TokenForge White Paper

VectorNine Systems Inc. | October 2026

TokenForge puts multi-terabyte unified memory for local AI inference on a single standard 120 V / 15 A wall outlet, with no connection to any cloud.

1. Executive Summary: The Enterprise Mandate for Localized Inference and Zero-Trust AI

Enterprises want frontier-scale AI working on their most sensitive data. Both of today's routes make that hard. A cloud API sends prompts, documents and decisions to a third party. A legacy liquid-cooled GPU cluster keeps the data home, but it costs millions of dollars and needs a purpose-built facility before the first model runs.

The root cause is memory. The largest open-weight foundation models need hundreds of gigabytes to terabytes of memory to run at all. Conventional accelerators sell that memory only when bundled with premium compute, three-phase power and liquid cooling, which is why capacity is so expensive to scale.

VectorNine Systems Inc. is engineering TokenForge to break that bundle. TokenForge is a dedicated inference appliance that pools multi-terabyte unified memory (VRAM) across modular compute blades. It runs from a single standard wall outlet and has no connection to any cloud.

The price-to-capability case

	Legacy liquid-cooled GPU rack	Desktop 1-Unit	Scaled array, wall outlet	Scaled array, rackmount
Configuration	One rack of enterprise GPUs	8 blades in one enclosure	Up to 56 blades, daisy-chained	56 blades in 7 Gen4 rackmount enclosures
Memory available to models	Up to 13.4 TB of GPU memory	Over 1.5 TB of unified VRAM	Over 9.2 TB of unified VRAM at 56 blades	Over 9.2 TB of unified VRAM
Capital cost	Upwards of $4,000,000 CAD	A fraction of the rack's capital cost	A fraction of the rack's capital cost	A fraction of the rack's capital cost
Facility buildout	Hundreds of thousands of dollars in three-phase power and plumbing	None	None	None beyond standard server-room rack PDUs
Power	More than 120 kW, three-phase	Under 1,440 W, on one 120 V / 15 A outlet	Under 1,440 W, on one 120 V / 15 A outlet	Roughly 10 kW
Execution	Not compared here	Full, unrestricted maximum throughput	Power-budgeted: total throughput bounded by the 1,440 W envelope	Unthrottled: 56 independent agents at once

	Legacy liquid-cooled GPU rack	Desktop 1-Unit	Scaled array, wall outlet	Scaled array, rackmount
Heat to remove	More than 120 kW, by liquid cooling	Under 1,440 W	Under 1,440 W	Roughly 10 kW
Operating cost	High, with recurring subscription fees	Minimal. Zero software subscription lock-in	Minimal. Zero software subscription lock-in	Minimal. Zero software subscription lock-in

Rack figures are industry baselines, drawn from published vendor specifications and published estimates of rack cost, power and cooling. TokenForge figures are VectorNine design targets.

TokenForge delivers multi-terabyte localized memory density at a fraction of the capital expenditure of traditional enterprise GPU racks, running with minimal OpEx and absolutely zero software subscription lock-in.

Three facts carry the case:

- **Roughly 1/80th of the power, on a wall outlet.** Run continuously at their ceilings for a year, a 120 kW rack consumes over one million kWh. A TokenForge array on one outlet consumes under 13,000 kWh.
- **Roughly 1/12th of the power, at full speed.** Even unthrottled, the 56-blade rackmount array draws about 10 kW, which is under 90,000 kWh a year.
- **Six times the memory on the same outlet.** Growing from 8 blades to 56 takes unified VRAM from over 1.5 TB to over 9.2 TB without a new circuit. On that outlet, total throughput stays bounded by the 1,440 W envelope.

These figures compare memory capacity, cost and facility requirements. They are not a throughput comparison.

What this means for the business

Executive priority	What TokenForge changes
Return on investment	Multi-terabyte model memory at a fraction of the upwards of $4,000,000 CAD that a legacy GPU rack costs. No facility buildout for power or plumbing, no per-token metering and no software subscription lock-in.
Data sovereignty	Model weights, prompts and retrieved documents stay inside the appliance. No third party processes, stores or logs them.
Zero cloud dependency	Inference needs no outbound connection, no license check-in and no telemetry endpoint. A provider outage or price change cannot stop your workloads.
Rapid deployment	An array of up to 56 blades runs from one standard 120 V / 15 A wall outlet. There is no electrical retrofit to schedule before the first model runs.

