation; memory bandwidth; energy efficiency; 3D integration; bisection bandwidth; scaling barriers

1. Introduction

The dominant research paradigm in AGI pursues scale: larger transformer models, more GPU hours, bigger datasets. This approach has produced remarkable capabilities—large language models that pass professional examinations, generate code, and engage in extended reasoning [1,2]. The implicit assumption is that these capabilities will converge on general intelligence as scale increases.

We argue that this assumption contains a structural flaw. The biological brain, the only known system that achieves general intelligence, differs from silicon processors not merely in degree but in *kind*. The brain is a three-dimensional, asynchronous, event-driven graph with ∼86 billion neurons [3], each connecting to ∼7,000 others [4], operating at ∼20 W total [5]. Contemporary CPUs and GPUs are two-dimensional, globally clocked, memory-bottlenecked fetch-execute machines. The mismatch between these architectures is not a quantitative gap that Moore's law can close—it is a *dimensionality gap*: a structural incompatibility between the topology, parallelism model, and energy budget of biology and those of planar silicon.

This paper makes two contributions. First, we formalise the dimensionality gap as three scaling barriers (Propositions 1–3)—lower bounds on the communication, memory bandwidth, and energy

 doi:10.20944/preprints202603.2196.v1

costs of emulating a brain-scale neural graph on von Neumann hardware. These bounds are derived from first principles of VLSI theory, memory architecture, and thermodynamics, and are evaluated at biological parameters. They establish that brain-scale emulation on 2D silicon is not merely expensive but *infeasible at any realistic engineering timescale*—the resource requirements exceed current technology by factors of 10^2–10^3 and improve only polynomially with process advances.

Second, we propose a constructive path forward: a hybrid neuromorphic–photonic architecture that addresses each barrier directly. Three-dimensional neuromorphic silicon restores the topological routing freedom lost in 2D fabrication (addressing Barrier 1). Integrated photonic interconnects provide the fan-out bandwidth and energy efficiency that electronic wires cannot (addressing Barriers 2 and 3). We specify quantitative targets for a first-generation device and identify the critical technology gaps.

We emphasise what this paper does *not* claim. It does not claim that brain emulation is sufficient for AGI—consciousness, grounding, and embodiment may require more. It does not claim that current deep learning is on the wrong track—transformer-based systems have genuine and remarkable capabilities that may not require brain-like hardware. It claims only that the specific path of "emulate the brain on GPUs" runs into three formal barriers that scaling alone cannot overcome, and that a different class of substrate is needed to make that path viable.

Related work.—The von Neumann bottleneck was identified by Backus in 1977 [12]. Neuromorphic computing was proposed by Mead in 1990 [13]. Modha et al. [15] and Davies et al. [14] have built large-scale neuromorphic chips. Shen et al. [17] demonstrated optical neural networks. Marković et al. [18] surveyed physics-based computing for AI. Our contribution is the formal quantification of the three barriers and the specific hybrid architecture that addresses all three simultaneously.

2. The Brain as a Computational Graph

We model the brain as a directed weighted graph $G = (V, E, w)$ with $|V| = N \approx 8.6 \times 10^{10}$ neurons [3], $|E| \approx 10^{14}$ synaptic connections [4], and continuously variable synaptic weights $w : E \to \mathbb{R}$. Each vertex $v \in V$ fires asynchronously when a threshold function over its weighted inputs is exceeded—a non-linear, analog, event-driven computation.

Three properties of G are computationally critical:

Non-planarity.—G is embedded in three-dimensional space and exploits all three dimensions for wiring. It is highly non-planar: the cortical white matter consists of ∼150,000 km of myelinated axons [7] carrying long-range connections that cross every conceivable 2D section. The crossing number of a random graph with N vertices and average degree F scales as $\Omega(N^2 \cdot F^2 / \log^2 N)$ when forced into a 2D embedding [8]. This non-planarity is not incidental; it enables the small-world connectivity (∼6 synaptic hops between any two neurons [9]) that underpins rapid information integration.

