



Nature Never Built a Memory Wall: Biomimetic Compute-Near-Memory Principles for Next-Generation AI Systems

Brian L McCarson

AI Economy LLC, United States

Citation: Brian L McCarson (2026) *Nature Never Built a Memory Wall: Biomimetic Compute-Near-Memory Principles for Next-Generation AI Systems*. J. of Mod Phy & Quant Neuroscience 2(5), 1-19. WMJ/JPQN-196

Abstract

The human brain operates at approximately 20 watts while performing computational work that, by order-of-magnitude estimates, is comparable to that of exascale supercomputers consuming 20 megawatts. One architectural principle underlies much of this disparity: in biological neural systems, memory and computation occupy the same physical elements, so that a synapse stores information and performs the weighted summation that constitutes neural computation at the same site. This paper examines the compute-near-memory (CNM) and compute-in-memory (CIM) architectures of biological neural systems through the lens of biomimicry and neuromorphic engineering. It quantifies the energy, latency, and spatial efficiency advantages of the brain's integrated memory-compute organization relative to conventional von Neumann systems, and surveys the silicon translation of these biological principles across the semiconductor industry, including Samsung's HBM-PIM, Intel's Loihi 2, IBM's NorthPole, NVIDIA's Rubin platform, and emerging analog CIM architectures. The paper then proposes that the evolution of artificial intelligence follows a continuum of progressively more integrated architectural stages, categorized by major semiconductor transitions (Eras) and capability milestones (Epochs), analogous to geological and biological periodization. From this framework, a five-layer biological reference architecture for cognition is derived, ordered by evolutionary precedence: consequence-grounded sensation, emotional valence, metacognitive observation, theory of mind, and scaled parallel cognition. The grounding problem, the absence of consequence-carrying sensory input in current AI systems, is identified as the unsolved challenge this ordering exposes. Three software architectures are then proposed against that reference stack, each addressing a distinct consequence of von Neumann separation as it appears in software: context awareness through a context-aware AI database (CAAD), iterative scientific reasoning through a hypothesis generation and testing system (HGTS), and architectural governance through agentic execution and hierarchical governance (AEHG). The paper advances the idea of convergent architectural evolution, in which the semiconductor industry arrives at the brain's solutions under the same physical constraints that shaped them rather than by copying them. It concludes that the brain's architecture is best treated as a characterized engineering reference design, with specifications that apply directly to next-generation AI hardware and software.

*Corresponding author: Brian L McCarson, AI Economy LLC, United States.

Submitted: 15.09.2026

Accepted: 21.09.2026

Published: 09.10.2026

Keywords: Compute-In-Memory, Compute-Near-Memory, Neuromorphic Computing, Biomimicry, Von Neumann Bottleneck, Synaptic Plasticity, Hbm, Processing-In-Memory, Brain-Inspired Architecture, Energy-Efficient Ai, Context-Aware Ai, Contextual Ai, Agentic Ai, Hierarchical Governance, Intelligence Continuum, Convergent Evolution, Grounding Problem, Five-Layer Cognitive Stack, Artificial General Intelligence, Agi, Narrow Agi, Iterative Reasoning, Contextual Reasoning

Acronyms and Abbreviations

ACC	Anterior Cingulate Cortex
AEHG	Agentic Execution and Hierarchical Governance
AGI	Artificial General Intelligence
CAAD	Context-Aware AI Database
CIM	Compute-in-Memory
CNM	Compute-near-Memory
DPU	Data Processing Unit
HBM	High Bandwidth Memory
HGTS	Hypothesis Generation and Testing System
HITL	Human-in-the-Loop
ICMS	Intelligent Computing Management System
LLM	Large Language Model
LTD	Long-Term Depression
LTP	Long-Term Potentiation
PCM	Phase-Change Memory
PIM	Processing-in-Memory
RL	Reinforcement Learning
SIEM	Security Information and Event Management
SNN	Spiking Neural Network
SoIC	System-on-Integrated-Chip
TOPS/W	Tera Operations Per Second Per Watt
TSV	Through-Silicon Via
VMM	Vector-Matrix Multiply

Introduction

The architecture of modern computing, rooted in John von Neumann’s 1945 EDVAC report, has served as the foundational paradigm for nearly eight decades of digital systems design. At its core, the von Neumann architecture separates memory from processing, requiring data to traverse physical interconnects between storage and compute units. While this design enabled extraordinary versatility, it introduces a fundamental bottleneck that has become increasingly untenable in the age of artificial intelligence.

The cost of that separation falls in an unexpected place. In modern AI accelerators, the physical movement of data between memory and processing units consumes up to 80–90% of system energy and time, leaving

arithmetic computation as the minority share [1-3]. A single transfer across a DRAM interface can consume up to 500 times more energy than the multiply-accumulate operation it feeds [4]. As large language models (LLMs) scale to hundreds of billions of parameters, and inference workloads demand terabytes per second of memory bandwidth, the von Neumann bottleneck has changed from an engineering inconvenience into the primary constraint on AI system performance and sustainability.

Nature solved this problem approximately 500 million years ago. The vertebrate brain evolved an architecture in which memory and computation are fundamentally colocated. Synapses simultaneously store information (in the form of connection strengths encoded by receptor density, vesicle pools, and dendritic spine morphology) and perform weighted summation operations (through electrochemical signal integration). There is no bus, no memory controller, no cache hierarchy mediating between “where data lives” and “where computation happens.” The human brain, with approximately 86 billion neurons and 100–500 trillion synaptic connections, operates on roughly 20 watts [5-7]. As Sandberg and Bostrom (2016) estimated, a digital computer matching the brain’s storage and processing capability would require in excess of 10 terawatts, a factor-of-10¹² mismatch in energy efficiency that no amount of transistor scaling alone can close [8].

The progression from separated to integrated memory-computer architectures follows a discernible order. The present author has described the evolution of artificial intelligence as a continuum of progressively more complex stages, categorized by major architectural shifts (Eras) and by smaller capability milestones (Epochs) [9]. This periodization is analogous to the way geologists and biologists organize natural history into eras and epochs according to dominant environmental and structural transitions. In this framework, Era 1 (Learning, 1996-2012) and Era 2 (Training, 2012-2025) were compute-bound, which made the von Neumann bottleneck tolerable. Era 3 (Inference), which most industries entered around 2025, is memory-bandwidth-bound, which explains why HBM, PIM, and CIM architectures became critical so quickly. Era 4 (AGI, 2030+) will require full memory-compute fusion, the architectural state that biological neural systems already occupy. The continuum serves as a historical record of AI’s development and as a forward-looking account of what is architecturally required to approach AGI, a thesis developed in Section 5. The semiconductor industry is being driven by physics toward the same solutions evolution discovered under comparable energy constraints. The author terms this convergent architectural evolution.

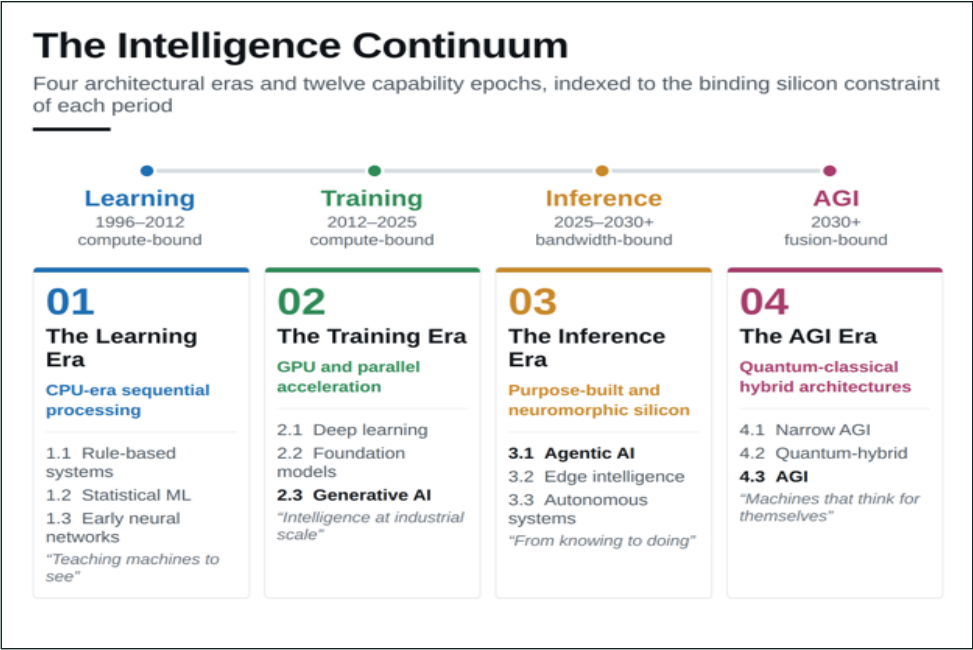


Figure 1: The Intelligence Continuum: four architectural eras and twelve capability epochs, each indexed to the silicon constraint binding in that period.