Why zero-trust AI points to local inference

A zero-trust posture assumes no external party is trustworthy by default. An inference service you do not operate is exactly such a party, and it sees your data in plaintext. TokenForge removes that party from the path instead of asking you to trust it.

2. The Edge Bottleneck: Cloud Latency and Shared-Accelerator Queuing

AI at the edge is limited by three things the buyer does not control: the network, the queue and the memory ceiling.

The network

Every cloud inference request is a round trip. The prompt leaves the building, crosses the public internet, waits at a provider's front door and returns. The model's own compute time is only part of what the user waits for, and the rest varies with conditions nobody on your team can tune. If the link drops, inference stops.

The queue

Shared accelerators serialize work. On a multi-tenant endpoint, or an in-house GPU pool shared between teams, a request waits behind whatever arrived first. Agentic workloads make this worse, because one task is a chain of model calls and each call joins the queue again. The delay compounds with every step.

The memory ceiling

A model that does not fit in memory does not run, or runs with part of itself offloaded to slower tiers and pays that penalty on every token. Accelerator memory is fixed at manufacture, so the conventional fix is to buy more accelerators to get more memory. At the scale of the largest models, that means a liquid-cooled rack drawing more than 120 kW of three-phase power.

How the three options compare

Constraint	Cloud inference API	Legacy liquid-cooled GPU cluster	TokenForge
Where data is processed	A third party's data center	Your site	Inside the appliance
Network in the inference path	Public internet round trip	Local network	None outside your site
Concurrency	Shared, multi-tenant queue	Shared between internal teams	Requests occupy different blades at the same time
Scaling memory	Buy a larger service tier	Buy more accelerators	Add blades and enclosures, up to 56 blades and over 9.2 TB
Power infrastructure	The provider's	More than 120 kW, three-phase	One standard 120 V / 15 A wall outlet, or standard rack PDUs for unthrottled 56-blade operation
Cost model	Metered per token, indefinitely	Upwards of $4,000,000 CAD in capital, plus facility buildout and recurring subscription fees	One capital purchase. Zero software subscription lock-in
Ongoing dependency	Provider uptime, pricing and terms	Vendor ecosystem	None for inference

3. The TokenForge Architecture

TokenForge is a set of modular compute blades, orchestrated by an Autonomous Head Node and joined by a proprietary high-bandwidth PCIe switching fabric. The internal design is proprietary. This section describes what the system does, not how it is built.

What is in the appliance

Element	What it gives you
Autonomous Head Node	Orchestrates the array and assigns workloads to blades
Modular compute blades	High-density enterprise APUs with dedicated AI acceleration. Each blade can run as an independent inference node
Unified VRAM pools	Memory on every blade, combined into one multi-terabyte pool for models
Dedicated local flash storage	Model weights reside on the blade that runs them, so nothing is fetched when a request arrives
Proprietary PCIe switching fabric	Joins the blades into one system and manages how workloads are assigned across them

Form factors and scale

Configuration	Blades	Unified memory	Power	Execution
Desktop 1-Unit Enclosure	8	Over 1.5 TB	Under 1,440 W, on one 120 V / 15 A wall outlet	Full, unrestricted maximum throughput
Scaled array on a wall outlet (power-budgeted swarm)	Up to 56, daisy-chained	Over 9.2 TB at 56 blades	Under 1,440 W, on one 120 V / 15 A wall outlet	Total throughput bounded by the 1,440 W envelope
Gen4 Rackmount Scaled Array	56, in 7 enclosures	Over 9.2 TB	Roughly 10 kW, across standard server-room rack PDUs	Unthrottled: 56 independent agents at once, at maximum speed

The architecture is the same in all three. What changes is how much power the array is given, and therefore how much of it can run at full speed at once. The upgrade path is modular: start with one desktop enclosure and add blades without changing the outlet. The Gen4 rackmount array is built for IT integration and for small-to-medium commercial and industrial deployments.

