Akamai x NVIDIA
The AI Grid Deep Dive

Six layers of NVIDIA's reference design, Akamai's first global-scale implementation, the NOC moat, the Anthropic economics, and the honest caveats.

DalalBytes Research · September 26, 2026 · Reference price $113.94 (September 25, 2026 close) · NASDAQ: AKAM

DALALBYTES TECHNICAL VERDICT

The AI Grid is a genuine full-stack architecture, not a marketing slide.

Akamai is the first and furthest along of NVIDIA's six launch partners. The durable edge is the orchestration software plus 25 years of NOC muscle. The open questions are deployment scale and customer diversification beyond Anthropic.

The story in 90 seconds

On March 16, 2026, Akamai announced the first global-scale implementation of NVIDIA's AI Grid reference design: thousands of Blackwell GPUs woven into 4,400+ edge locations, governed by an intelligent orchestrator that routes each AI request to the right tier of compute by cost, latency, and throughput. Six months later, Anthropic committed $11.6 billion over seven years (expandable by $9 billion) for Akamai's distributed cloud. The stock spiked from a $110.41 close to a $129.16 intraday peak, then faded to $113.94 as the market weighed $5.5 billion of capex arriving two years before the revenue. This report goes deep on the architecture: what NVIDIA designed, what Akamai built, what the numbers say, and what remains unproven.

Timeline: from CDN to AI Grid

Date	Milestone
1998	Akamai founded out of MIT; the original distributed edge network is born.
October 28, 2025	**Inference Cloud launches**: 20 initial GPU regions across ~4,200 points of presence.
November 2025	Early traction reported in video, gaming, and retail; partners cite frame-accurate processing near viewers.
March 3, 2026	Procurement of **thousands of NVIDIA Blackwell GPUs** disclosed for the distributed cloud.
March 16, 2026	**AI Grid intelligent orchestration launches**, the first global-scale implementation of the NVIDIA AI Grid reference design, plus a **$200 million, four-year** metro-edge GPU agreement (customer unnamed).
~August 6, 2026	Q2 earnings: **$600M+, four-year robotics infrastructure deal** with an unnamed U.S. technology company; all GPU capacity sold out; $2.8B in 2026 CIS bookings to date.
September 24, 2026	**Anthropic commits $11.6B over seven years** (+$9B expansion option) for distributed cloud capacity, explicitly for CPU-heavy workloads.

Part I: NVIDIA's blueprint, six layers deep

NVIDIA's AI Grid reference design is a full hardware-and-software stack for turning distributed sites into one orchestrated inference platform. The six building blocks, from NVIDIA's own documentation:

Layer	What it is
1. GPUs	Rack-scale GB300 NVL72 systems for centralized AI factories; **RTX PRO 6000 Blackwell Server Edition** GPUs for distributed grid sites, designed to fit existing footprints with minimal retrofit.
2. Spectrum-X Ethernet	RDMA over Converged Ethernet (RoCE) with adaptive routing and congestion control; NVIDIA claims it accelerates storage performance by nearly 50% and cuts communication bottlenecks.

Layer	What it is
3. BlueField DPUs	Data processing units that offload, accelerate, and isolate networking, security, and infrastructure services in hardware, letting multiple tenants share common infrastructure securely.
4. TensorRT-LLM	Open-source library for high-performance real-time LLM inference on NVIDIA GPUs: modular Python runtime, PyTorch-native authoring, tuned for throughput and cost.
5. Dynamo	NVIDIA's distributed inference-serving framework. This is the deep one: it **disaggregates prefill from decode**, optimizes routing across nodes, and tiers KV-cache to cost-effective storage. Genuinely hard distributed-systems work.
6. NIM microservices	Containerized, pre-optimized inference microservices for deploying foundation models at enterprise scale.

The design goal, in NVIDIA's words: run every AI workload in its optimal location, with predictable latency, better token economics, higher utilization across pooled sites, and concurrency at scale. If a site fails, workloads rebalance automatically.

Part II: What Akamai built on top

NVIDIA supplies the blueprint and the silicon. Akamai's implementation adds the layer only a 25-year edge operator could build:

The intelligent orchestrator. A workload-aware control plane that acts as a real-time broker for AI requests, routing each one across edge, regional, and core tiers. It optimizes what Akamai calls **"tokenomics"**: cost per token, time-to-first-token, and throughput. Three techniques do the work:

Semantic caching. Instead of caching exact prompt matches (which almost never repeat in personalized AI), the orchestrator caches by *meaning* using embedding similarity: a new question close enough to an answered one gets served from cache without touching the model. This is the direct answer to the objection that AI output is uncacheable. **Flag: Akamai discloses no hit rates.**

Model affinity. Requests route to nodes where the right model (including fine-tuned or sparsified variants for the long tail) is already loaded and warm, avoiding cold-start latency.