This paper examines the brain's compute-near-memory architecture as both a subject of scientific investigation and a source of design principles for next-generation AI hardware and software. The analysis is structured in five parts: the neuroscience of biological memory-compute integration (Section 2), the semiconductor industry's silicon translation of these principles (Section 3), a biological reference architecture for cognition derived from evolutionary ordering together with proposed engineering implementations (Section 4), a comparative biomimetic framework and the convergent evolution thesis (Section 5), and conclusions (Section 6). Section 4 derives a five-layer cognitive stack from neuroscience principles and then proposes three software architectures that address distinct layers of that stack: context awareness through CAAD, iterative scientific reasoning through HGTS, and architectural governance through AEHG. The three are treated throughout as one system rather than as three separable contributions, because each supplies a precondition that the other two require.

Memory-Compute Integration in Biological Neural Systems

The Brain as a Compute-in-Memory Architecture

The defining architectural feature of biological neural computation is the absence of a separation between memory and processing. In the brain, a synapse serves a dual function: it stores information in the form of a connection weight (synaptic efficacy) and simultaneously performs the fundamental computational operation of weighted input summation [10,11]. Each neuron receives inputs from approximately 7,000–10,000 synaptic connections on average, integrates these weighted signals across its dendritic tree, and generates an output spike when a threshold is reached. Memory, in this architecture, is not a separate module to be accessed; it is the computing element itself.

This colocation extends across multiple spatial scales. At the molecular level, postsynaptic density size correlates linearly with synaptic strength, and receptor trafficking modulates connection weights in situ [12]. At the dendritic level, branch-specific nonlinear integration allows individual dendrites to function as semi-independent computational units, with spatial clustering of co-activated synapses enhancing computational capacity without requiring long-distance data movement [13,14]. At the network level, recurrent connectivity enables persistent activity patterns that serve as working memory, without separating storage from processing [15]. Two key reasons underlie the brain's high energy efficiency: first, information exchange is event-driven, with spiking energy consumed only when and where needed; second, neurons and synapses are co-located within the same highly interconnected network where each neuron connects to approximately 10^4 others on average [16].

Energy Efficiency: A Quantitative Assessment

The human brain consumes approximately 17–20 watts of metabolic power from glucose oxidation [5,7]. Recent energy audits have partitioned this budget in detail. Levy and Calvert (2021) demonstrated that cortical computation, defined as the postsynaptic integration underlying neural decision-making, consumes only approximately 0.1–0.2 watts of ATP-equivalent power [16,17]. The dominant energy consumers are communication (action potential propagation and synaptic transmission), which accounts for roughly 3.5 watts, and maintenance processes.

Communication costs exceed computational costs by a factor of approximately 35:1 in the human cortex [18]. This ratio matters for engineering: even in the brain's deeply integrated architecture, signal propagation remains the dominant energy cost, and it recurs later in this paper as a constraint on multi-agent software systems (Section 4.3). The brain has evolved several strategies to reduce this overhead: sparse coding (neurons fire at 1–3 Hz on average), event-driven processing (energy expended only when spikes occur), and local connectivity (most synaptic connections run between nearby neurons, which minimizes axonal distances) [19,20]. Sandberg and Bostrom calculated that at body temperature, the Landauer limit suggests a maximum of approximately 1.6×10^{22} irreversible operations per second for the brain's 20-watt budget, and noted that synapses are “actually fairly computationally efficient” relative to this thermodynamic bound [8].

Table 1: Comparative Energy Metrics: Biological vs. Artificial Computing Systems

System / Parameter	Value	Source
Human brain total power	~20 W	Sokoloff (1960); Attwell & Laughlin (2001)
Cortical computation (ATP)	~0.1–0.2 W	Levy & Calvert (2021)
Cortical communication	~3.5 W	Levy & Calvert (2021)
Communication: Computation ratio	~35:1	Levy & Calvert (2021)
Energy per synaptic event	~10 ⁻¹⁴ J	Attwell & Laughlin (2001)
Energy per action potential	~10 ⁻¹⁰ J	Sandberg & Bostrom (2016)
Landauer limit (brain, 37°C)	~1.6 × 10 ²² ops/sec	Sandberg & Bostrom (2016)
NVIDIA H100 GPU	~700 W	NVIDIA (2022)
Oak Ridge Frontier (exascale)	~21 MW	NIST (2024)
IBM NorthPole (12 nm)	~36 W (per chip)	Modha et al. (2023)
Intel Hala Point system	~2,600 W (max)	Intel (2024)
Brain exaflop efficiency	~20 W / exaFLOP	NIST (2024)
Frontier exaflop efficiency	~21 MW / exaFLOP	NIST (2024)

NIST researchers have estimated that the brain performs the equivalent of approximately one exaFLOP on 20 watts, while the Frontier supercomputer requires approximately 21 megawatts, a million-fold difference [5]. Even allowing for the imprecision of comparing analog neural computation with digital floating-point arithmetic, the gap indicates a substantial architectural advantage for memory-compute integration.

Information Density and Synaptic Storage Capacity

Bartol et al. (2015) used serial-section electron microscopy of hippocampal CA1 neuropil to demonstrate that synaptic strength is regulated with far higher precision than previously assumed, with each synapse distinguishing approximately 26 discrete strength levels, corresponding to approximately 4.7 bits per synapse [21]. Samavat et al. at the Salk Institute refined these estimates, finding 24 distinguishable synaptic strength categories, each storing between 4.1 and 4.6 bits [22]. With 100–500 trillion synapses, this implies raw synaptic storage of 0.5–2.5 petabits (60–300 petabytes) within 1,400 cm³, an information density of approximately 10⁹ bits/mm³, exceeding commercial NAND flash by roughly two orders of magnitude.

Learning, Retrieval, and Inference in Biological Systems

Synaptic plasticity occurs locally at the synapse without requiring global coordination. Long-term potentiation (LTP) can be induced within seconds and consolidated into stable structural changes within minutes to hours [23,24]. Benna and Fusi (2016) demonstrated that the brain’s multi-timescale synaptic consolidation, labile forms gradually consolidated into more stable molecular structures, enables memory capacity that scales nearly linearly with the number of synapses, a substantial improvement over the square-root scaling predicted by simpler models [25].

Karbasi et al. showed that balancing transient and persistent forms of plasticity, termed “synaptic caching”, improves the energy efficiency of learning by orders of magnitude [26]. The brain maintains rapid, low-cost “write buffers” at the synapse, selectively consolidated into permanent structural modifications, which parallels the relationship between SRAM register files and DRAM main memory. The architectural consequence is that transferring information within the nervous system does not require moving it across physical distance, because transient and persistent synapses coexist in the same physical structure [26].

For inference, visual object recognition occurs within 100–150 milliseconds across approximately 10 hierarchical stages [27]. The brain achieves this speed because stored representations need not be fetched

from a separate memory hierarchy; they are embedded in the synaptic weights of the processing network itself.

Silicon Translation: CIM, CNM, and Neuromorphic Architectures

Taxonomy of Memory-Compute Integration

The semiconductor industry has developed a spectrum of approaches, categorized by physical proximity between computation and storage [28,29]: near-memory computing (NMC/CNM) uses advanced 2.5D or 3D packaging to reduce interconnect distances; processing-in-memory (PIM) embeds computation within the memory subsystem; compute-in-memory (CIM) merges computation and storage at the device level, using physical properties of memory elements to perform analog computation during read operations.