Power: a wall outlet, not a facility project

Unlike a single liquid-cooled rack of traditional enterprise GPUs, which requires more than 120 kW of three-phase facility power, TokenForge is engineered to operate strictly within the National Electrical Code (NEC) and Canadian Electrical Code (CEC) 80% continuous load limits. How much of the array runs at once depends on the power it is given.

Desktop 1-Unit, on a wall outlet. The 8-blade base unit draws under 1,440 W at full, unrestricted maximum throughput. It runs on a standard 120 V / 15 A wall outlet with zero facility electrical upgrades.

Scaled array, on a wall outlet (power-budgeted swarm). A user can daisy-chain up to 56 blades on a single residential outlet. Proprietary power management holds the array inside the 1,440 W envelope by bounding total throughput. Massive foundation models and large agentic swarms both run within that envelope. Adding blades on the outlet adds memory and agents, not total throughput.

Scaled array, in a rack (about 10 kW). For unthrottled, simultaneous execution of 56 independent agents at maximum speed, the Gen4 rackmount array scales power linearly. It draws roughly 10 kW, distributed across standard server-room rack PDUs.

The arithmetic is simple. Both codes limit a continuous load to 80% of the breaker rating. On a 15 A, 120 V circuit that is 12 A, or 1,440 W. That ceiling covers everything on the circuit, so an array on a wall outlet should have the circuit to itself. Seven enclosures at full speed draw seven times that figure, which is where the 10 kW comes from.

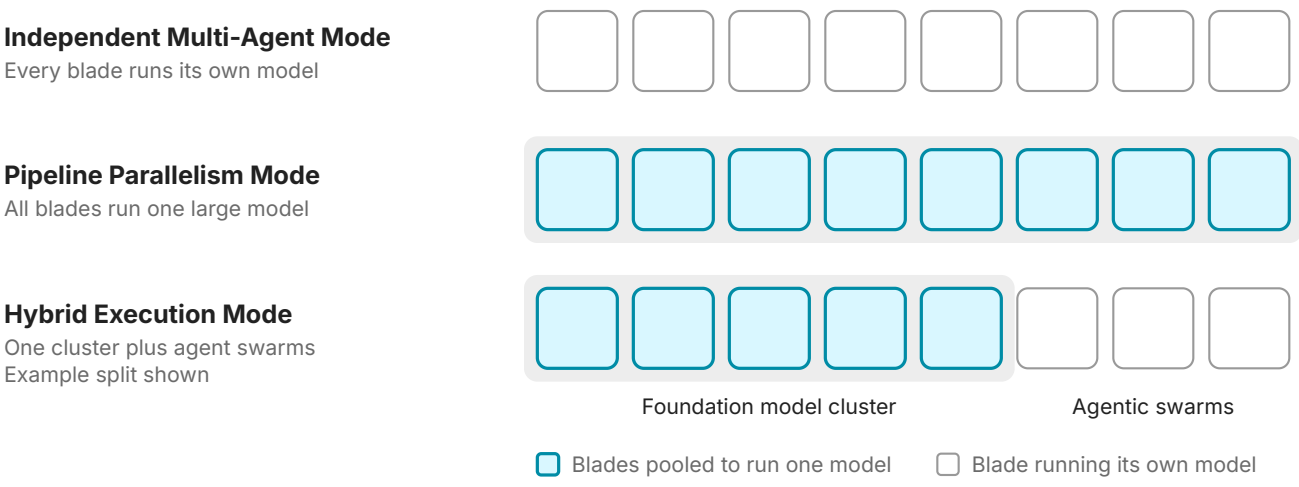
Heat follows power. Everything an appliance draws leaves it as heat. On a wall outlet the array gives off no more than 1,440 W, about what one portable space heater produces. The unthrottled rackmount array gives off about 10 kW. That calls for server-room cooling, and it is still roughly 1/12th of the heat a liquid-cooled rack has to pump away.

