AEVION QVenture — Investment Memo

Generated 2026-07-10 · AEVION AI Investment Analyst · not investment advice

Nova Compute

AI Infrastructure / Tooling · seed · US · raising $6,000,000

Score 65.5/100 — WATCH (conviction: medium)

Investment memo

Verdict: watch, with a conditional small check rather than a lead. Nova has demonstrated something rare at seed—real paying demand and a measurable 41% cost reduction with 99.95% uptime, proving the arbitrage genuinely works and customers will pay for it. But the single decisive reason against is structural: this is a commoditizing middle layer where hyperscalers and inference incumbents (Bedrock, Vertex, Together, Fireworks) can absorb cross-cloud routing as a native feature overnight, and a single cloud revoking spot-reseller access via ToS could collapse the business with no recourse. The "data scale" moat is not real at 22 teams. That risk-reward doesn't justify leading. Concretely: pass on the lead, but offer a $1.5–2M participating ticket for ~7–9% ownership, structured in two tranches—half now, half released on hitting $150k MRR with disclosed net revenue retention above 110% and at least two enterprise anchor contracts. Reserve pro-rata for follow-on. Cap total exposure at $3M and hold to roughly 2% of the portfolio.

Narrative engine: live model (anthropic)

Entry strategy

Ticket: $2,239,500 target (range $1,119,750–$3,000,000)
Target ownership: 10%
Valuation band (pre-money): $8,372,000 / $16,395,000 / $32,790,000
Return: 6.87x expected (16.4x base) · ~31.7% IRR over 7yr · loss prob 59%

Deployment schedule:

- 40% — Entry: On close, after founder + IP + cap-table diligence.
- 35% — Milestone: Product-market fit signal (retention cohort / first repeatable revenue).
- 25% — Pro-rata: Reserve for next priced round to defend ownership.

Portfolio: Size at ~2.1% of a diversified venture portfolio (fractional-Kelly, conviction-scaled). Reserve 3,359,250 USD for pro-rata follow-on.

Score breakdown

Market size & growth — 61/100 (weight 20%)
~$158B TAM, 24% CAGR (AI Infrastructure / Tooling).

Timing / tailwinds — 80/100 (weight 10%)
Sector growth 24% vs. 12% neutral baseline.

Moat / defensibility — 78/100 (weight 15%)
Dominant defensibility here: data scale.

Unit economics potential — 55/100 (weight 15%)
~65% mature gross margin, capital intensity 70%.

Team / execution signal — 68/100 (weight 12%)
revenue/customers cited

Scientific / tech feasibility — 69/100 (weight 10%)
inference cost curves, sparse/MoE architectures, eval + safety tooling

Regulatory / legal headroom — 77/100 (weight 9%)
Regulatory intensity 35% (higher = more legal drag).

Competitive headroom — 37/100 (weight 9%)
Competitive intensity 90%. foundation-model commoditization compressing the middle layer.

Analyst council

Research Scientist — Nova's cross-cloud spot inference router is technically sound but sits in a commoditizing middle layer with fragile moat

+ Core approach is well-grounded engineering, not novel science: latency-SLA-aware placement over multi-cloud spot GPUs is a constrained optimization/bin-packing + predictive latency modeling problem with credible prior art (Kubernetes scheduling, spot arbitrage, vLLM/Ray Serve). No breakthrough required — execution risk, not feasibility risk.

+ 41% cost reduction and 99.95% uptime over 90 days are plausible and consistent with real spot-vs-on-demand price gaps (often 60-70% raw, netting to ~40% after failover overhead and SLA padding). Numbers are credible, not hand-wavy.

+ Claimed 'data scale' moat is thin: the routing telemetry (per-model latency, spot preemption patterns, price feeds) is genuinely accumulable and improves placement decisions, but the dataset is small ($48k MRR = ~months of traffic) and providers publish much of the underlying price/capacity signal directly.

