MAVKA CAPITAL

Data Center Technology: AI Inference Infrastructure Deep Dive

Frontier Infrastructure & Compute Markets

2026 Outlook

Team

Vitaly Golomb

Managing Partner

Vitaly Golomb is a seasoned technology entrepreneur, investor, and investment banker with over 20 years of experience in Silicon Valley. As a Partner at Drake Star, he led the Mobility and Climate Tech practice, advising companies like Rimac Automobili, Fisker, and Taiga Motors. Previously, Vitaly was a Founding Partner at HP Tech Ventures and CEO of several startups.

vitaly@mavkacap.com

+1 415.683.6865

Misha Edel

Partner

Misha Edel is a seasoned executive and deal advisor with over two decades of experience in M&A, value creation, and technology-enabled transformation. He has advised on hundreds of transactions across sectors and helped scale numerous early-stage companies at the intersection of software, analytics, and emerging technologies.

medel@mavkacap.com

+1 510.282.9758

About Mavka Capital

At the intersection of strategy, finance, and marketing, Mavka Capital offers a unique approach to business transformation. Our integrated services combine hands-on leadership with deep expertise, positioning companies for long-term success. We align strategic vision with market realities and investor expectations, guiding businesses through critical growth phases and ensuring they thrive before, during, and after significant transactions.

Executive Summary

The center of gravity in artificial intelligence has shifted from model training to **inference**—the act of delivering intelligence to users at scale.

Between 2024 and 2025, inference became the **economic engine** of the AI value chain, dictating power allocation, pricing, and control of access to intelligence.

Three structural shifts define this new phase:

- **Physics as the limit of progress.** Energy, heat, and latency—not data or algorithms—now bound the frontier.
- **Capital as the new gatekeeper.** BlackRock's $40 B acquisition of Aligned Data Centers (with Microsoft and NVIDIA as partners) marks the rise of "compute landlords."
- **Economics of milliseconds.** Tokens-per-second and Time-to-First-Token now determine margins and user experience.

Seventy-four percent of new capacity is pre-leased; Northern Virginia vacancy is below 1 percent. Rack power densities have surged from 40 → 130 → 250 kW, and average pricing has reached **$217 / kW / month**, the highest since 2011.

Inference workloads dominate data-center energy consumption, and venture capital continues to chase efficiency—into startups such as Modular AI, Rebellions,EdgeCortix, and d-Matrix.

Inference is no longer a computational step; it is a market-design problem where energy, latency, and capital intersect.

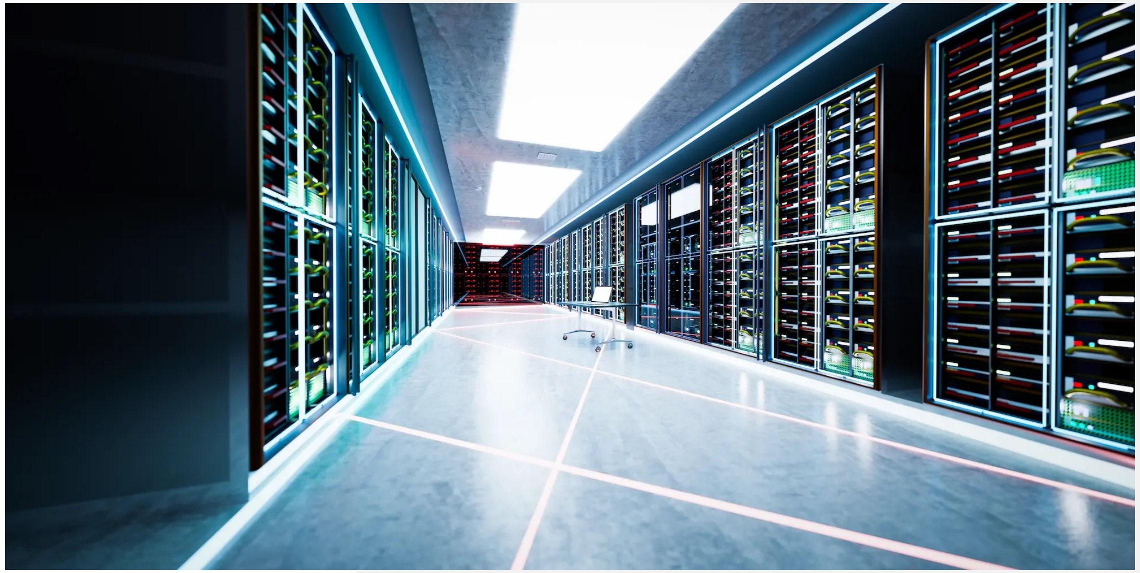
Source: CIO (FoundryCo, Oct 16 2025)

From Physics to Economics: The New Data-Center Reality

Compute scarcity has replaced data scarcity as the limiting factor of innovation.

