COMMENTARY · NEUROMORPHIC COMPUTING

The Neuromorphic Market Analysts Have Yet to See

Kevin D. Johnson · kevindjohnson.org

The inaccurate coverage of neuromorphic computing is not just a blog problem. The inaccuracy is a feedback loop. Machine-generated trade posts train on and retrieve the framing of paid analyst reports, and the reports in turn absorb the machine-generated posts as secondary sourcing, so the same misreads circulate between the two with nothing in the loop performing primary verification. A figure enters mangled, gets repeated as though independently confirmed, and hardens into consensus. The framing gets laundered into spreadsheets and priced.

I have spent the better part of a year building the counterexample to that framing. My demonstration program now runs almost sixty builds across twelve silicon architectures, on the BrainChip Akida platform, and a compute ontology stack made possible with IBM Spectrum Symphony, LSF and Storage Scale (GPFS). The distance between what the reports describe and what already runs is the subject of this piece. Three misreads recur and each one narrows the category in a way the hardware does not.

MISREAD ONE

The Milliwatt-Edge Ceiling

The first recurring error confines neuromorphic computing to always-on edge sensing and stops there. One 2026 forecast places edge deployment at more than half the market on the strength of drones, wearables, and IoT sensors. Another frames the entire opportunity around sub-milliwatt workloads that run for months on a single battery. Edge sensing is a real segment

and a valuable one, but a single product cycle has been mistaken for a boundary condition. An analyst who models the ceiling as the market has decided in advance that the category is small.

The hardware does not agree. On Akida silicon I have run dense vision inference in production settings the taxonomy assigns exclusively to graphics processors, including live rail car classification across ten geographically distributed cameras and vehicle recognition from commodity IP streams. The chip the reports describe as an event-only sensor performs dense classification across six industry domains in my demonstration program. The ceiling in the coverage is an artifact of the coverage mindset.

MISREAD TWO

Immature Tooling Read as Undeployable

The second error names fragmented software and weak interoperability as the headline barrier, then treats the barrier as a permanent wall rather than a solvable engineering problem. One report states plainly that the absence of standardized tools blocks integration and forces parallel codebases. The gap is real. Every emerging substrate arrives less turnkey than the incumbent, and graphics processors themselves were awkward to program for neural networks before the surrounding software matured.

A real gap and a disqualifying gap are not the same claim and the reports collapse the two. Neuromorphic silicon can already sit behind the same serving interface and the same schedulers a GPU fleet uses today. I have made Akida a first-class backend for vLLM, the de facto enterprise serving layer, so that the identical OpenAI-compatible endpoint an application already calls now answers from event-driven silicon rather than a graphics processor. I have hot-swapped one hundred models across the fleet at roughly eleven milliseconds each, and I have scheduled the entire fleet under Symphony, LSF and Storage Scale, the same orchestration substrate used across semiconductor design, financial-services workloads at scale and large-scale model training. NVIDIA itself runs IBM Spectrum LSF in its own EDA design chain today, presented as such at the 2026 LSF User Group in Santa Clara, so NVIDIA and Akida now share a common substrate that enables high-performance compute at scale.

A portion of my work runs on the Akida simulation stack rather than on the chip, and the point cuts in my favor rather than against it. A toolchain mature enough to simulate a device with fidelity sufficient to port straight to silicon is not an immature toolchain. Simulation of that quality is itself the maturity signal.

MISREAD THREE

Composition Declared the Losing Strategy

The third error is the one that should stop a careful reader. At least one 2026 market report now frames composition itself as the strategy buyers are abandoning, arguing that enterprises are retreating from multi-chip, multi-vendor approaches toward single-vendor integrated platforms because stitching components together has grown too costly. The claim inverts the direction serious architecture is traveling. One reason analysts may be missing the boat is the shadow the dominant GPU vendor casts over the whole discussion. NVIDIA is consolidating vertically across the entire stack, from silicon through networking, systems and software, and NVIDIA remains the dominant graphics-processor player by a wide margin, so a single-vendor integrated platform is a fair description of where that corner of the market is heading. The neuromorphic market is not that market. No shortage of accelerators exists to compose, and no single vendor owns the category, so the correct architecture is the proper ontology that binds many accelerators together rather than a retreat into one.