+ Real de-risking milestone: demonstrate the router adds value NOT reducible to raw price arbitrage — e.g., predictive preemption avoidance, sub-100ms reroute on failover, or KV-cache-aware placement that competitors can't trivially copy. That converts a spreadsheet feature into defensible IP.

! Structural commoditization: hyperscalers and inference-serving incumbents (Together, Fireworks, Baseten, plus AWS/GCP native routing) can absorb cross-cloud routing as a feature; competitive intensity scored 90/100. A router that only arbitrages price is a margin-thin passthrough with 65% claimed gross margin likely eroding.

! Spot-capacity fragility at scale: multi-cloud spot availability for scarce high-end GPUs (H100/H200) is thin and correlated — preemptions cluster during demand spikes, precisely when SLAs matter most. 99.95% over 90 days at 22 small teams may not hold at 100x volume when Nova's own demand moves spot prices against it.

! Customer concentration and switching: 22 teams / $48k MRR (~$2.2k avg) implies SMB/experimental buyers who churn as they standardize on a single provider or bring routing in-house; no evidence of enterprise anchor contracts or net revenue retention.

Data Analyst — Nova Compute: real early traction on inference cost arbitrage, but thin moat in a commoditizing middle layer

+ Traction is credible but tiny: $48k MRR ($576k ARR) across 22 teams = ~$2.2k ACV, sub-seed scale; 41% cost reduction and 99.95% uptime are strong proof points but need to hold at 10x volume and larger enterprise SLAs.

+ TAM framing is inflated: $158B is total AI infra; Nova's actual SAM is the routing/optimization take-rate on multi-cloud inference spend — realistically low-single-digit % of serving budgets, a far smaller SOM (~$2-5B) not disclosed.

+ Unit economics ambiguous: 65% mature gross margin conflicts with 70% capital intensity — if Nova touches GPU capacity or passes through spot compute, margins compress toward 20-40%; pure software router margins would be 75%+. Need clarity on billing model (take-rate vs. resale).

+ CAC/LTV, payback, net revenue retention, and gross margin per request are ALL missing — the core metrics that confirm/kill the thesis. At $2.2k ACV, CAC must be <$5k for viable payback.

! Structural: foundation-model providers and clouds (Bedrock, Vertex, OpenRouter, Together) can bundle routing/failover natively, compressing Nova's middle layer to zero take-rate; competitive intensity scored 90/100.

! Moat claim ('data scale') is weak at 22 teams — routing telemetry advantages don't compound until thousands of workloads; spot-arbitrage logic is replicable and not defensible IP.

! Concentration/churn risk unquantified: 22 teams means a few logos likely drive most MRR; single-customer loss could swing ARR 10-20%, and NRR is undisclosed.

Economist — Nova arbitrages GPU spot markets for 41% inference savings — real value, but the middle layer is being squeezed from both sides.

+ Demand is highly elastic to price: inference cost is a top-3 line item for AI-native teams, so a verified 41% reduction is a strong wedge. $48k MRR across 22 teams (~$2.2k ARPA/mo) with 99.95% uptime shows the product survives contact with production SLAs — the hard part in routing.

+ Economic rent lives in the arbitrage spread across 5 clouds' spot markets plus SLA-aware placement data — a two-sided information moat that compounds with request volume (the 78/100 data-scale signal). But this spread is inherently mean-reverting: as spot pricing efficiency improves and clouds add their own routing, the arbitrage narrows structurally.

+ Competitive equilibrium is the core problem (37/100 headroom, 90% intensity): hyperscalers (AWS Bedrock, Azure) and inference platforms (Together, Fireworks, OpenRouter) all want this layer, and each cloud is incentivized to keep spot arbitrage inside its own walls. Nova's neutrality is its differentiation but also its structural fragility.