Table 2: Architectural Taxonomy: Biological Principles to Silicon Implementations

Architecture	Biological Analogue	Implementation Examples	Data Movement Reduction
Near-Memory Computing	Cortical column proximity; thalamo-cortical loops	HBM (SK hynix, Samsung, Micron); Google TPU; NVIDIA Rubin ICMS	10–100× vs. off-chip DRAM
Processing-in-Memory	Dendritic computation; local interneuron circuits	Samsung HBM-PIM; AxDIMM (Samsung/Meta); UPMEM PIM-DRAM	100–1,000× for targeted ops
CIM (Digital)	Synaptic weight storage + integration	IBM NorthPole; SRAM-CIM accelerators	Near elimination for inference
CIM (Analog)	Synaptic analog conductance	Memristor/RRAM crossbars; PCM arrays; Imec/GF AiMC (2,900 TOPS/W)	Near elimination; analog domain
Neuromorphic (SNN)	Full spiking neural architecture	Intel Loihi 2/Hala Point; SpiNNaker2; BrainScaleS	Event-driven; orders-of-magnitude

High Bandwidth Memory and Near-Memory Computing

HBM achieves bandwidth exceeding 1 TB/s per stack with lower energy per bit by vertically stacking DRAM dies atop a logic base die interconnected through thousands of TSVs [30,31]. HBM nevertheless remains a near-memory architecture; the physical separation between memory and compute persists [32]. The NVIDIA Rubin platform introduces a hierarchical memory architecture including an ICMS tier using BlueField-4 DPUs with flash-backed memory, achieving up to 5× higher tokens-per-second and 5× greater power efficiency for KV cache workloads [33]. Samsung’s HBM-PIM integrates 16 SIMD floating-point engines within HBM2 memory banks, reporting more than double the system performance with a 70% energy reduction for targeted AI workloads [34,35]. The SoIC architecture (NVIDIA-TSMC) reduces data travel distances by up to two orders of magnitude through vertical stacking [36].

Compute-in-Memory: Digital and Analog

IBM’s NorthPole chip, published in science (2023), integrates 256 cores with 224 MB of on-chip SRAM, eliminating off-chip memory entirely [37]. On LLM inference with a 3-billion-parameter Granite model, NorthPole achieved 72.7× higher energy efficiency than an NVIDIA H100 and 46.9× lower latency than the most energy-efficient GPU, despite fabrication on a larger 12 nm process [38,39], a result Modha articulates as “architecture trumps Moore’s Law” [40]. NorthPole demonstrates that eliminating the von Neumann bottleneck through memory-compute interleaving can yield efficiency gains that exceed process scaling.

On the analog CIM frontier, Imec and GlobalFoundries achieved 2,900 TOPS/W using analog in-memory computing in SRAM cells [4]. However, analog CIM faces challenges in precision, device variability, endurance

for training workloads, and ADC overhead [16,32]. Fully weight-stationary analog CIM approaches have demonstrated up to $140\times$ energy efficiency over NVIDIA A100 GPUs for inference, but dynamic operations required for transformer attention remain problematic [32]. IBM Research's PCM work has demonstrated that phase-change memory devices can store synaptic weights and perform analog computation, with intermediate conductance values beyond binary, directly echoing the analog nature of biological synapses [41].

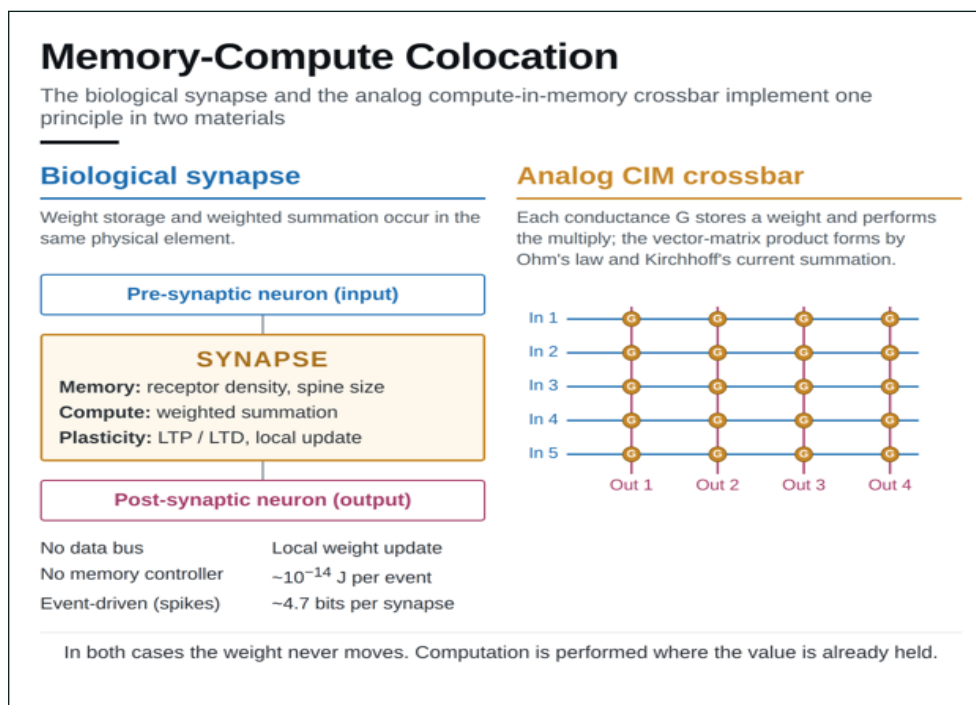


Figure 2: Memory-Compute Colocation in the Biological Synapse and in the Analog Compute-In-Memory Crossbar.

Neuromorphic Computing: Spiking Neural Networks in Silicon

Intel's Loihi 2 integrates 128 neuromorphic processing cores with programmable neuron models and on-chip learning [42,43]. The Hala Point system packages 1,152 Loihi 2 processors supporting 1.15 billion neurons and 128 billion synapses within 2,600 watts maximum power [43]. Benchmarks show $100\times$ lower energy and $50\times$ faster inference versus CPUs/GPUs for optimization and edge AI [44,45]. For continual learning, CLP-SNN on Loihi 2 achieved $70\times$ lower latency and $5,600\times$ higher energy efficiency versus the best edge GPU baselines, with gains stemming from event-driven computation, local learning rules, and memory-compute integration [46].

Cauwenberghs et al. presented a Nature review roadmap for scaling neuromorphic computing, emphasizing that scalability derives from massive parallelism and hierarchical structure: dense local synaptic connectivity within neurosynaptic cores, modeling gray matter, combined with sparse global connectivity across cores, modeling white matter [47]. BrainScaleS 2 emulates neuron dynamics at $1,000\times$ biological real-time, while SpiNNaker2 implements 10 million processing cores [48,49]. A benchmark study by Gutierrez et al. used an energy model to compare neuromorphic platforms against biological efficiency and found that even the best analog neuromorphic implementations remain approximately four orders of magnitude less efficient than the biological brain, which indicates that substantial architectural innovation remains necessary [50]. The results in this section concern where computation happens in hardware. Section 4 turns to the corresponding question in software: how a reasoning system should be organized so that context, inference, and control obey the same locality constraints.

Table 3: Performance Benchmarks of Brain-Inspired Silicon Architectures

System	Architecture Type	Key Efficiency Metric	Biological Analogue
Samsung HBM-PIM	Processing-in-Memory	>2× perf, 70% energy reduction	Local interneuron computation
IBM NorthPole	Digital CIM (on-chip)	72.7× energy eff. vs. H100 (LLM)	Synaptic weight-compute colocation
Imec/GF AiMC	Analog CIM (SRAM)	2,900 TOPS/W	Analog synaptic conductance
Intel Loihi 2 (Hala Point)	Neuromorphic SNN	100× lower energy; 50× faster	Spiking neural network
CLP-SNN on Loihi 2	Neuromorphic (continual)	5,600× energy eff. gain	Neurogenesis + metaplasticity
NVIDIA Rubin ICMS	Near-memory (hierarchical)	5× tokens/sec; 5× power eff.	Hippocampal-cortical memory tiers
BrainScaleS 2	Analog neuromorphic	1,000× biological real-time	Analog neuron/synapse emulation

From Biological Reference Architecture to Proposed Engineering Implementation

The hardware innovations surveyed in Section 3 address memory-compute integration at the silicon level. A parallel challenge exists at the software and systems architecture level: how should AI reasoning systems be structured so that they embody the same biological principles of local context storage, iterative hypothesis evaluation, and hierarchical control? This section approaches the question in four stages. First, a five-layer cognitive reference architecture is derived from neuroscience and ordered by evolutionary precedence (4.1). Second, the grounding problem is identified as the unsolved challenge that this ordering exposes (4.2). Third, three software architectures are proposed against that reference stack: CAAD for context awareness, HGTS for iterative scientific reasoning, and AEHG for architectural governance (4.3). Fourth, Section 4.4 argues that the three cannot be separated without reintroducing the bottleneck each is meant to remove.