Execution modes

TokenForge supports three execution modes. They are logical workload assignments managed by the proprietary PCIe switching fabric. Changing mode changes which blades work together. It does not change the hardware.

Mode	Built for	Workload assignment
Independent Multi-Agent Mode	Many specialized models at once	Each blade operates as an independent node running dedicated models in parallel swarms.
Pipeline Parallelism Mode	The largest foundation models	Distributes unified memory across blades to run massive frontier foundation models as a single pipeline.
Hybrid Execution Mode	Mixed workloads	Dynamically partitions the array, running large foundation models across a blade cluster while remaining blades execute independent agentic swarms simultaneously.

One array, three ways to assign its blades



The diagram shows one 8-blade enclosure under each mode. The hybrid split is an example. The array partitions dynamically. On a power-budgeted array, the modes run inside the 1,440 W envelope.

Precision and scope

TokenForge is a dedicated inference architecture. It natively accelerates BF16, FP16, INT8 and INT4 workloads, executing both full-fidelity half-precision models and highly optimized 8-bit and 4-bit quantized models.

- **Half precision (BF16, FP16).** Full-fidelity execution, at the precision most open-weight models are released in.
- **Quantized (INT8, INT4).** 8-bit and 4-bit models take a half or a quarter of the memory. Quantization trades a small, measurable amount of accuracy for that density, and the amount depends on the model and the method.
- **Not a training system.** TokenForge does not perform ground-up training of foundation models. That remains the work of large training clusters.

Enterprises choose the precision that fits the job: half precision where no accuracy can be sacrificed, quantization to maximize VRAM density. For a trillion-parameter model:

Precision	Weights for a 1-trillion-parameter model	Desktop 1-Unit (over 1.5 TB)	56-blade array (over 9.2 TB)
BF16 / FP16, 16 bits per parameter	2 TB	Exceeds capacity	Fits
INT8, 8 bits per parameter	1 TB	Fits	Fits
INT4, 4 bits per parameter	0.5 TB	Fits	Fits

These are weight sizes only. Working memory for context is additional. On a 56-blade array, a trillion-parameter model runs at full-fidelity half precision, without sacrificing accuracy.

Multimodal workloads

The unified VRAM pools and parallel execution logic are designed for more than large language models (LLMs). They are also built for heavy multimodal inference, including spatial diffusion, text-to-image and image-to-video generative workloads.

Bare-metal advantages

- **No shared tenant.** The appliance serves one owner. No other customer's workload competes for its memory or its compute.
- **Weights are resident.** Request latency never includes a download or a cold start from remote storage.
- **Deterministic hardware routing.** Every device comes up in the same place on every boot, so behaviour your team characterizes once stays characterized.

4. True Air-Gapped Security

TokenForge performs inference with zero telemetry and zero outbound calls: no license check-in, no usage reporting and no runtime negotiation with any external service.

What stays inside

- **Model weights** are held on each blade's dedicated local flash storage.
- **Prompts and responses** are processed in the appliance's own memory.
- **Retrieved documents** used for vector retrieval never leave local storage.

Deployment in three steps

1. **Power.** Connect the array to one standard 120 V / 15 A wall outlet, or to rack PDUs for unthrottled 56-blade operation.
2. **Load.** Place your open-weight models on the appliance's local flash storage. This is the only time weights move.
3. **Run.** Operate the appliance on a network segment with no route to the internet. Inference does not attempt to reach one.

Nothing in that sequence involves a vendor account, an activation server or a usage report.

Zero-trust by construction

Most zero-trust programs spend their effort verifying external services that handle sensitive data. With TokenForge there is no external service in the inference path to verify. Prompts, documents and outputs never cross the boundary of the appliance, so they never appear in a provider's logs, a transit network or another jurisdiction.

What an air gap does and does not cover

An air gap removes the network path, which is the path most data leaves by. Physical access to the appliance and the provenance of the models you load remain under your own controls, where a zero-trust program expects them to be.

TokenForge specifications in this paper are forward-looking design targets and are subject to change. Detailed electrical and installation specifications will be provided in the Technical Data Sheet and installation manuals at commercial release.