Intelligent routing to right-sized compute. Premium GPU cycles are reserved for workloads that demand them; everything else drops to the cheapest tier that meets the SLA. Underpinning the economics: Akamai Cloud's open-source infrastructure and generous egress allowances, a direct attack on hyperscaler egress fees for data-intensive AI workloads.

The continuum of compute. The edge (4,400+ locations) handles rapid response for physical AI and agents via serverless Akamai Functions (WebAssembly) and EdgeWorkers. Akamai Cloud IaaS and dedicated multi-thousand-GPU clusters handle heavy post-training and multi-modal inference. One platform, three tiers, one SLA.

> **Claim check:** Akamai cites internal testing of up to 2.5x latency reduction and up to 86% inference cost savings versus hyperscalers. **No independent third-party benchmarks have validated these numbers.** Treat them as vendor claims until proven.

Part III: The hardware on the ground

What is disclosed:

Thousands of RTX PRO 6000 Blackwell Server Edition GPUs	**20** initial GPU regions	**41** datacenters in the AI platform expansion	**4,400+** edge locations in the routing fabric	**100%** of GPU capacity sold out (Q2 2026)

Per industry analysis of the deployment: cluster configurations run up to 8x RTX PRO 6000 GPUs with 128 vCPUs, NVMe storage, and BlueField-3 DPUs for networking offload, under managed Kubernetes with vLLM and KServe for model serving and NVIDIA NIM microservices. For the Anthropic build, Akamai signed hardware supply and procurement agreements with Lenovo and Jabil to pre-purchase memory and server components.

> **Not disclosed:** the exact GPU count (only "thousands"), and how many of the 4,400+ locations host inference GPUs versus routing-only. The footprint is real but concentrated; "global scale" today means global routing with regional GPU density.

Part IV: The NOC, the unparalleled strength

GPUs get the headlines. The Network Operations Command Center is what makes the grid operable, and it may be the most underpriced asset in the entire thesis.

200,000+ servers under management	**One-third** of global web traffic touched	**788 TB** of first-party data analyzed daily	**9 PB** threat-intelligence database	**900 GB** peak DDoS attack absorbed	**~10** people per shift, hundreds of alerts/hour

That last ratio is the whole story. It is not a room full of people watching screens; it is 25 years of automation density. The Cambridge NOCC is the mother ship among several command centers worldwide, running 24/7, and it has steered planetary-scale traffic through every record-breaking event and every record-breaking attack for a quarter century.

Why it matters for the grid: distributed inference with strict SLAs is an *operations* problem before it is a hardware problem. The orchestrator routes the request; the NOC keeps the grid alive when sites fail, attacks land, and demand spikes tenfold in a minute. Nobody buys that capability. You accumulate it.

Why telco competitors cannot copy it quickly: AT&T, Comcast, and Spectrum all run NOCs, but theirs are built to keep *connectivity* up. Akamai's is built to run a *multi-tenant distributed compute platform* with per-request

routing SLAs across thousands of networks it does not own. Buying RTX 6000s is a purchase order; operating them as one coherent system under SLA is institutional muscle.

There is also a flywheel the others do not get: seeing a third of web traffic means seeing every new attack and anomaly first, which feeds the threat models, which makes the security products better, which wins more traffic. The NOC is not just operations. It is the sensor network for the $2.24 billion security business.

Part V: The competitive field

Akamai was first, but it is a six-player launch cohort, not a solo act. At GTC 2026, NVIDIA unveiled the AI Grid concept with six initial network operators, each building to its own footprint:

Player	Grid status
Akamai	First global-scale implementation; live Inference Cloud product; $200M four-year metro-edge GPU agreement; $600M robotics deal; $11.6B Anthropic commitment. Furthest along.
Comcast	Validated that edge inference beats centralized on cost and throughput during demand spikes; now in field trial (personalized ad agents, local small-language-model concierge, gaming latency).
Spectrum (Charter)	1,000+ edge data centers and hundreds of megawatts within 10ms of 500 million devices; initial focus on GPU rendering for media production.
Indosat Ooredoo Hutchison	Sovereign AI grid across Indonesia; running the Bahasa Indonesia Sahabat-AI platform inside national borders.
T-Mobile	Exploring with NVIDIA; developers piloting smart-city, industrial, and retail workloads at cell sites.
AT&T	Signed on with Cisco partnership; enterprise and IoT inference focus.