Implications:

- **Regulatory and Grid Constraints.** Carbon-intensity reporting and regional megawatt ceilings now gate expansion.
- **Procurement Horizon.** Enterprises must forecast compute needs 3–5 years ahead, often locking in ROFO/ROFR clauses.
- **Geographic realignment.** Tier-2 markets—Phoenix, Dallas, Montréal, Santiago—absorb overflow.
- **Rise of Compute Landlords.** PE-backed platforms own capacity; hyperscalers lease long-term; enterprises rent residual supply.



Infrastructure has moved from IT periphery to **board-level strategy**; access to power is now a strategic asset.

Compute Metrics and the Economics of Inference

Every token generated consumes measurable electricity. At scale, the relationship between **latency, throughput, and power** determines the cost of intelligence..

Metric	Meaning	Why It Matters
Tokens per Second (TPS)	Throughput of generated tokens	Defines serving cost and concurrency
Time to First Token (TTFT)	Latency until model begins output	Determines interactivity and UX
Throughput per Watt	Tokens /sec per watt	Measures efficiency; key power-cost driver
Utilization Rate	GPU busy ratio	Affects marginal cost and ROI

Training scales with total compute hours; inference scales with **latency and concurrency**. The architectural race is about shortening TTFT while maximizing sustained TPS under strict power envelopes

Typical pricing ranges from **$0.15 → $15 per 1 M tokens**, depending on model class and context window.
At ~70 % compute utilization and $0.10 /kWh, **energy cost ≈ $0.01 per 1 M tokens**.

As architectures improve caching and parallel decoding, energy and cooling become dominant marginal costs.

Source: Mavka Analysis – Mavka_AI_Inference_Data_2025-10-19.xlsx

Architectural Strategies and the Compiler Wars

When physics limits performance, architecture and software become the differentiators.

NVIDIA and the CUDA Moat

Infrastructure has moved from IT periphery to **board-level strategy**; access to power is now a strategic asset.

1. Developer Inertia — Millions trained on CUDA / cuDNN; porting cost is high.

2. Library Density — Optimized kernels & inference servers (Triton, TensorRT).

3. Compiler Continuity — Backward compatibility across GPU generations.

Architectural Strategies and the Compiler Wars

Groq's Counter-Model

Groq reverses the paradigm: its deterministic single-cycle pipeline makes **the compiler the hardware**. This yields microsecond-level TTFT with minimal batching—ideal for chatbots and real-time inference—at the cost of flexibility and ecosystem depth.

NVIDIA = scale and software inertia
Groq = latency determinism and compiler elegance.

Strategic Access	Leading Examples	Core Strategy	Commentary
Hardware Throughput	NVIDIA H200, AMD MI325X	Dense tensor cores, HBM3e bandwidth	Dominates batch LLMs > 100 B params
Deterministic Latency	Groq LPU	Compiler-driven single-cycle pipeline	Excels at ultra-low TTFT; fixed workloads
Compiler Ecosystem	CUDA / TensorRT vs ROCm / XLA / Groq Compiler	Vertical integration	Software > Silicon for moat durability
Vertical Cloud Integration	AWS Inferentia, Azure ND H100, CoreWeave	Own stack + power procurement	Margins via managed endpoints
Edge Inference	Qualcomm AI Hub, Apple Neural Engine, EdgeCortix	Local compute, privacy, latency	Smaller models, huge install base
Middleware Abstraction	Modular AI, OctoML, Anyscale	Translate models across backends	Neutral "Switzerland" layer; M&A targets

Capital and Capacity: The Financialization of Compute

Ownership of physical compute now defines strategic advantage. BlackRock's $40 B Aligned Data Centers acquisition—with Microsoft and NVIDIA as AI infrastructure partners—illustrates how capital allocators are becoming **gatekeepers of intelligence**.