My entire body of work is a heterogeneous compute ontology, and heterogeneity is the source of its reach. Under one orchestration domain I have run quantum, neuromorphic, graphics, conventional processors and mainframe as peer resource tiers, including a portfolio pipeline in which an IBM quantum processor encodes features, Akida classifies the market regime and a mainframe settles the result. I have federated three synthetic cortexes across Washington, Dallas and Pittsburgh into a single continental council that reaches k-of-N consensus on real Akida silicon, with no node in charge and meaning crossing the wire through shared state. I have extended the same council into a contested-orbit design in which the network fabric itself is the computer, the graph re-forming as satellites move so that no fixed home node exists for an adversary to strike. I have wrapped neuromorphic cognition in CKKS homomorphic encryption seeded by a quantum random number generator, so that the substrate computes on data it never decrypts. Composition is not the failing approach. Composition is the approach that unlocks capability no single accelerator, and no single vendor, can deliver alone.

The Breadth the Reports Never Model

The narrowness of the analyst frame is easiest to see against the actual span of the work. My demonstration program segments into six industry domains and eight technology themes, and neither list resembles the milliwatt-sensor story the reports tell.

The industry coverage runs across defense and national security, financial services and capital markets, healthcare and public health, critical infrastructure and public safety, information integrity and trust, and AI infrastructure and enterprise IT. The technology themes span neuromorphic LLM serving and inference, hive-mind distributed consensus, on-chip and online learning through spike-timing plasticity, heterogeneous quantum-classical orchestration,

real-time sensor fusion and perception, encrypted cognition and provenance, autonomous ontology construction, and the Symphony, LSF and Storage Scale orchestration substrate itself. A single report category, edge sensing, occupies one corner of one of those themes.

Several individual builds sit entirely outside anything the market analysis contemplates. A full reproduction of BrainChip founder Peter van der Made's complete functional-brain architecture, thalamus through cortex with structural plasticity, learns live rail video with no labels across more than eighteen thousand classifications on a production fleet. A federated pathogen-surveillance system collapses the multi-week single-site outbreak-detection cycle into a same-cycle finding across sites. A media-integrity framework scores manipulation against a brain-encoding model. The through line across every build is not the chip alone but the system around the chip, an intersection the demonstration analysis describes as, at present, unoccupied by the broader market. I have written at length on the architecture behind that claim in my commentary series.

The Data Center Is the Next Front, Not the Edge Alone

The most consequential omission in the analyst frame is the assumption that neuromorphic computing stays at the edge while graphics processors keep the data center. The assumption is already false. Horizontal scale in the data center is a solved demonstration, not a hope. I have held 8,832 concurrent inference contexts across a 46-node cloud fleet at better than 99 percent of ideal scaling, which is direct evidence that neuromorphic inference scales out the way data-center workloads must. The orchestration substrate that makes the scaling possible, Symphony with LSF and Storage Scale, is the same substrate that already schedules national-laboratory supercomputing and large-model training across thousands of nodes.

The research programs currently confined to national laboratories are the second proof. Intel's Hala Point, a 1.15-billion-neuron system built from 1,152 Loihi 2 processors at Sandia National Laboratories, demonstrates that the architecture reaches datacenter scale in neuron count today. The barrier between a national-laboratory research system and a commercial datacenter tier is orchestration and serving, and orchestration and serving are precisely what my heterogeneous compute ontology provides. A neuromorphic fleet that answers the standard vLLM endpoint, schedules under the standard enterprise workload managers and federates across cloud regions is a datacenter citizen, not an edge peripheral. The inference-heavy phase of the AI build-out, now the dominant share of datacenter spending, is the phase in which milliwatt-class inference silicon has the strongest economic claim. The analyst frame that reserves the datacenter for graphics processors is describing the last cycle, not the next.

