Vector+ and LatticeRunner

A Neuromorphic Memory Engine for the Next Wave of AI Infrastructure

EXECUTIVE SUMMARY

Modern AI systems are increasingly bottlenecked not by compute, but by memory. Retrieval-augmented generation (RAG) pipelines struggle with slow vector search, unpredictable recall, and rapidly escalating GPU memory and power costs. As context windows grow and inference workloads scale, vector databases, embedding lookups, and KV caches are becoming dominant cost drivers in AI infrastructure.

Vector+ and **LatticeRunner** address this problem with a neuromorphic semantic memory engine designed to replace the most expensive part of today's retrieval pipelines while running efficiently on existing commodity GPUs.

Vector+ enters the market immediately as an acceleration layer for local and enterprise RAG systems, targeting the hot path of retrieval: identifying relevant memories quickly and predictably. LatticeRunner extends the same memory substrate to datacenter-scale inference, offering a path toward dramatically lower memory footprints, reduced power consumption, and new approaches to long-context inference beyond traditional KV caches.

Together, they form a unified roadmap as a revenue-generating product today, and a scalable memory technology positioned at the center of the next generation of AI infrastructure.

The Problem: Memory Is Now the Bottleneck

The dominant constraints in AI systems have shifted:

- **Vector search is expensive:** As embedding counts grow, similarity search consumes increasing GPU time, memory bandwidth, and power.
- **Recall is unpredictable:** Approximate nearest neighbor (ANN) indexes trade accuracy for speed, producing variable and sometimes unstable retrieval behavior.
- **GPU memory is scarce and costly:** Every gigabyte used for retrieval reduces capacity available for model weights, batch size, or throughput.
- **Context windows don't scale cleanly:** KV caches grow linearly with context length, driving up memory use and inference cost.

These issues compound at scale. In retrieval-heavy systems, even small inefficiencies on the datacenter side translate into higher power bills, increased data footprints, lower throughput, potential network congestion, and reduced utilization of expensive GPU hardware. The drive to scale by adding more GPUs is a hard pressure. On the user side it means higher cloud bills and slower access times when seconds actually count.

The Solution: A Physics-Inspired Memory Substrate

Vector+ and LatticeRunner replace traditional vector similarity search with a neuromorphic, attractor-based memory system inspired by associative memory and energy-based models.

Rather than repeatedly computing distances between embeddings, the system encodes semantic information into a structured memory lattice. Queries converge toward stored patterns through deterministic dynamics, producing predictable recall behavior even under noise or partial input.

Key properties:

- **Predictable recall:** Memory retrieval converges toward stable attractors rather than relying on approximate distance heuristics.
- **Lower power usage:** Retrieval avoids repeated high-cost vector distance calculations.
- **Sublinear Memory Growth:** Leverages a shared memory substrate and sparse data representations to store thousands of semantic patterns within a fixed, compact GPU footprint.
- **Noise tolerance:** Retrieval remains robust under partial or corrupted inputs.

This approach has been validated experimentally, achieving strong correlation against FAISS-based SBERT benchmarks as well as Nomic embedding models even under substantial noise, and improving as lattice scale increases.

PRODUCT OVERVIEW

Vector+ (Near Term Product)

Vector+ integrates into existing RAG pipelines by accelerating the core retrieval step, the part of the system responsible for identifying relevant memories. It is the top layer powered by the LatticeRunner technology (which is sort of the “Intel Inside” of the product).

In practical terms, Vector+ replaces the slowest and most expensive part of a vector database, the similarity search. Existing metadata filters, rerankers, and fallback logic remain unchanged. Teams do not need to rewrite their data models or application logic.

Vector+ has been benchmarked today at 100,000 stored patterns on a 16GB NVIDIA 4080 Super. GPU. It has sub-second access times for recall and no indexing step. Multi-lattice configurations can linearly scale this further while preserving predictable recall behavior. The upper bounds of LatticeRunner storage have yet to be discovered experimentally limited only by available VRAM.

This already covers a large fraction of real-world RAG deployments, particularly local, on-prem, and hybrid systems where GPU memory is a limiting factor. This system will work on currently in place commodity GPU models.

LatticeRunner (Datacenter-Scale Extension)

LatticeRunner is the memory substrate underlying the entire system. It scales linearly with the neuron grid size and can be expanded (theoretically) to multi-GPU and multi-node environments. It targets the largest and fastest-growing costs in AI infrastructure:

- GPU memory footprint during inference
- Retrieval-heavy workloads at scale
- Power and thermal budgets in inference clusters

At datacenter scale, even modest efficiency gains matter. Reducing retrieval overhead by single-digit percentages can save millions annually across large inference fleets. By reducing memory bandwidth pressure and repeated distance computation, LatticeRunner offers a new optimization layer for inference pipelines.