Around them, an ecosystem is forming: **Cisco** (full-stack partner), **HPE** (infrastructure), and **Armada, Rafay, and Spectro Cloud** building control-plane software for grids. If the control plane gets productized well by third parties, Akamai's orchestration advantage narrows; that is the competitive risk to monitor.

Cloudflare is the parallel-track competitor. Workers AI runs serverless inference on NVIDIA H100 GPUs across 200+ GPU cities, billed in "Neurons" ($0.011 per 1,000), with 50+ open-source models and small models (up to ~8B parameters) kept permanently warm at the edge with ~45ms time-to-first-token reported from Tokyo. The positioning differs: Cloudflare serves developers with serverless inference; Akamai serves enterprises with SLA-bound distributed inference plus a $2.24B security portfolio. Both can win; they are playing different games on the same field.

> **Remember:** NVIDIA is Switzerland. It sells the reference design, the GPUs, and the software stack to everyone, including Akamai's competitors. The blueprint is not the moat; the implementation and the operations are.

Part VI: The money

The Anthropic deal. $11.6 billion over seven years for distributed cloud capacity, explicitly for CPU-heavy workloads, with an option to expand by up to $9 billion (potential ~$20B total). It is the second act of a relationship: a May 2026 master services agreement worth $1.8 billion, expanded ~7x by two new project plans dated September 18. Anthropic received a warrant for up to 7.7 million shares (~5% of outstanding) at $111.33; about 2% vested at signing, the rest vests per additional $3 billion of Anthropic spend. Attached capex: ~$5.5 billion, with the 2026 portion reported as $1.6 billion by some outlets and $1.7 billion by others. **Flag: discrepancy unresolved.**

What analysts say the deal is worth. Guggenheim (via Barron's): annual revenue around 40% of Akamai's $4.2B total and more than 5x 2025 CIS revenue of $314M. Piper Sandler: implies ~$1.66B in annual recurring revenue at full run rate, expected by Q4 2028, and sees CIS surpassing Security by late 2028. JPMorgan: Anthropic is now ~93% of the $14.4B in new deals Akamai signed in 2026, and wants diversified customer drivers.

The bookings cadence. $200M (Blackwell cluster, March) to $1.8B (frontier provider, May) to $600M+ (robotics, August) to $11.6B (Anthropic, September). Before Anthropic, 2026 CIS bookings already exceeded $2.8 billion.

Price action. September 24 close: $110.41 (-6.78% on the day). After-hours: ~+20% on the announcement. September 25: opened $125.23-125.41 (+12.9% gap), peaked at $129.16 (+17%) at 9:18am ET, faded to a **$113.94 close (+3.20%)**. Volume: 31.6M shares, ~5.6x the average. The fade is the market's verdict on timing: $5.5B of spending arrives two years before any revenue. Management reiterated full-year guidance; the deal adds cost now, revenue later.

Q2 2026 snapshot (reported ~August 6). Revenue $1.10B (+5% YoY); Security $604M (+10%); CIS $99.3M (+39%, a slight miss vs ~$103.8M expected); Delivery $396M (-6%). Non-GAAP EPS $1.59 (-8%). Operating cash flow $326M. Cash and securities $4.616B; debt $7.93B including $3.5B in converts raised to fund the buildout. FY2026 guidance: revenue $4.445-4.530B, non-GAAP EPS $6.40-7.05, CIS growing at least 50% in constant currency and accelerating in 2027. All GPU capacity sold out.

Analyst targets after the deal (all September 25, 2026):

Firm	Rating	Target
Guggenheim	Buy	**$225** (raised from $190)
Evercore ISI	Outperform	**$175** (maintained; called the deal "landmark")
Oppenheimer	Outperform	**$180** (maintained)
BofA	Buy	**$185** (raised from $175)
JPMorgan	Neutral	**$167** (raised from $158; wants diversification)
Piper Sandler	Overweight	**$158** (raised from $125)

Firm	Rating	Target
TD Cowen	Hold	**$149** (raised from $140)
UBS	Neutral	**$148** (reiterated)
RBC	Sector Perform	**$135** (reiterated)

Consensus: Hold, ~$147.71 (13 Buy / 9 Hold / 2 Sell). The street is impressed but split: the bulls see a value asset turning into a hypergrowth one; the skeptics see concentration risk and margin pressure.

Part VII: Honest caveats

A deep report earns trust by saying what is not proven. Here is the full list:

1. No independent benchmarks. The 2.5x latency reduction and 86% cost savings are Akamai's internal testing against hyperscalers. No third party has validated them.