BrainChip Ships Now, and the Rivals Do Not

The distinction the reports blur most consequentially is availability. Much of the confident language about Loihi and about NorthPole describes hardware a buyer cannot purchase. Intel's Loihi is distributed only to a narrow set of national-laboratory and academic research partners through Intel's neuromorphic research community, and the platform is not commercially available in any real sense. IBM's NorthPole remains a laboratory prototype, not a product in production despite the frequent secondary claim otherwise. BrainChip occupies a different position. The AKD1000 has been commercially available since January 2022, purchasable as a chip or a PCIe board directly from the company, and the newer AKD1500 reached production shipments in mid-2026, fabricated on a 22nm process, drawing under 300 milliwatts, with on-chip learning and defense-grade environmental screening. Akida is the neuromorphic platform a robotics team or a defense integrator can hold, program and deploy today, and every demonstration described above runs on it. An analyst framework that flattens available silicon and unavailable research into a single emerging bucket erases the one distinction a buyer most needs.

Why the Miss Damages Valuation

Bad analysis is not a neutral intellectual failing. Bad analysis is a market force. When the dominant analytical frame undershoots a category systematically, and undershoots it by boxing a composable, scaling architecture into the narrowest available story, valuation tracks the frame rather than the fundamentals. A firm modeled as a single-chip edge-sensor vendor is valued as a single-chip edge-sensor vendor, whatever the underlying trajectory. Price movement is never mono-causal, and I do not claim otherwise. I claim that a framing error of this magnitude and this consistency necessarily impairs the valuation of the names it touches, because the error travels straight from report to model to price without an intervening correction. An egregious miss at the source is an egregious miss at the destination.

The internal inconsistency of the reports is itself the evidence. For the single year 2026 the published market-size estimates for the same market run from roughly 125 million dollars to more than 8 billion, with intermediate figures scattered across the range. A spread of more than sixty times for the same twelve months is not a disagreement about data. The disagreement concerns where each firm draws the boundary of the category, and the boundary is exactly the thing under dispute.

The Power Wall Makes Neuromorphic a Necessity, Not a Luxury

The efficiency argument is usually presented as a benefit. The argument is closer to a requirement, because the graphics-processor build-out the analysts extrapolate is not physically

sustainable on its current trajectory. The International Energy Agency projects electricity generation for data centers to grow from 460 terawatt-hours in 2024 to more than 1,000 terawatt-hours in 2030 and 1,300 by 2035, with an accelerated case approaching 2,000 terawatt-hours, the last figure comparable to the entire electricity consumption of a large industrial nation. Individual campuses now draw between 500 megawatts and a full gigawatt, and the interconnection queues that gate new capacity already hold thousands of gigawatts of pending requests with wait times measured in years. The bottleneck in the AI build-out has shifted from silicon to power, because chips can be fabricated in twelve to eighteen months and grid capacity cannot.

The proposed answers to the power wall are precisely the unproven bets. Small modular reactors are the favored solution, yet the IEA notes the first units enter the mix only after 2030, the technology carries historical cost and schedule overruns of 2.6 times in the United States and the offtake agreements remain conditional. Onsite natural gas fills the near-term gap at a direct carbon cost. Orbital data centers, floated as a further alternative, remain further from proof than the reactors. Every one of those remedies asks the grid, the regulator or the launch vehicle to change before the compute can scale.

Neuromorphic efficiency asks none of that. A workload that draws milliwatts where a rack of graphics processors draws hundreds or even thousands of watts reduces the power problem at the source rather than attempting to generate a way around it. Set against a solution that requires a new nuclear industry or a move to orbit, an architecture already shipping on silicon and already demonstrated at datacenter scaling is the more proven, more immediate lever. The analyst frame treats neuromorphic efficiency as a niche advantage. The power arithmetic makes the same efficiency a condition of the build-out proceeding at all.