Massive fan-out.—The average neuron projects to $F \approx 7{,}000$ postsynaptic targets [4]. When a neuron fires, all F downstream neurons receive input *simultaneously* via physical axonal branching—a parallel broadcast at the speed of axonal propagation, with no serialisation.

Extreme energy efficiency.—A single synaptic transmission consumes ∼1–10 fJ [6]. The entire brain achieves $\sim 10^{14}$ synaptic operations per second at ∼20 W—an energy cost of ∼0.2 pJ per operation including all metabolic overhead [5,6].

3. Three Scaling Barriers

We now derive three lower bounds on the cost of emulating G on von Neumann hardware. Each barrier arises from a different axis of the dimensionality gap: topology, parallelism, and energy.

3.1. Barrier 1: Communication Complexity

Contemporary processors are fabricated as 2D planar silicon dies. Inter-unit communication must be routed on a flat surface. The fundamental constraint is the *bisection bandwidth*: the maximum data rate across any cut that partitions the chip into two halves of equal compute. For a 2D chip of area A, the bisection bandwidth scales as $\mathcal{O}(\sqrt{A})$—the perimeter of the bisection line [19,20].

 doi:10.20944/preprints202603.2196.v1

The brain's graph G, by contrast, exploits three-dimensional wiring. The bisection bandwidth of a 3D-embedded graph with N nodes scales as $\mathcal{O}(N^{2/3})$ [20], which for a 2D chip with the same node count drops to $\mathcal{O}(N^{1/2})$. The ratio represents data movement that cannot be handled in-plane and must be serialised through the 2D bisection.

Proposition 1 (Communication complexity barrier)**.** *Any 2D substrate emulating one timestep of a neural graph G with N vertices, average degree F, and 3D bisection bandwidth $B_{3D} = \Theta(N^{2/3})$ must perform at least*

$$C = \Omega\left(\frac{N^{2/3} \cdot F}{N^{1/2}}\right) = \Omega(N^{1/6} \cdot F) \tag{1}$$

additional data-movement operations per timestep beyond what a 3D substrate would require, where the $N^{1/2}$ denominator is the 2D bisection bandwidth.

The derivation (Methods, §7) follows from the bisection bandwidth framework of Thompson [19] and Leighton [20]. At brain parameters ($N = 8.6 \times 10^{10}$, $F = 7{,}000$), the overhead factor is

$$C \approx (8.6 \times 10^{10})^{1/6} \times 7{,}000 \approx 660 \times 7{,}000 \approx 4.6 \times 10^{6} \tag{2}$$

additional data movements per emulated neuron per timestep. Across the $\sim 10^{10}$ neurons that fire per second (at average $\sim$10 Hz firing rate across 10^{11} neurons active at any moment, with sparse coding), this amounts to $\sim 4.6 \times 10^{16}$ excess data movements per second—orders of magnitude beyond the $\sim 10^{12}$ cache-line transfers per second achievable on current high-end GPUs [21].

Interpretation.—This barrier does not claim that 2D emulation is *impossible* in principle. It establishes that the communication overhead grows as $\Omega(N^{1/6} \cdot F)$ and cannot be eliminated by algorithm design; it is a property of the dimensional mismatch between the graph and the substrate. Process scaling (smaller transistors) increases $N^{1/2}$ slowly and does not close the gap.

3.2. Barrier 2: Serialisation Bandwidth

When neuron v fires in the brain, it updates F downstream neurons *simultaneously* via axonal branching. On a von Neumann machine, this fan-out must be serialised: F sequential (or SIMD-batched) memory writes. The required aggregate memory bandwidth is:

Proposition 2 (Serialisation barrier)**.** *Any von Neumann emulation of a neural graph with N neurons, average fan-out F, and average firing rate f requires sustained memory bandwidth*

$$B \geq N \cdot f \cdot F \cdot b\,, \tag{3}$$

where b is the bytes per synaptic update (minimum: weight read + state write $\geq$ 8 bytes for 32-bit precision).