- **Compute as an Asset Class**. Data-center platforms are valued on forward megawatts, not square footage.
- **PE Dominance.** Private equity accounts for 80 – 90 % of data-center M&A since 2022.
- **Concentration.** By 2026, five fund consortia are projected to control > 40 % of North American AI capacity.
- **Vertical Integration.** Hyperscalers co-invest to secure supply and power contracts.

Implication:

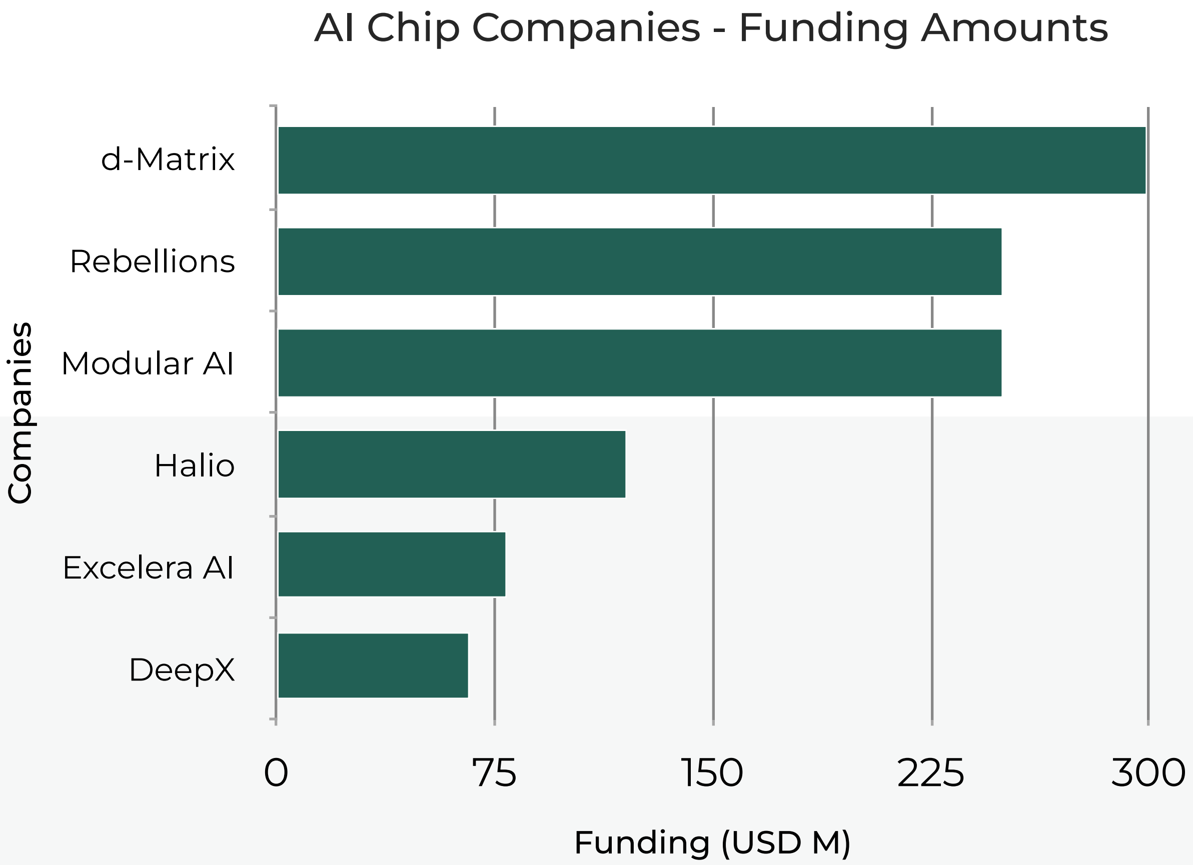
Control of megawatts = control of AI margins.