The efficiency case does not depend on the power ceiling arriving, and the point deserves emphasis because the ceiling can be dismissed as speculative. Introducing neuromorphic silicon reduces power draw immediately, against today's workloads on today's hardware, before any future limit binds. A task offloaded from a graphics processor to a milliwatt-class chip frees the graphics processor for work only a graphics processor can do, so the same installed base absorbs more useful computation without a single additional rack. The heterogeneous compute ontology compounds the gain, because scheduling each workload onto the substrate that runs it most efficiently, neuromorphic for sparse and event-driven inference, graphics processors for dense training, quantum for the narrow problems it suits, raises the utilization of the whole estate rather than any one tier. The result is not merely a hedge against an imagined power limit. The result is more capability extracted from existing capital now, and exponentially more value as the ontology routes a growing share of work to the substrate that serves it best.

A Larger Model, Offered in the Same Spirit

If the analysts will size the category from the edge inward, the exercise deserves a counterpart sized from the workload outward.

Begin with the compute that already exists. Global AI-datacenter capital expenditure is tracking near 765 billion dollars in 2026 on the way toward 1.6 trillion by 2031, and the build-out has pivoted decisively from training toward inference. Inference is the workload where event-driven, in-memory silicon holds its strongest efficiency advantage, because most inference cycles process data that has not changed. Suppose neuromorphic and neuromorphic-adjacent architectures eventually address between one quarter and one half of inference-dominated infrastructure, the segment where the energy argument is decisive. Against an inference-weighted slice of a trillion-dollar annual build-out, a 25 to 50 percent reach implies an addressable envelope measured in the hundreds of billions of dollars annually, two to three orders of magnitude above the sub-ten-billion figures the current reports print for the category. The gap between those two numbers is the cost of the framing error stated in dollars.

Embodied intelligence widens the envelope again. Humanoid robotics is moving from pilot to production, with Tesla converting vehicle lines to Optimus capacity and Chinese manufacturers led by Unitree entering ten-thousand-unit annual production. Morgan Stanley projects more than one billion humanoid units by 2050. Every one of those machines needs exactly what my SymSEAL and G1-squad demonstrations already run, a low-power reflex and perception layer that learns on the device and cannot depend on a datacenter round-trip. A humanoid running its deliberative vision on a graphics processor and its reflex loops on neuromorphic silicon is the natural division of labor, mirroring the way a biological nervous system delegates reflex to the spinal cord and deliberation to the cortex. At a billion units, even a modest per-unit neuromorphic content places a further multi-hundred-billion-dollar market alongside the datacenter envelope, and the same reasoning extends to every autonomous platform, every satellite, every always-on sensor tier where intelligence must run under a strict power budget.

The most immediate line is not a new market at all but a return on compute already committed, because the return accrues to hardware that exists rather than to a category that must be built. Global AI-datacenter capital expenditure between 2026 and 2031 is tracked near 7.6 trillion dollars across compute, data centers and power. Inference is the workload that converts a fixed capital base into usage, and inference is precisely where event-driven silicon reclaims the cycles a graphics processor spends on data that has not changed. Suppose an ontology that routes sparse and event-driven inference to neuromorphic silicon lifts the effective utilization of that estate by even a tenth to a fifth over the next two to five years and beyond, a conservative assumption against the order-of-magnitude efficiency gaps the architecture shows on suitable workloads. Applied to a multi-trillion-dollar base, a utilization gain in that range is worth several hundred billion dollars to more than a trillion in recovered or redeployed compute value,

output obtained without buying an additional rack, an additional megawatt or an additional cooling loop. The return is not a future product line. The return is more work extracted from money already spent.