At biological parameters ($N = 8.6 \times 10^{10}$, $f \approx 1$ Hz average across all neurons including quiescent populations [10], $F = 7{,}000$, $b = 8$ bytes):

$$B \geq 8.6 \times 10^{10} \times 1 \times 7{,}000 \times 8 \approx 4.8 \times 10^{15} \text{ bytes/s}\,. \tag{4}$$

This is approximately 4.8 PB/s. For comparison, the NVIDIA H100 GPU achieves $\sim$3.35 TB/s HBM bandwidth [21]—a factor of $\sim$1,400 shortfall.

Even at a reduced firing rate estimate of 0.1 Hz (accounting for sparse cortical activity [11]), the requirement is $\sim$480 TB/s—still $\sim 140\times$ current GPU bandwidth. Memory bandwidth has historically improved at $\sim$25%/year, giving a $\sim$20–30 year timeline to close a $140\times$ gap on this axis alone, assuming current architectural paradigms.

 doi:10.20944/preprints202603.2196.v1

Counterargument and response.—One might object that sparse firing patterns reduce the effective bandwidth. This is already accounted for: the 0.1–1 Hz average rate reflects cortical sparsity. The deeper issue is that even sparse updates are *scatter-pattern* memory accesses (each firing neuron writes to F non-contiguous memory locations), which achieve only ∼1–10% of peak bandwidth on GPU memory systems [22]. The effective bandwidth gap is therefore 10–100× larger than the raw numbers suggest.

3.3. Barrier 3: Energy Scaling

Proposition 3 (Energy barrier). *Silicon emulation of a neural graph with synaptic throughput S operations/s at per-operation energy η_{Si} requires power*

$$P_{\text{Si}} = S \cdot \eta_{\text{Si}} \,. \tag{5}$$

At brain-scale throughput $S \approx 10^{14}$ ops/s and current GPU efficiency $\eta_{\text{Si}} \approx 60$ pJ/op (NVIDIA H100 achieves ∼60 pJ per 16-bit multiply-accumulate at the chip level [21,23]), the required power is

$$P_{\text{Si}} \approx 10^{14} \times 6 \times 10^{-11} \approx 6 \times 10^{3}\ W = 6\ kW \text{ per chip equivalent.} \tag{6}$$

Accounting for memory access energy (∼10× the compute energy for off-chip DRAM access at brain-scale working sets [23]), the system-level power is ∼60 kW–6 MW, versus ∼20 W for the biological brain.

The energy ratio $\eta_{\text{Si}}/\eta_{\text{brain}} \approx 60\,\text{pJ}/0.2\,\text{pJ} = 300$ quantifies the efficiency gap. This ratio has improved at approximately 2× per chip generation (∼3 years), or ∼25%/year—similar to the Dennard scaling rate, which ended circa 2006 [24,25]. Post-Dennard, energy efficiency gains have slowed to ∼10–15%/year. At 15%/year, closing a 300× gap requires ∼40 years.

Thermodynamic context.—The Landauer limit for irreversible bit erasure is $k_B T \ln 2 \approx 3 \times 10^{-21}$ J at room temperature [26]. Current CMOS operates at $\sim 10^4$–$10^5\times$ Landauer, while biological synapses operate at $\sim 10^2$–$10^3\times$ Landauer. Biology is not at the fundamental limit either, but it is ∼100× closer than silicon. This suggests that the energy gap has both an engineering component (improvable) and a physical-substrate component (requiring new materials or mechanisms).

4. Quantitative Comparison

Table 1 summarises the dimensionality gap across eight metrics for five systems: the biological brain, a current CPU, a current GPU, a state-of-the-art neuromorphic chip, and the proposed hybrid architecture of Section 5.3.

Table 1. The dimensionality gap: quantitative comparison across five computational substrates. Brain parameters are drawn from [3–6]. GPU parameters are for NVIDIA H100 [21]. Neuromorphic parameters are for Intel Loihi 2 [14,16]. Hybrid targets are from Section 5.3.