+ Unit economics are middling: 65% mature GM with 70% capital intensity suggests Nova either passes through GPU cost (thin margin) or takes inventory risk on committed capacity. A pure software router should hit 80%+ GM — the 65% figure hints they may be absorbing capacity risk, which caps operating leverage.

! Disintermediation by supply side: foundation-model providers and hyperscalers commoditize the middle layer (the flagged structural risk). If Bedrock/Azure ship native cross-pool routing, Nova's arbitrage evaporates and it becomes a feature, not a company.

! Margin compression from macro + hardware: falling inference cost curves (MoE, cheaper GPUs, Blackwell supply) shrink the absolute dollar savings Nova can capture — the 41% reduction becomes 41% of a smaller number, weakening the value prop as the market matures.

! Multi-cloud dependency risk: Nova's model depends on 5 clouds tolerating an intermediary that optimizes them

into a price war. Any single provider revoking spot API access or tightening terms damages the routing graph and reliability guarantees.

Corporate & Regulatory Lawyer — Nova Compute: light-touch US regulatory posture, but multi-cloud spot ToS and data-flow exposure are the real legal surface

+ Licensing/compliance: no sector license required for a routing layer, but Nova acts as a data processor under CCPA/CPRA and (for EU customer traffic) GDPR — needs DPAs, SCCs for cross-border inference, and clarity on whether prompts/completions transit or are logged; SOC 2 Type II is effectively table-stakes for the 22 enterprise-ish teams and a common Series A diligence gate.

+ Cloud dependency risk is contractual, not just technical: reselling/arbitraging spot capacity across five clouds (AWS/GCP/Azure/etc.) must be checked against each provider's ToS and acceptable-use/reseller terms — some prohibit resale or capacity brokering, creating termination-for-convenience and unilateral price/eviction exposure that Nova cannot control.

+ IP posture: core value is routing/placement heuristics and the cost/latency dataset — likely trade-secret-protected (weak patent story given prior art in load balancing); confirm all founder/contractor IP assignments executed and no open-source license contamination (AGPL/GPL) in the router that would compromise proprietary claims.

+ Deal structure for a $6M seed: standard priced round or capped SAFE with pro-rata, a MFN, information rights, and IP reps & warranties; add specific reps on cloud-ToS compliance and data-processing lawfulness, plus a founder vesting reset (4yr/1yr cliff) — regulatory headroom scored 77/100 so terms, not regulation, are where investor protection is earned.

! Foundation-model providers and the hyperscalers themselves can embed cheapest-pool routing natively (commoditizing the middle layer, competitive intensity 90/100) — and a cloud can simply revoke spot-reseller access via ToS, collapsing the business overnight with no legal recourse.

! Data/privacy liability if prompts/completions containing PII or regulated data (health/financial) are logged or routed cross-border without proper DPAs/SCCs — GDPR exposure up to 4% of global turnover and CCPA statutory damages, magnified because Nova touches every customer request.

! IP defensibility is thin: routing logic is hard to patent and easy to reverse-engineer; the only durable moat is the proprietary cost/latency dataset, which erodes if a larger player replicates it at scale — the 'data scale' moat claim (78/100) is fragile at $48k MRR.

Market data sources

- BCC Research (2025) — AI infrastructure $158.3B in 2025 !' $418.8B by 2030 at 21.5% CAGR
  https://www.bccresearch.com/market-research/artificial-intelligence-technology/ai-infrastructure-market.html
- Grand View Research (2025) — To $223.4B by 2030 at 30.4% CAGR
  https://www.grandviewresearch.com/industry-analysis/ai-infrastructure-market-report

Assumptions & limitations

- Market size / growth for AI Infrastructure / Tooling is anchored to BCC Research (2025): AI infrastructure $158.3B in 2025 !' $418.8B by 2030 at 21.5% CAGR. Full citations are listed under "Market data sources".
- Stage norms reflect US-market seed deals; adjust for geography "US".
- Score is a screening signal, not a substitute for legal, financial, and technical due diligence.