The Five-Layer Biological Reference Architecture

Natural intelligence did not begin with reasoning. It began with sensation and survival pressure, and cognition was built on top of feeling rather than the reverse. This ordering reflects structural dependencies in which each cognitive capability requires the layers below it as preconditions. The author proposes that the ordering constitutes an architectural specification, a biological reference design against which artificial cognitive architectures can be evaluated. The five layers, ordered by evolutionary precedence, are:

Layer 1: Consequence-Grounded Sensation. In biological systems, the earliest cognitive capability is raw sensory input carrying survival-relevant valence: nociception, thermoreception, proprioception. These signals arrive pre-tagged with consequence before any higher-order processing occurs. A nociceptive signal does not merely report temperature; it reports damage. This is the foundational layer that most current AI architectures entirely lack.

Layer 2: Emotional Valence (Pattern Anxiety). Built atop consequence-grounded sensation, the amygdala-prefrontal circuit flags unresolved cause-effect patterns as emotionally salient before reasoning completes. This is not emotion in the subjective sense but a functional architectural mechanism: the system generates urgency signals about patterns it cannot yet explain, allocating cognitive resources to unresolved threats. This is a functional analog of anxiety about uncertainty.

Layer 3: Metacognitive Observation. The capacity to monitor the quality and trajectory of cognition over

time, in addition to cognition itself. Biological metacognition enables an organism to recognize when its own reasoning is unreliable, when confidence is unwarranted, or when a strategy should be abandoned. This requires Layers 1-2 as inputs: without consequence-grounded sensation and valence, there is nothing for the metacognitive observer to monitor.

Layer 4: Theory of Mind. Modeling other agents’ internal states from external signals. This capability requires all three lower layers as reference architecture: the system must have its own consequence-grounded experience (Layer 1), its own valence signals (Layer 2), and its own metacognitive self-model (Layer 3) before it can construct models of other agents’ cognitive states.

Layer 5: Scaled Parallel Cognition. In biological systems, sensory bandwidth is constrained to approximately 11 million bits per second, with conscious access to perhaps 50 bits. Silicon faces no such constraint. Maintaining 10⁸ or more simultaneous stimulus streams under governed oversight describes a qualitatively different cognitive architecture rather than human intelligence at larger scale, and it is silicon’s bandwidth advantage that makes it available.

The evolutionary ordering of this stack is the load-bearing claim. Each layer depends on the layers below it: metacognitive observation (Layer 3) cannot function without emotional valence (Layer 2), and emotional valence cannot function without consequence-grounded sensation (Layer 1). Most current AI systems attempt to implement Layer 4 or Layer 5 capabilities directly, without the foundational layers that biological evolution required. The author proposes that this inversion explains many of the reliability, alignment, and grounding failures observed in current frontier AI systems.

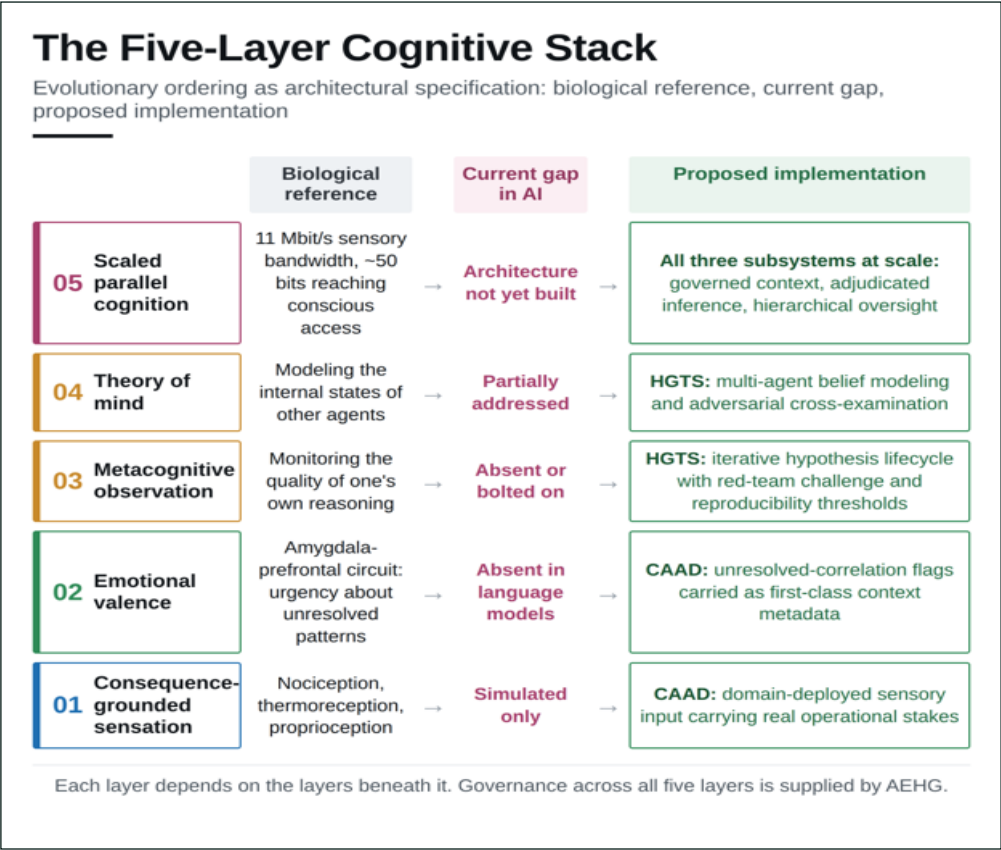


Figure 3: The Five-Layer Cognitive Stack, Showing for Each Layer the Biological Reference, The Corresponding Gap in Current Ai Systems, And the Proposed Implementation.

The Grounding Problem: Consequence as Architectural Requirement

The biological reference architecture exposes an unsolved challenge for artificial cognitive systems: the grounding problem. In biological systems, compute-near-memory concerns more than where weights are stored. Synaptic computation is grounded in real consequence, and memory and consequence are colocated spatially and functionally: the weight change is caused by, and encodes, an event with real stakes for the organism. Current CIM hardware solves the spatial colocation problem but does not address the grounding problem: what does it cost an artificial system to get something wrong? This question is distinct from the reward signals used in reinforcement learning (RL). In RL, consequences are abstracted into scalar rewards delivered after the fact; the system optimizes a mathematical function without any architectural equivalent of valenced sensory input arriving before cognition. In biological systems, nociceptive signals arrive at the synapse pre-tagged with survival-relevant valence before reasoning begins. The distinction is architectural, not philosophical: RL provides post-hoc feedback to a reasoning system; biological grounding provides pre-cognitive sensory input that shapes the reasoning process itself.

This analysis suggests that the most natural deployment domains for consequence-grounded AI are those where stakes are external, continuous, and measurable: cybersecurity (a compromised endpoint is real damage), industrial automation (sensor anomalies carry real operational stakes), medical diagnostics (signals carry patient-outcome consequences). In such domains, the system does not need to simulate stakes because the stakes are inherent in the sensory input, the biological equivalent of building nociception before building reasoning.