Metric	Brain	CPU (AMD EPYC)	GPU (NVIDIA H100)	Neuromorphic (Intel Loihi 2)	Hybrid (target) (§5.3)
Neurons / nodes	8.6×10^{10}	∼10^1 cores	∼10^4 cores	∼10^6	≥10^9
Connections / edges	∼10^{14}	∼10^3 (bus)	∼10^6 (NoC)	∼10^9	≥10^{13}
Average fan-out	∼7,000	N/A (shared mem)	N/A (shared mem)	∼4,000	≥5,000
Topology	3D, non-planar	2D planar	2D planar	2D planar	3D stacked
Parallelism	Async, spatial	Pipelined	SIMD	Async, spatial	Async, spatial
Energy / op	∼0.2 pJ	∼500 pJ	∼60 pJ	∼10 pJ	≤5 pJ
Total power	∼20 W	∼300 W	∼700 W	∼1 W	≤500 W
Clock / timing	None (async)	∼4 GHz	∼2 GHz	Async	Async

Several features of Table 1 deserve comment. The brain exceeds all silicon systems in neuron count, connection count, and energy efficiency by orders of magnitude. Current neuromorphic chips (Loihi 2, IBM NorthPole [15]) close the energy-per-operation gap to ∼50× and achieve asynchronous operation, but remain 2D planar and orders of magnitude short of brain-scale node counts. No existing system addresses all three barriers simultaneously. The hybrid architecture (Section 5.3) is the first proposal that targets all eight dimensions concurrently.

 doi:10.20944/preprints202603.2196.v1

5. A Path Forward

The three barriers identify the axes along which a brain-scale substrate must improve. We now propose three complementary technologies, culminating in a hybrid architecture.

5.1. 3D Neuromorphic Hardware

Barrier 1 arises from 2D fabrication. The remedy is 3D integration: stacking compute layers vertically so that connections can exploit all three spatial dimensions, partially recovering the routing freedom of biological tissue.

The technology exists in embryonic form. High-Bandwidth Memory (HBM) stacks up to 12 DRAM layers with through-silicon vias (TSVs) [27]. Monolithic 3D integration, where transistor layers are fabricated sequentially on a single wafer [28], promises >100 vertical connections per μm^2—far denser than TSV-based stacking.

A brain-scale 3D neuromorphic chip would require: (i) at least 10–100 vertical compute layers, each containing $\sim 10^7$–10^8 neuromorphic cores; (ii) asynchronous, event-driven operation with no global clock; (iii) on-chip analog or mixed-signal synaptic weight storage to avoid the memory-access energy penalty of digital SRAM/DRAM.

Intel Loihi 2 [14] and IBM NorthPole [15] implement (ii) and partially (iii) in 2D. BrainScaleS-2 [29] uses analog circuits for synaptic dynamics. None has achieved (i). Monolithic 3D integration [28] is the most promising path to vertical scaling, but thermal management at >10 layers remains an open engineering challenge.

5.2. Photonic Interconnects

Barriers 2 and 3 arise from the serialisation cost and energy dissipation of electronic interconnects. Photonic (optical) communication offers three properties that directly address these:

Parallel broadcast.—A single optical waveguide carrying W wavelengths via wavelength-division multiplexing (WDM) can broadcast to W destinations simultaneously, providing physical fan-out without serialisation. At $W = 100$ channels per waveguide (achievable with current silicon photonics [30]), each waveguide replaces 100 electronic wires.

Low energy.—Optical switching in microring resonators achieves $\sim$1–10 fJ per bit [30,31], approaching the synaptic energy of biology.

No crosstalk.—Photons do not interact with each other in linear media, enabling dense waveguide routing without the signal integrity problems of electronic interconnects at high density.

Optical neural networks have been demonstrated by Shen et al. [17] using Mach–Zehnder interferometer meshes, and photonic tensor cores by Feldmann et al. [31] using phase-change materials. The critical gap is *optical memory*: current photonic systems rely on electronic memory for weight storage, creating an optical–electronic conversion bottleneck. Until all-optical or hybrid opto-electronic memory matures, photonic systems are best suited for the *communication* layer (fan-out and long-range routing) rather than the *compute* layer (synaptic weight storage and integration).