Longer-term, LatticeRunner provides a path toward augmenting or partially replacing KV caches, enabling larger effective context windows without linear growth in memory usage. Because of its unique design focused on complete binary functionality, there is a direct line into hardware as well. At the moment it lives on the GPU taking advantage of the massive parallelism by bit-packing its values into pixels.

LatticeRunner is designed for binary functions, no floating points (which take up space and processing time) just fixed-point integers in INT8. The target goal is eventual progression into hardware such as FPGA chips.

MARKET OPPORTUNITY

Near-Term Market: Vector+ (RAG Infrastructure)

The market for Retrieval Augmented Generation is expanding rapidly as enterprises deploy AI systems dependent on fast, accurate retrieval. Analyst forecasts placed the RAG market at approximately $2 billion in 2025, growing toward $10 billion by 2030.

Vector+ targets the segment of this market tied directly to retrieval infrastructure: vector search, embedding storage, and GPU-accelerated lookup systems.

- **Serviceable Available Market (SAM):** Estimated at $2.5-$3.5 billion under conservative assumptions.
- **Serviceable Obtainable Market (SOM):** Early adoption by local RAG developers, small and mid-sized enterprises, and on-premises deployments support $12-$35 million in revenue within three to five years, with upside beyond $40 million. Even one tenth of that project is still profitable.

Vector+ provides the potential for immediate value, early revenue, and validation of the underlying memory technology.

Long-Term Market: LatticeRunner (Datacenter AI Memory)

AI datacenters are entering a historic expansion cycle. Global AI infrastructure spending already reaches hundreds of billions annually and is projected to grow into the trillions over the next decade.

The portion of this spend tied to memory footprint, retrieval efficiency, inference throughput, and power consumption represents the Total Addressable Market for LatticeRunner.

- **Conservative SAM:** $50-$150 billion
- **Upside SAM:** $200-$300 billion as long-context and retrieval-heavy architectures become standard

Even modest penetration of this market translates into substantial revenue opportunities. Early deployments focused on high-volume inference and retrieval workloads alone represent tens to hundreds of millions in potential revenue.

GO-TO-MARKET STRATEGY

Phase 1: Local RAG (Year 1)
Ship a working Vector+ demo, release an open-source core, and offer a paid performance tier. Target developers already experiencing vector search bottlenecks. Benchmarks and a live demo dashboard drive early adoption. The first prototype is nearly ready.

Phase 2: Enterprise RAG (Year 1-2)
Convert developer traction into enterprise pilots. Offer on-prem and VPC deployments, enterprise features, and hands-on support. Produce case studies demonstrating latency, memory, and cost improvements.

Phase 3: Datacenter Optimization (Year 2-3)
Scale the substrate to multi-GPU environments via LatticeRunner. Integrate with inference frameworks and pursue partnerships with cloud providers and hardware vendors.

Phase 4: LLM Memory Integration (Year 3+)
Position LatticeRunner as a memory layer for next-generation LLMs, augmenting or replacing KV caches and enabling larger effective context windows.

IP STRUCTURE AND FUNDING

The core memory technology is held in a dedicated IP-owning entity, probably as an LLC. Vector+ is the operating company commercializing the first product built on this substrate.

Both entities are founder-controlled, ***with contractual guarantees ensuring Vector+ holds a perpetual, preferential license to the core LatticeRunner technology.***

We are seeking $25,000-$50,000 in early angel funding to complete the Vector+ demo, finalize benchmarks, and prepare for a formal pre-seed round. This funding supports three to five months of focused development.

With a working demo in hand, the company will be positioned to raise a $250,000 pre-seed to expand enterprise pilots and advance the datacenter roadmap. There is already interest from the Portland State Business Accelerator and a few other local organizations. But they need to see traction and demos to move forward. Plus, the fastest funding process with any of them is six months to a year out. Although they will supply office space and connections as well as access to tools and other resources at steep discounts.

VISION

Vector+ delivers immediate value as a practical replacement for the most expensive part of today's RAG systems. LatticeRunner scales that same memory engine to the datacenter.

As inference workloads grow and memory becomes the dominant cost driver, this technology positions the company at the center of the next wave of AI infrastructure by reducing power, improving throughput, and extending the useful life of existing GPU deployments.

# # #