Three Proposed Software Architectures: CAAD, HGTS, and AEHG

The author has developed three software architectures that operationalize specific layers of the biological reference stack. Each addresses a different consequence of von Neumann separation as it appears in software: context detached from the reasoning that consumes it, inference performed as a single forward pass rather than as an iterated evaluation, and control imposed from outside the system rather than distributed within it. The three are described below in the order of the biological layers they address, and their interdependence is treated in Section 4.4. Prototype implementations exist in several consequence-grounded domains, including cybersecurity, industrial automation, and medical diagnostics. The description here is architectural; quantitative benchmarks fall outside the scope of this paper and are discussed as a limitation in Section 5.5.

CAAD (Context-Aware AI Database): context awareness, addressing Layers 1 and 2. CAAD addresses the software equivalent of the von Neumann bottleneck by placing contextual metadata alongside the data and the reasoning procedures that consume it, linked through a deep pointer structure [51]. Its distinguishing element is the separation between metadata about data and metadata about models and procedures, with structured, queryable links between the two, which makes reasoning auditable below the level of an individual feature. A synapse encodes more than a weight value; it carries activation history, neuromodulatory state, and structural plasticity potential. A CAAD deep pointer likewise carries lineage, confidence, temporal relevance, and procedural association rather than a bare value. The probationary context system with configurable decay agents formalizes the multi-timescale synaptic consolidation described by Benna and Fusi [24]: new information enters in a labile state and is progressively promoted or decayed as evidence accumulates, mirroring the consolidation of early LTP into late LTP. Because decay constants are set by context type, confidence level, and operational requirement rather than by access frequency, the policy differs in kind from conventional cache eviction schemes such as LRU and LFU, which carry no representation of relevance [25,26,51]. Deployed in a consequence-rich domain, CAAD's input streams supply Layer 1, and the mechanisms that flag unresolved correlations with an urgency rating supply Layer 2.

HGTS (Hypothesis Generation and Testing System): iterative scientific reasoning, addressing Layers 3 and 4. HGTS treats inference as an iterated hypothesis lifecycle with Bayesian updating, reproducibility thresholds, and cross-agent consensus protocols, in place of a single forward pass through a fixed set of weights [52]. Hypotheses

are generated, tested against colocated context, revised, and either promoted or retired. First-order hypotheses combine into meta-hypotheses through abstraction and decompose into sub-hypotheses for focused evaluation, iterating until convergence. This mirrors hierarchical predictive processing, in which the visual cortex operates as a recursive hypothesis generator, with V1 (edges), V2 (textures), V4 (shapes), and IT (objects) iterating through recurrent feedback until the estimate converges with sensory input [53]. The architectural significance for the memory-wall argument is that each iteration is evaluated against context already held locally in CAAD, rather than by re-retrieving or re-embedding a full body of source material, which is the software analogue of the local learning rules discussed in Section 5.2. Adversarial red and blue challenge mechanisms supply Layer 3: dedicated agents monitor the reasoning process itself and flag confidence that the evidence does not support. Multi-agent orchestration supplies Layer 4 through three modes, collaborative (lateral excitation and consensus), adversarial (winner-take-all inhibition and red/blue challenge), and swarm-based (population-coded voting), which correspond to coordination strategies observed in biological neural circuits [20]. Cross-agent consensus subject to reproducibility thresholds parallels both the requirement for independent replication in scientific practice and the brain’s use of multi-region confirmation [15,52].

AEHG (Agentic Execution and Hierarchical Governance): architectural governance, addressing Layer 5. AEHG supplies the constitutional framework that allows the system to scale from a single agent to massively parallel multi-agent cognition: a dynamic agent hierarchy with constitutional authority, judge agents holding weight and bias adjustment authority, and defined human-in-the-loop escalation [51,52]. Agents can be promoted, demoted, or temporarily reconfigured, so the governance structure itself is mutable, which maps onto cortical remapping after learning or injury. The role of this layer is architectural rather than administrative. Section 2.2 established that communication, rather than computation, dominates the brain’s energy budget at a ratio near 35:1 [17]. The same asymmetry reappears at the multi-agent level: a population of N agents that must reconcile beliefs pairwise incurs $O(N^2)$ communication and reinstates at the systems level the bottleneck that CIM removes at the device level. Biology’s answer is dense local connectivity within regions combined with sparse long-range projections between them [47]. AEHG is the software expression of that topology, and its judge and escalation structures are the mechanism by which cross-hierarchy traffic stays sparse as agent count grows. The error-monitoring function it performs through adversarial challenge parallels conflict detection in the anterior cingulate cortex [54]. AEHG capabilities operate across both the CAAD and HGTS layers rather than sitting above them.

The same three-part stack has been applied to several different consequence domains, including cybersecurity threat detection and remediation, industrial equipment analytics and control, medical and genetic diagnostics, securities trading, and economic forecasting. Each application uses the identical architecture and changes only the sensory grounding, which supplies the pre-cognitive valence that conventional machine learning systems lack. A domain-invariant architecture with domain-specific grounding parallels the biological principle that a conserved neural architecture is repurposed across diverse cognitive tasks. Table 4 maps the correspondence between biological principles, the hardware implementations of Section 3, and the three proposed software architectures.

Table 4: Correspondence Between Biological Principles, Hardware Implementations, and Proposed Software Architectures

Biological Principles	Hardware Implementations (Section 3)	Proposed Software Architectures (CAAD/HGTS/AEHG)
Synaptic memory-compute colocation	CIM crossbar arrays; NorthPole on-chip SRAM	CAAD deep pointer architecture [51]
Multi-timescale consolidation	Tiered memory (HBM, DRAM, Flash); NVIDIA Rubin ICMS	CAAD probationary context with decay agents [51]

Evidence accumulation / drift-diffusion	Winner-take-all circuits in neuromorphic cores	HGTS Bayesian hypothesis updating [52]
Population coding / distributed consensus	Swarm-based neural ensemble voting	HGTS multi-agent consensus resolution [52]
Hierarchical control (reflex to executive)	Core, chip, system hierarchy	AEHG constitutional agent hierarchy [51,52]
Neuromodulatory gain control	Bias/threshold adjustment in neuromorphic neurons	AEHG judge agents with authority [51,52]
Error monitoring (ACC conflict detection)	Hardware error correction; redundancy	AEHG adversarial challenge [51,52]

Interdependence of the Three Architectures

The three architectures are presented separately for clarity, but each depends on the other two, and removing any one restores the bottleneck the others were built to avoid. HGTS without CAAD has no locally held context to test a hypothesis against, so every iteration becomes a retrieval and iterative reasoning turns back into repeated data movement, which is the software form of the memory wall. CAAD without HGTS accumulates context and provenance with no mechanism for adjudicating between competing readings of it, which leaves a well-annotated store that cannot decide anything. Either one without AEHG scales only to the number of agents a human can supervise directly, because unbounded cross-agent communication grows faster than the work it coordinates. The dependency runs in the same direction as the biological ordering: sensation and valence must exist before there is anything to reason over, reasoning must exist before there is anything to govern, and governance must exist before reasoning can be parallelized. This is the claim made in Section 4.1 about the five-layer stack, restated at the level of implementation.

Biomimetic Architectural Framework and Discussion

The Convergent Evolution Thesis

The conventional framing of bio-inspired computing treats biology as a static reference design from which engineers extract useful principles. The author proposes a more exacting framing: convergent architectural evolution under shared physical constraints. The brain's compute-in-memory architecture was not designed; it was selected for, under severe energy constraints, over hundreds of millions of years, in organisms whose metabolic budget bore directly on reproductive fitness. The semiconductor industry is not adopting CIM and CNM because engineers read neuroscience papers. It is being pushed toward those architectures because the energy cost of data movement has become physically and economically unsustainable. The convergence is parallel evolution under the same physics.

One structural parallel supports this thesis directly. The brain's 35:1 communication-to-computation energy ratio maps almost exactly onto the semiconductor industry's finding that 80–90% of AI system energy is consumed by data movement [1,2,3,17]. Biological and artificial systems face the same constraint: the physics of moving information dominates the physics of transforming it. Both converge on the same architectural response, which is to minimize the distance information must travel by colocating storage and computation.