5.3. Hybrid Neuromorphic–Photonic Architecture

We propose a hybrid architecture that assigns each technology to the task it handles best:

Electronic neuromorphic cores handle local synaptic computation: weight storage, dendritic integration, spike generation. These are fabricated as 3D-stacked layers of asynchronous, mixed-signal neuromorphic circuits (Section 5.1), providing the topology and energy density needed for local processing.

Photonic interconnects handle long-range communication: high-fan-out spike broadcast between cores on different layers or at large intra-layer distances. WDM waveguides integrated between neuromorphic layers provide the bandwidth of Barrier 2 at the energy cost of Barrier 3.

Architecture specification:

- **Compute:** $\geq 10^9$ neuromorphic neurons across $\geq$10 vertically stacked layers, with on-chip analog synaptic weights.

 doi:10.20944/preprints202603.2196.v1

- **Interconnect:** Integrated silicon photonic waveguides with $\geq$100 WDM channels per waveguide, providing aggregate fan-out bandwidth $\geq 10^{14}$ events/s.
- **Energy:** $\leq$5 pJ per synaptic operation (electronic compute + photonic communication), system power $\leq$500 W at 10^9-neuron scale.
- **Timing:** Fully asynchronous—no global clock.

This architecture addresses each barrier: 3D stacking reduces the communication overhead of Barrier 1 by restoring $\mathcal{O}(N^{2/3})$ bisection bandwidth; photonic fan-out eliminates the serialisation of Barrier 2 by providing physical parallel broadcast; and the combination of analog neuromorphic compute ($\sim$1–10 pJ/op) and photonic communication ($\sim$1–10 fJ/bit) approaches the energy budget of Barrier 3.

The first-generation device targets 10^9 neurons—approximately 1% of brain scale. This is sufficient to emulate a complete cortical column or a small mammalian brain (mouse-scale), providing a biologically meaningful testbed. Scaling to 10^{11} (full human brain) requires $\sim$100$\times$ more layers or lateral area, which we regard as a second-generation challenge contingent on advances in monolithic 3D integration and thermal management.

6. Limitations and Scope

Brain emulation is not sufficient for AGI.—The three barriers concern the feasibility of brain-scale *emulation* on von Neumann hardware. They do not address whether emulation is sufficient for general intelligence. Consciousness, embodied cognition, developmental learning, and social grounding may require mechanisms beyond what any hardware substrate provides.

Current deep learning may not need brain-like hardware.—Large language models achieve impressive capabilities on conventional GPUs without brain-like connectivity. Our argument does not claim that GPU-based AI is a dead end—only that the specific path of "scale up brain emulation on GPUs" encounters formal barriers. Alternative paths to AGI (e.g., via scaling transformers, symbolic-neural hybrids, or novel algorithms) are not addressed by our analysis.

The barriers are lower bounds, not impossibility proofs.—Each proposition establishes a *minimum* resource requirement, not an absolute impossibility. A sufficiently large GPU cluster could, in principle, brute-force past Barriers 1–3. The argument is that the required cluster size ($\sim 10^4$–10^6 current GPUs, consuming megawatts) is impractical and that a purpose-built substrate is the more viable path.

The proposed architecture is speculative.—The hybrid neuromorphic–photonic architecture of Section 5.3 specifies targets, not a demonstrated system. Key technology gaps remain: monolithic 3D integration at >10 layers, large-scale photonic–electronic co-integration, thermal management of 3D stacks, and scalable fabrication. We present it as a *research direction*, not a product roadmap.

The brain may not be optimal.—We use the brain as a benchmark because it is the only known generally intelligent system, not because we claim its architecture is optimal. Alternative substrates (quantum, chemical, biological) may achieve general intelligence through entirely different mechanisms.

7. Conclusion

The question animating this paper is not "How many more GPUs do we need?" but "Are GPUs the right substrate at all?" The three scaling barriers—communication complexity, serialisation bandwidth, and energy—establish that the answer is no, at least for the brain-emulation path to AGI. The dimensionality gap between 2D silicon and 3D biology is structural, not quantitative, and it improves only polynomially with process scaling.