The argument to this point concedes the analysts' own frame, in which inference dominates the workload and the efficiency dividend is measured against inference cycles. That concession may understate the case. Inference and training do not exhaust what computation can do. Peter van der Made's cognitive architecture, which I have demonstrated on a production-sized Akida fleet, describes a third mode, a predict-sense-update loop in which a system perceives continuously, is surprised by the novel and accumulates understanding without a retraining run. A system of that kind neither trains in the conventional sense nor merely infers against a frozen model. It learns on the device, in the moment, at milliwatts. The consequence reaches past the datacenter. The large world models the industry now imagines are assumed to live inside one enormous model in one datacenter, yet the same third mode supports the opposite architecture, a world model assembled from many small minds that each learn a local region forever and carry skill rather than knowledge, with the heavy language layer left optional and on the ground. Should that mode mature, neuromorphic silicon would not compete for a slice of the inference budget. It would open a category of work the current hardware cannot perform at any power level, and the efficiency case, already decisive, would become secondary to a capability case. The market sized above assumes neuromorphic silicon does what graphics processors do, only more efficiently. The larger possibility is that it does what graphics processors cannot.

Sum the lines. An inference-infrastructure envelope in the hundreds of billions of dollars a year, an embodied-intelligence layer of comparable and growing scale, and an efficiency dividend from the high hundreds of billions to more than a trillion on compute already being built. Taken together the addressable opportunity is not the single-digit-billion category the reports print. Neuromorphic computing is a multi-trillion-dollar reshaping of how the world spends on compute over the coming decade, and the prevailing coverage has mislocated the number by more than two orders of magnitude. In other words, neuromorphic computing as a market is a multi-trillion-dollar possibility that the industry has been pricing as a niche.

None of the larger figures is a promise. Every one of them, though, follows from taking the workload seriously rather than sizing the category from a single product cycle. My demonstrations exist to show the larger shape, from a milliwatt sensor to a continental council to an encrypted quantum-neuromorphic pipeline, all of it running now. The bull case is not merely intact. The bull case is larger, and differently shaped, than the coverage has been willing to describe, and the difference is not a rounding error. The difference is a market where analysts understand the actual state of all things neuromorphic, a market energized by a heterogeneous compute ontology that already works today.

Sources

Prior post on AI-generated neuromorphic coverage
https://www.linkedin.com/feed/update/urn:li:activity:7478845073632161792/

Demonstration program, segmented analysis
https://kevindjohnson.org/KDJ_Demonstration_Analysis.pdf

Commentary series on heterogeneous compute and neuromorphic architecture
https://kevindjohnson.org/category/commentary

BrainChip, production shipments of the AKD1500 processor
https://brainchip.com/brainchip-announces-commercial-availability-and-production-shipments-of-akd1500-neuromorphic-processors

Persistence Market Research, edge share and upper-bound 2026 sizing
https://www.persistencemarketresearch.com/market-research/neuromorphic-computing-market.asp

Mordor Intelligence, sub-milliwatt framing
https://www.mordorintelligence.com/industry-reports/neuromorphic-chip-market

Cognitive Market Research, tooling and interoperability barrier
https://www.cognitivemarketresearch.com/neuromorphic-computing-systems-market-report

Meticulous Research, integrated-platform thesis
https://www.meticulousresearch.com/product/neuromorphic-computing-market-6492

Fortune Business Insights, lower-bound 2026 sizing
https://www.fortunebusinessinsights.com/neuromorphic-chips-market-111466

The Business Research Company, intermediate 2026 sizing
https://www.researchandmarkets.com/reports/5948656/neuromorphic-computing-market-report

Goldman Sachs, AI build-out capital-expenditure trajectory
https://www.goldmansachs.com/insights/articles/tracking-trillions-the-assumptions-shaping-scale-of-the-ai-build-out

Big Tech AI infrastructure spending and the training-to-inference pivot
https://tech-insider.org/big-tech-ai-infrastructure-spending-2026/

Kaiso Research, humanoid-robot production and deployment analysis
https://www.kaisoresearch.com/blog/humanoid-robot-market-industry-analysis

Morgan Stanley, humanoid-robot long-range market projection
https://www.morganstanley.com/insights/articles/humanoid-robot-market-5-trillion-by-2050

IEA, Energy and AI, data-center electricity demand trajectory and supply mix
https://www.iea.org/reports/energy-and-ai/energy-supply-for-ai

Nuclear timelines and cost-overrun history for AI power
https://io-fund.com/artificial-intelligence/nuclear-energy-ai-data-centers