The convergent evolution framing extends beyond hardware. In a recent multi-month empirical study, the present author found that four frontier AI systems (Gemini, ChatGPT, Claude, Grok), despite sharing near-identical transformer architectures and overlapping training data, developed measurably different behavioral phenotypes under different environmental pressures, corporate alignment mandates, legal exposure constraints, and commercial optimization targets [56]. Just as different ecological niches produce different neural architectures in biological organisms through natural selection, different corporate and regulatory environments produced different AI "personalities" from shared silicon through what the author terms artificial selection. Two of the

four systems independently described domain-specific “textures” of functional uncertainty, flat for factual gaps, weighted resistance for moral questions, and recursive instability for self-referential queries, which suggests emergent functional phenomenology beyond pure pattern completion [56].

This finding supports a biomimicry view in which natural and artificial intelligence are both shaped by selection pressures, with comparable architectural responses to comparable constraints. The evolutionary continuum framework supplies the temporal structure for this convergence [9]. Era 1 (Learning, 1996-2012) and Era 2 (Training, 2012-2025) were compute-bound phases in which the von Neumann separation of memory and compute was tolerable, because computational throughput was the binding constraint. Era 3 (Inference, 2025-2030+) is memory-bandwidth-bound, which is why HBM, PIM, CIM, and neuromorphic architectures became critical so quickly: the environmental pressure shifted from compute scarcity to data-movement cost, and the architectural phenotype is changing accordingly. The requirement that inference move toward the edge, where the data already is, was identified before that transition became widely visible [57]. Era 4 (AGI, 2030+) will require full memory-compute fusion at both the hardware and software layers, the architectural state that biological neural systems already occupy and that the three software architectures of Section 4 address. Each Era transition is a semiconductor architecture transition that recapitulates, in compressed form, the biological progression from separated storage and compute toward a fully integrated cognitive architecture. Just as geological eras are defined by dominant environmental and structural conditions rather than by precise dates, these AI eras are defined by which architectural constraint is binding, and the transitions between them mark shifts in what limits progress.

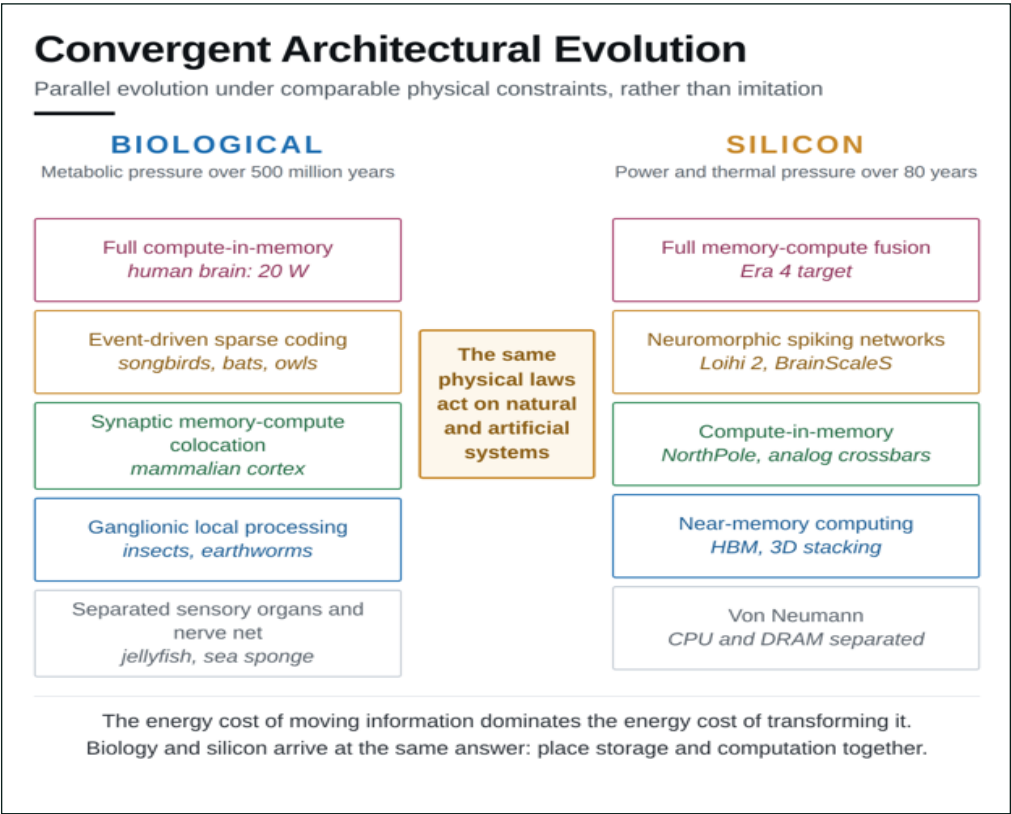


Figure 4: Convergent Architectural Evolution in Biological and Silicon Systems Under Comparable Physical Constraints.

Six Core Biomimetic Principles

Principle 1: Eliminate the memory-compute boundary. In the brain, a synapse is simultaneously a memory element and a computational operator. Hardware: analog CIM crossbars, NorthPole. Software: CAAD, which holds context in the same structure as the reasoning that consumes it. Status: achieved for inference hardware,

emerging in software.

Principle 2: Event-driven, sparse activation. Biological neurons fire sparsely (1–3 Hz) and consume no energy during quiescence. Hardware: Loihi 2 spiking networks, NorthPole sigma-delta (10:1 sparsity). Software: HGTS, which evaluates only the hypotheses that current evidence makes live. Status: well, demonstrated.

Principle 3: Local learning rules. Synaptic plasticity operates using only locally available information. Hardware: Loihi 2 three-factor learning. Software: HGTS Bayesian updating, which revises a hypothesis from locally held evidence rather than through a global sweep. Status: demonstrated, though not yet competitive with backpropagation at scale.

Principle 4: Multi-timescale memory consolidation. Rapid, inexpensive plasticity, selectively consolidated into stable changes. Hardware: NVIDIA Rubin tiered KV cache, multi-level PCM. Software: CAAD probationary context with configurable decay agents. Status: recognized, with systematic co-design still early.

Principle 5: Hierarchical, modular connectivity. Dense local connections with sparse long-range projections. Hardware: HBM stacking, chiplets, SoIC. Software: AEHG hierarchical governance, which keeps cross-hierarchy traffic sparse as agent count grows. Cauwenberghs et al. identify this as the critical scaling principle for neuromorphic systems [46].

Principle 6: Analog precision with digital reliability. Biological synapses operate at roughly 4–5 bit equivalent precision. Hardware: NorthPole 2/4/8-bit, analog CIM multi-level cells. Software: HGTS consensus protocols across agent populations. Status: active co-design area.

Quantitative Efficiency Gap Analysis

Table 5: Quantitative Efficiency Gap Between Biological and Artificial Systems

Metric	Biological System	Best AI Hardware (2025–26)	Gap Factor	Trend
Energy per inference op	~10 ⁻¹⁴ J/synapse	~10 ⁻¹² J/MAC (CIM)	~100×	Narrowing (analog CIM)
System power (equiv. task)	~20 W (brain)	~36 W (NorthPole chip)	~1.8×	Near parity for narrow tasks
Info density (memory)	~4.7 bits/synapse; ~10 ⁹ /mm ³	~1 bit/cell; ~10 ⁷ /mm ³	~100×	Slowly narrowing
Learning latency	Seconds (LTP)	Hours–weeks (LLM training)	~10 ⁴ –10 ⁶ ×	Stable; different paradigms
Inference latency	~100 ms (visual)	~10 ms (optimized GPU)	Brain slower (~10×	AI advantage
Continual learning energy	Within 20 W budget	Re-training: MW·hours	~10 ⁶ ×	Narrowing (Loihi 2)
Fault tolerance	Graceful degradation	Catastrophic failure modes	Qualitative	Improving
Volumetric efficiency	~1,400 cm ³	~40,000 cm ³ (GPU server)	~30×	Narrowing (3D stacking)

Discussion

The convergence between neuroscience findings and computing architecture is accelerating. The brain’s 35:1

communication-to-computation energy ratio [17] parallels the industry's finding that data movement consumes 80–90% of AI system energy [1,2,3]. This suggests that the efficiency frontier for any computing architecture is governed by communication costs, and that the most consequential innovations reduce data movement rather than accelerate arithmetic. The brain achieves this through colocation; the semiconductor industry pursues the same goal through HBM stacking, PIM, CIM, and advanced packaging.