The hybrid neuromorphic–photonic architecture we propose is not a finished solution. It is a specification of what a brain-scale substrate must look like: three-dimensional, asynchronous, massively parallel, with photonic long-range communication and analog local computation. Building it is an engineering challenge at least as ambitious as the Manhattan Project or the Human Genome Project. But the scaling barriers suggest it is a challenge that cannot be bypassed—only met.

 doi:10.20944/preprints202603.2196.v1

Methods

Communication Complexity Derivation

The bisection bandwidth framework [19,20] establishes that for a graph G with N nodes embedded in d dimensions, the bisection bandwidth—the minimum number of edges crossing any equipartition—scales as $\Theta(N^{(d-1)/d})$. In 3D ($d = 3$), this gives $B_{3D} = \Theta(N^{2/3})$; in 2D ($d = 2$), $B_{2D} = \Theta(N^{1/2})$.

When a 3D-embedded graph is emulated on a 2D substrate, any timestep requiring data movement across the 3D bisection must route that traffic through the 2D bisection. The congestion factor is $B_{3D}/B_{2D} = \Theta(N^{2/3}/N^{1/2}) = \Theta(N^{1/6})$. Each of the B_{3D} edges in the 3D bisection carries $\mathcal{O}(F)$ messages per timestep (one per outgoing synapse for each neuron whose axon crosses the bisection), giving a total communication overhead of $\Omega(N^{1/6} \cdot F)$ additional operations per timestep.

This is a lower bound: it counts only the bisection congestion and does not account for multi-hop routing within the 2D substrate, which adds further overhead.

Serialisation Bandwidth Derivation

The bandwidth requirement follows from a counting argument. In each simulated second, $N \cdot f$ neurons fire. Each firing neuron must update F synaptic targets. Each update requires at minimum a weight read (4 bytes for 32-bit float) and a state write (4 bytes), totalling $b = 8$ bytes. The aggregate bandwidth is $B = N \cdot f \cdot F \cdot b$.

This is a lower bound because it assumes perfect memory locality. In practice, the F targets of a given neuron are distributed across the graph with low spatial locality, resulting in scatter-pattern memory access. On current GPU architectures, scatter accesses achieve approximately 1–10% of peak memory bandwidth [22], further widening the gap.

Energy Scaling Derivation

The power requirement $P_{\text{Si}} = S \cdot \eta_{\text{Si}}$ is a direct product of throughput and per-operation energy. The per-operation energy $\eta_{\text{Si}} \approx 60\,\text{pJ}$ for the NVIDIA H100 is computed as total chip power ($\sim$700 W) divided by peak 16-bit throughput ($\sim 10^{13}$ ops/s), consistent with Horowitz's analysis of energy costs at each level of the memory hierarchy [23]. The biological figure $\eta_{\text{brain}} \approx 0.2\,\text{pJ}$ is computed as total brain power ($\sim$20 W) divided by estimated synaptic throughput ($\sim 10^{14}$ ops/s) [6].

The system-level power includes memory access energy. At brain-scale working sets ($\sim 10^{14}$ synaptic weights $\times$ 4 bytes = $\sim$400 TB), the data cannot reside in on-chip SRAM and must be accessed from off-chip DRAM or HBM. Horowitz [23] estimates off-chip DRAM access at $\sim$640 pJ per 64-bit word, which is $\sim 10\times$ the compute energy. Including memory access raises the system-level power by an order of magnitude, yielding the 60 kW–6 MW range quoted in Proposition 3, depending on the cache hit rate and memory hierarchy design.