2. GPU count undisclosed. "Thousands" of Blackwell GPUs is all we get. No exact count, no deployment timeline by site.

3. GPU density undisclosed. How many of the 4,400+ locations host inference GPUs versus routing-only is not public. Earlier descriptions referenced an initial rollout across ~20 sites.

4. No semantic caching metrics. Semantic caching is disclosed as a capability; hit rates, cost-savings attribution, and benchmarks specific to it are not. (Industry context: embedding-keyed semantic caches can reach 60-70% hit rates on natural-language LLM traffic where exact-match caching yields near zero, but that is not Akamai's number.)

5. Unnamed customers. The $200M metro-edge agreement and the $600M robotics deal both have unnamed counterparties. The $11.6B deal is the first named at scale.

6. Concentration. Anthropic is ~93% of the $14.4B in new deals signed in 2026 (JPMorgan). One customer, SLA-gated payments, termination on material outage.

7. Cost before revenue. ~$5.5B of capex lands before the revenue, which ramps in 2027-2028. The September 25 fade from $129 to $114 is the market pricing exactly this.

8. NVIDIA sells to everyone. The reference design, the GPUs, and Dynamo are available to all six launch partners and the ecosystem. Akamai's edge must come from implementation and operations, not the blueprint.

What to watch

The numbers that decide whether this thesis compounds:

Quarterly CIS revenue ex-Anthropic. Is the engine real beyond one customer? Piper Sandler expects CIS to surpass Security by late 2028; the quarterly prints will show whether that trajectory holds.
A second frontier-scale customer. JPMorgan's diversification point is the single most important catalyst.

One more $1B+ name changes the concentration math.

Deployment disclosures. Any hard GPU counts, site counts, or utilization figures in earnings calls or the Q3 10-Q agreement texts.

Independent benchmarks. The first credible third-party test of the latency and cost claims.

Capex as a share of revenue. Is the build converting to contracted cash flow, or is the fortress being spent without the revenue following?

2027 delivery. Management's line is growth accelerating from single digits to low-teens in 2027 on contracted commitments. That is the prove-it year.

Bottom line

The architecture is genuinely deep: six layers from NVIDIA, a workload-aware orchestrator with semantic caching and model affinity from Akamai, thousands of Blackwell GPUs, and a 25-year NOC that no competitor can replicate on a venture timeline. Akamai earned first-mover status among six launch partners and is the furthest along, with a live product, sold-out GPU capacity, and the largest disclosed grid contract on the board.

The investment question is not whether the technology works. It is execution and diversification: delivering $5.5B of capex on schedule, converting contracted revenue to cash in 2027-2028, and signing a second frontier-scale customer so that Anthropic is the first proof rather than the only proof. The market's fade from $129 to $114 says the skepticism is about timing, not direction. That is usually the right skepticism to have, and the right one to be proven wrong about.

Sources

Akamai press release, "Akamai Launches AI Grid" (March 16, 2026): akamai.com/newsroom

NVIDIA AI Grid for Telecommunications and building blocks: nvidia.com/en-in/industries/telecommunications/ai-grid

RCR Wireless, "Nvidia and global telcos are building AI grids" (GTC 2026): rcrwireless.com/20260319/ai/nvidia-telco-ai-grid

TelecomTV, "NVIDIA, telecom leaders build AI grids" (GTC 2026): telecomtv.com

Barron's on the Anthropic deal and Guggenheim math (Sept 25, 2026): barrons.com

Investopedia on the Anthropic deal and warrants (Sept 25, 2026): investopedia.com

Piper Sandler via Stocktwits (Sept 25, 2026): stocktwits.com

Akamai Q2 2026 results via StockTitan: stocktitan.net

Akamai AI Grid launch details via StockTitan: stocktitan.net

HostingJournalist on Blackwell deployment: hostingjournalist.com

AICerts on the inference grid architecture: aicerts.ai

Nasdaq press release, "Akamai Sharpens Its AI Edge" (March 27, 2025): nasdaq.com

Akamai NOCC (AVNetwork): avnetwork.com · Haivision case study: haivision.com

Akamai scale and threat intelligence (solution brief): akamai.com/resources

LogicMonitor case study on Akamai monitoring: logicmonitor.com

Cloudflare Workers AI practitioner review: adamarant.com

InPlay Finance on September 25 price action: inplay.finance

Insider Monkey on the cost-before-revenue fade: insidermonkey.com

MarketsandMarkets AI inference sizing; Futurum edge/hybrid growth estimate (via search).

Research for informational purposes only, not investment advice. Investing involves risk. This analysis discusses the ticker only and contains no portfolio or position information.