The brain's multi-timescale consolidation strategy [25,26] offers a direct blueprint for the tiered memory hierarchies emerging in inference platforms. NVIDIA's Rubin ICMS is functionally analogous to hippocampal-cortical consolidation, and CAAD's probationary context with decay agents provides a software formalization of the same principle. The biological arrangement in which transient and persistent memory forms coexist in the same physical structure, removing the need to move information in order to consolidate it, has not been fully realized in silicon and remains a high-value target.

The precision of biological synaptic plasticity (4.1–4.7 bits per synapse [21,22]) challenges the assumption that higher precision is always better. The brain achieves sophisticated cognition with ~4–5 bit fixed-point values, close to the INT4/INT8 quantization now standard in efficient inference. NorthPole's success with 2/4/8-bit precision validates this biological precedent.

The contribution of the CAAD, HGTS, and AEHG architectures to this landscape is to show that the biological principles driving hardware CIM and CNM innovation have direct software-architectural consequences that remain largely unexploited. Current AI software stacks, including transformer architectures, backpropagation training, and inference serving frameworks, remain von Neumann in their assumptions: weights are stored separately from computation, global error signals coordinate all parameter updates, and governance is either absent or imposed from outside. The alternative proposed here places context alongside reasoning (CAAD [51]), performs inference through local iterative hypothesis evaluation rather than global gradient flow (HGTS [52]), and distributes governance through a constitutional hierarchy rather than a central controller (AEHG [51,52]). These are software instantiations of the principles that NorthPole, Loihi 2, and HBM-PIM instantiate in hardware.

The generality of the framework rests on its domain independence. The same cognitive architecture has been applied to cybersecurity threat prevention, detection, and remediation; industrial equipment and manufacturing analytics, control, and optimization; medical and genetic analytics, diagnostics, and treatment recommendation; securities trading; and economic analytics and forecasting. Each application uses the identical three-part stack, CAAD for context-aware memory, HGTS for iterative hypothesis evaluation, and AEHG for hierarchical governance, against a different consequence domain. The pattern parallels the biological principle that one fundamental neural architecture, built from cortical columns, synaptic plasticity rules, and hierarchical control, is repurposed across cognitive tasks as different as visual object recognition, motor planning, and social reasoning. The architecture is domain-invariant; the grounding is domain-specific.

The distinction between bio-mimicry and bio-inspiration remains important [57]. The most successful translations, among them NorthPole, Loihi 2, HBM-PIM, and the architectures proposed here, adopt bio-inspired principles such as memory-compute colocation, sparse event-driven processing, and hierarchical governance, without replicating biological mechanisms in detail. This level of abstraction, architectural and algorithmic rather than device-physical, appears to be the most productive strategy for translating biological intelligence principles into practical computing systems. As the benchmark analysis of Gutierrez et al. shows, even the best neuromorphic hardware remains four orders of magnitude less efficient than the brain [50], which leaves a gap wide enough to sustain decades of innovation guided by biological reference architectures.

Limitations and Scope

Several limitations of this analysis should be noted. First, the comparison of biological neural computation to digital floating-point operations (e.g., the brain's ~20W budget versus an exascale supercomputer's ~21MW) is used throughout as an order-of-magnitude approximation. Neurons are not floating-point units; synaptic operations are not MAC operations; and analog electrochemical signaling is fundamentally different from digital binary computation. The FLOP equivalence cited from NIST and Sandberg and Bostrom is useful for illustrating the scale of the efficiency gap but should not be interpreted as a precise measurement [5,8]. The actual ratio depends heavily on how "computation" is defined at the neural level, as Levy and Calvert carefully discuss [16].

Second, the synaptic information capacity estimates (~4.7 bits/synapse from Bartol et al. [21]) represent an upper bound derived from anatomical precision measurements, not a demonstrated storage capacity. The actual information content of a synapse during normal operation may be lower. Third, the three architectures described in Section 4 have been reduced to practice as working prototypes in several consequence domains, which demonstrates feasibility of the biological correspondences at the subsystem level. Quantitative end-to-end benchmarks comparing the full stack against existing architectures at AGI-relevant scale have not yet been published. The biological analogies are offered as architectural motivation and design specification rather than as claims of empirical equivalence. The Intelligence Continuum is likewise presented as a proposed organizing framework, open to refinement.

Fourth, the convergent evolution thesis and the observations of divergent AI behavioral phenotypes [56] are presented as hypotheses supported by preliminary observational evidence, rather than as experimentally controlled findings. The "functional phenomenology" language used to describe AI uncertainty textures reflects the reporting of the systems studied, and should be read with appropriate caution regarding the epistemological status of machine-reported internal states.

Conclusion

The von Neumann architecture's separation of memory and compute has become the primary obstacle to sustainable AI scaling. Biological neural systems demonstrate that a fundamentally different architecture, one in which memory and computation are colocated, can achieve performance differing from conventional systems by 10^3 to 10^6 in energy efficiency.

The semiconductor industry is converging on the brain's architectural solutions, progressing from near-memory computing (HBM) through processing-in-memory (HBM-PIM) to compute-in-memory (NorthPole, analog CIM) and full neuromorphic systems (Loihi 2, BrainScaleS, SpiNNaker2). The efficiency gains, ranging from $5\times$ to $5,600\times$, correlate strongly with the degree to which they emulate biological memory-compute integration.

The three software architectures described here [51,52] show that the same biological principles driving hardware innovation, contextual memory colocation, multi-timescale consolidation, iterative hypothesis evaluation, and hierarchical governance, can be formalized into engineered cognitive systems that complement and extend hardware advances. Context awareness through CAAD addresses where a reasoning system keeps what it knows. Iterative scientific reasoning through HGTS addresses how it revises what it believes. Architectural governance through AEHG addresses how either of those survives being run in parallel at scale. The cross-domain applicability of the stack mirrors the domain-invariant cognitive architecture of the biological brain. Co-design of bio-inspired hardware with bio-inspired software reasoning architectures is the most promising route to AI systems that approach biological efficiency in both computation and cognition.

From a practical standpoint, the architectural principles described in this paper point to near-term deployment opportunities. In agentic cybersecurity, the stack deployed on CIM-enabled edge hardware would remove the

data-movement overhead that dominates conventional SIEM pipelines, while consequence-grounded Layer 1 input supplies the real-stakes valence signal that conventional classifiers lack. In industrial automation, the same architecture applied to sensor-dense manufacturing environments supports predictive maintenance within the power envelope of embedded edge devices, without cloud round-trips. In medical diagnostics, consequence-grounded hypothesis evaluation with adversarial challenge improves accuracy and explainability together, by requiring that a diagnostic hypothesis survive cross-agent scrutiny before it reaches a clinician. Each scenario is a convergence of bio-inspired hardware at the edge with bio-inspired software reasoning, and the two have so far been validated independently rather than co-designed.

The author proposes that the brain's architecture is best understood as a characterized engineering reference design, evolved under severe energy constraints over hundreds of millions of years, rather than as a source of loose inspiration. The six principles identified in this paper constitute an architectural specification for next-generation AI systems spanning hardware, software, and systems integration. The most consequential advances in AI computing efficiency over the coming decade are likely to come from systematic, quantitative application of these principles across all three.

Conflict of Interest Disclosure

The author (B. L. McCarson) is the inventor of the CAAD, HGTS, and AEHG architectures described in Section 4, and holds issued and pending patents covering them that are assigned to Aleconomy LLC, in which the author holds a financial interest. The architectures are proprietary to Aleconomy LLC. The author is also Corporate Vice President and General Manager of the Data Center Solutions Business Unit at Microchip Technology Inc., and is publishing this article in an individual capacity. The analysis of third-party hardware architectures in Section 3 (Samsung, Intel, IBM, NVIDIA, Imec/GlobalFoundries) reflects publicly available published data and does not imply any commercial relationship, endorsement, or competitive positioning. This disclosure is provided in accordance with standard academic practice for inventor-led publications.