References

1. T. Brown *et al.*, "Language models are few-shot learners," Adv. Neural Inf. Process. Syst. **33**, 1877 (2020).
2. OpenAI, "GPT-4 technical report," arXiv:2303.08774 (2023).
3. F. A. C. Azevedo *et al.*, "Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain," J. Comp. Neurol. **513**, 532 (2009).
4. J. DeFelipe, L. Alonso-Nanclares, and J. I. Arellano, "Microstructure of the neocortex: Comparative aspects," J. Neurocytol. **31**, 299 (2002).
5. M. E. Raichle and D. A. Gusnard, "Appraising the brain's energy budget," Proc. Natl. Acad. Sci. USA **99**, 10237 (2002).
6. S. B. Laughlin and T. J. Sejnowski, "Communication in neuronal networks," Science **301**, 1870 (2003).
7. L. Marner, J. R. Nyengaard, Y. Tang, and B. Pakkenberg, "Marked loss of myelinated nerve fibers in the human brain with age," J. Comp. Neurol. **462**, 144 (2003).
8. J. Spencer and G. Tóth, "Crossing numbers of random graphs," Random Struct. Algorithms **21**, 347 (2002).
9. O. Sporns, D. R. Chialvo, M. Kaiser, and C. C. Hilgetag, "Organization, development and function of complex brain networks," Trends Cogn. Sci. **8**, 418 (2004).

 doi:10.20944/preprints202603.2196.v1

10. C. Koch, *Biophysics of Computation* (Oxford University Press, 1999).
11. S. Shoham, D. H. O'Connor, and R. Segev, "How silent is the brain: is there a dark matter problem in neuroscience?" J. Comp. Physiol. A **192**, 777 (2006).
12. J. Backus, "Can programming be liberated from the von Neumann style?" Commun. ACM **21**, 613 (1978).
13. C. Mead, "Neuromorphic electronic systems," Proc. IEEE **78**, 1629 (1990).
14. M. Davies *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," IEEE Micro **38**, 82 (2018).
15. D. S. Modha *et al.*, "Neural inference at the frontier of energy, space, and time," Science **382**, 329 (2023).
16. G. Orchard *et al.*, "Efficient neuromorphic signal processing with Loihi 2," in *Proc. IEEE Workshop on Signal Processing Systems* (2021).
17. Y. Shen *et al.*, "Deep learning with coherent nanophotonic circuits," Nat. Photonics **11**, 441 (2017).
18. D. Marković, A. Mizrahi, D. Querlioz, and J. Grollier, "Physics for neuromorphic computing," Nat. Rev. Phys. **2**, 499 (2020).
19. C. D. Thompson, "A complexity theory for VLSI," Ph.D. thesis, Carnegie Mellon University (1980).
20. F. T. Leighton, *Complexity Issues in VLSI* (MIT Press, 1983).
21. NVIDIA Corporation, "NVIDIA H100 Tensor Core GPU datasheet," (2024).
22. J. L. Hennessy and D. A. Patterson, *Computer Architecture: A Quantitative Approach*, 6th ed. (Morgan Kaufmann, 2017).
23. M. Horowitz, "Computing's energy problem (and what we can do about it)," in *IEEE Int. Solid-State Circuits Conf.* (2014), pp. 10–14.
24. R. H. Dennard *et al.*, "Design of ion-implanted MOSFET's with very small physical dimensions," IEEE J. Solid-State Circuits **9**, 256 (1974).
25. H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, and D. Burger, "Dark silicon and the end of multicore scaling," in *Proc. 38th Int. Symp. Computer Architecture* (2011), pp. 365–376.
26. R. Landauer, "Irreversibility and heat generation in the computing process," IBM J. Res. Dev. **5**, 183 (1961).
27. J. C. Lee, J. Kim, K. W. Kim, Y. J. Ku, D. S. Kim, and C. Cho, "High bandwidth memory (HBM) with TSV technique," in *Proc. IEEE Int. SoC Design Conf.* (2016), pp. 181–182.
28. M. M. Shulaker *et al.*, "Three-dimensional integration of nanotechnologies for computing and data storage on a single chip," Nature **547**, 74 (2017).
29. C. Pehle *et al.*, "The BrainScaleS-2 accelerated neuromorphic system with hybrid plasticity," Front. Neurosci. **16**, 795876 (2022).
30. B. J. Shastri *et al.*, "Photonics for artificial intelligence and neuromorphic computing," Nat. Photonics **15**, 102 (2021).
31. J. Feldmann *et al.*, "Parallel convolutional processing using an integrated photonic tensor core," Nature **589**, 52 (2021).