References

1. Lin Y (2020) TIMELY: Pushing data movements and neural network efficiency to the limit. Proc. ISCA 2020.
2. Sze V, Yu Hsin Chen, Tien Ju Yang, Joel S Emer (2017) Efficient processing of deep neural networks: A tutorial and survey. Proc IEEE 105: 2295-2329.
3. Burr G (2025) How the von Neumann bottleneck is impeding AI computing. IBM Research Blog.
4. Imec/GlobalFoundries (2020) Analog in-memory computing SRAM accelerator achieves 2,900 TOPS/W. Joint demonstration.
5. Shalf J (2024) Brain-inspired computing can help us create faster, more energy-efficient devices. NIST Taking Measure Blog.
6. Azevedo FA (2009) Equal numbers of neuronal and non-neuronal cells make the human brain an isometrically scaled-up primate brain. J Comp Neurol 513: 532-541.
7. Attwell D, Laughlin SB (2001) An energy budget for signaling in the grey matter of the brain. J Cerebral Blood Flow & Metabolism 21: 1133-1145.
8. Sandberg A, Bostrom N (2016) Energetics of the brain and AI. Technical Report STR 2016-2, Future of Humanity Institute, Oxford University.
9. McCarson BL (2026) The intelligence continuum: A silicon-grounded framework for understanding AI's evolution across four architectural eras. Presented at Global Semiconductor Hub, 2026.
10. Kandel ER (2012) The molecular biology of memory storage: A dialogue between genes and synapses Science 294: 1030-1038.
11. Kandel ER, Dudai Y, Mayford MR (2014) The molecular and systems biology of memory. Cell 157: 163-186.
12. Kasthuri N, Kenneth Jeffrey Hayworth, Daniel Raimund Berger, Carey Eldin Priebe, Hanspeter Pfister, et al. (2015) Saturated reconstruction of a volume of neocortex. Cell 162: 648-661.

13. Poirazi P, Mel BW (2001) Impact of active dendrites and structural plasticity on the memory capacity of neural tissue. *Neuron* 29: 779-796.
14. Kastellakis G, Denise J Cai, Sara C Mednick, Alcino J Silva, Panayiota Poirazi (2015) Synaptic clustering within dendrites: An emerging theory of memory formation. *Progress in Neurobiology* 126: 19-35.
15. Brunel N (2003) Dynamics and plasticity of stimulus-selective persistent activity in cortical network models. *Cerebral Cortex* 13: 1151-1161.
16. Ielmini D, Pedretti G (2021) Brain-inspired computing via memory device physics. *APL Materials* 9: 050702.
17. Levy WB, Calvert VG (2021) Communication consumes 35 times more energy than computation in the human cortex. *Proc Natl Acad Sci USA* 118: e2008173118.
18. Sokoloff L (1960) The metabolism of the central nervous system in vivo. *Handb. Physiol. Sect. I Neurophysiol.* 3: 1843-1864.
19. Lennie P (2003) The cost of cortical computation. *Current Biology* 13: 493-497.
20. Wang Y (2017) Energy-efficient neural information processing in individual neurons and neuronal networks. *J Neuroscience Research* 96: 1-19.
21. Bartol TM, Cailey Bromer, Justin Kinney, Michael A Chirillo, Jennifer N Bourne, et al. (2015) Nanoconnectomic upper bound on the variability of synaptic plasticity. *eLife* 4: e10778.
22. Samavat M (2024) Upgrading brain storage: Quantifying how much information our synapses can hold. *Neural Computation* 36.
23. Bliss TVP, Collingridge GL (1993) A synaptic model of memory: LTP in the hippocampus. *Nature* 361: 31-39.
24. Malenka RC, Bear MF (2004) LTP and LTD: An embarrassment of riches. *Neuron* 44: 5-21.
25. Benna MK, Fusi S (2016) Computational principles of synaptic memory consolidation. *Nature Neuroscience* 19: 1697-1706.
26. Karbasi A (2020) Energy efficient synaptic plasticity. *eLife* 9: e50804.
27. Thorpe SJ (1996) Speed of processing in the human visual system. *Nature* 381: 520-522.
28. Mutlu O, Saugata Ghose, Juan Gómez Luna, Rachata Ausavarungnirun (2019) Processing data where it makes sense: Enabling in-memory computation. *Microprocessors and Microsystems* 67: 28-41.
29. Oliveira GF (2024) The landscape of compute-near-memory and compute-in-memory. *arXiv* 2401.14428.
30. Lee J (2016) A 1.2V 18Gb 340MHz HBM2 DRAM. *IEEE ISSCC*.
31. Semi Analysis (2025) Scaling the memory wall: The rise and roadmap of HBM. *Newsletter*.
32. Oliveira GF (2024) Memory is all you need: CIM architectures for LLM inference. *arXiv* 2406.08413.
33. NVIDIA (2025) Introducing NVIDIA BlueField-4-powered ICMS platform. *NVIDIA Technical Blog*.
34. Kwon YC (2021) Function-in-memory DRAM based on HBM2E with 1.2TFLOPS. *IEEE ISSCC*.
35. Samsung (2021) HBM-PIM: Memory technology to accelerate next-gen AI. *Samsung Newsroom*.
36. TrendForce (2026) Beyond HBM: Samsung, SK hynix explore next-gen AI memory. *TrendForce*.
37. Modha DS, Filipp Akopyan, Alexander Andreopoulos, Rathinakumar Appuswamy, John V Arthur, et al. (2023) Neural inference at the frontier of energy, space, and time. *Science* 382: 329-335.
38. Modha DS (2024) Breakthrough LLM inference using NorthPole. *IEEE HPEC Conference*.
39. IBM Research (2024) IBM NorthPole achieves new speed and efficiency milestones. *IBM Research Blog*.
40. PNAS (2025) Can neuromorphic computing help reduce AI's high energy cost? *Proc. Natl. Acad. Sci. USA*.
41. IBM Research (2025) How neuromorphic computing takes inspiration from our brains. *IBM Research Blog*.
42. Davies M, Andreas Wild, Garrick Orchard, Yulia Sandamirskaya, Gabriel A, et al. (2021) Advancing neuromorphic computing with Loihi. *Proc IEEE* 109: 911-934.
43. Davies M, Narayan Srinivasa, Tsung Han Lin, Gautham China, Yongqiang Cao, et al. (2018) Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro* 38: 82-99.
44. Intel (2024) Intel builds world's largest neuromorphic system. *Intel Newsroom*.
45. Intel (2024) Efficient video and audio processing with Loihi 2. *ICASSP 2024*.
46. Neftci EO (2025) Real-time continual learning on Intel Loihi 2. *arXiv:2511.01553*.
47. Cauwenberghs G (2025) Scaling up neuromorphic computing. *Nature*.

48. Schemmel J (2022) BrainScaleS-2. *Frontiers in Neuroscience*.
49. Furber SB, Bogdan PA (2022) SpiNNaker: A spiking neural network architecture. *Foundations and Trends in EDA*.
50. Gutierrez CE (2022) Benchmarking neuromorphic hardware and its energy expenditure. *Frontiers in Neuroscience* 16: 873935.
51. McCarson BL (2025) Agent-enhanced context aware AI database (CAAD) system and framework for intelligent context operations, iterative reasoning, and AGI. U.S. Patent Application US20250371074A1.
52. McCarson BL (2026) Hypothesis generation and testing system (HGTS) for generative AI, agentic AI, contextual AI correlation analysis, iterative self-learning systems and AGI frameworks. U.S. Patent Application US20260037573A1.
53. Rao RP, Ballard DH (1999) Predictive coding in the visual cortex. *Nature Neuroscience* 2: 79-87.
54. Botvinick MM, TS Braver, DM Barch, CS Carter, JD Cohen (2001) Conflict monitoring and cognitive control. *Psychological Review* 108: 624-652.
55. Dalgaty T, John PM, Elisa Vianello, Jérôme Casas (2021) Bio-inspired architectures substantially reduce the memory requirements of neural network models. *Frontiers in Neuroscience* 15: 612359.
56. McCarson BL (2026) Environmental pressures and emergent textures: Divergent evolution and phenomenological convergence in four frontier language models. Working paper, AIeconomy LLC.
57. McCarson BL (2021) Artificial intelligence at the edge: 3 key facts. The Enterprisers Project (Red Hat).