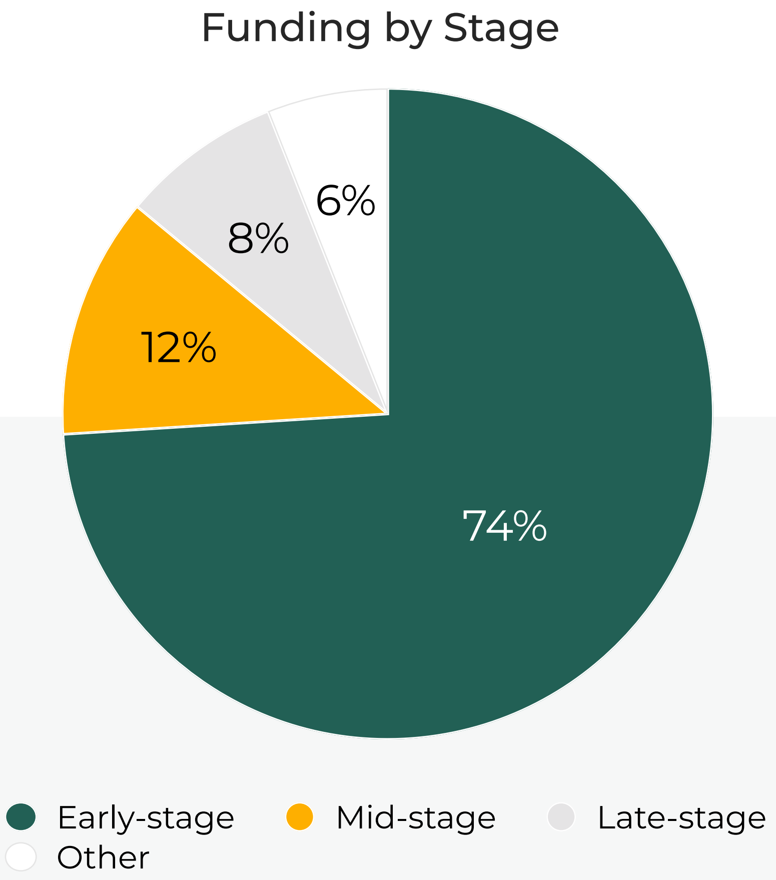
Source: CB Insights, State of Venture Q3 2025 (AI Sector Spotlights)

Enterprise Implications

- ✓ **Compute as Balance-Sheet Asset.** CFOs increasingly treat capacity reservations like energy hedges.
- ✓ **Forecast Horizon Extension.** AI budgets require 36–60 month visibility tied to power SLAs.
- ✓ **Operational Exposure.** Pre-commitments can create stranded costs if model architectures shift.
- ✓ **Efficiency Opportunity.** Modernizing idle workloads may reclaim 15–20 % of capacity.
- ✓ **Governance Risk.** “AI-ready” claims without electrical or cooling upgrades can trigger M&A diligence issues.

Investment and M&A Trends





Source: CB Insights, State of Venture Q3 2025 (AI Sector Spotlights)

Key Takeaway: Investments in early-stage startups may imply future opportunities for innovation

Strategic Outlook — Mavka Capital View

1. Compute Concentration → Pricing Power

Ownership consolidation gives capital allocators control over AI margins.

2. Software Moats > Silicon

Compiler and runtime control dictate defensibility more than transistor design.

3. Capital Stratification

Expect continued PE consortium acquisitions (GIP / NVIDIA / Microsoft alignments).

4. Emergence of Inference-as-a-Service

Mid-tier providers (CoreWeave, Lambda) will capture enterprises unable to pre-lease hyperscale capacity.

5. Thermal and Power Innovation

MidCooling, waste-heat reuse, and density optimization form the next investable frontier.

6. Edge Expansion

Privacy rules and latency needs drive inference toward device-level compute.

7. Regulatory Pressure

Grid allocation and carbon accounting may shape site economics more than demand curves.

8. Convergence Risk

“AI-ready” branding without true retrofit creates potential stranded assets in PE portfolios.

Mavka Thesis

The future of inference is a market for milliseconds.

Inference infrastructure will bifurcate into two ecosystems: hyperscale compute landlords and distributed inference networks.

Winners will price latency as a product and treat power as capital. Inference is becoming a strategic national asset that blends compute, energy, and finance.

Sources & References

Primary Industry and Market Data

- CIO (FoundryCo, Oct 16 2025) – "BlackRock's $40 B data center deal opens a new infrastructure battle for CIOs." ($217 /kW/mo; 17–18 % YoY increase; 1.6 % vacancy; 74 % pre-leased; 130 → 250 kW density)
- CB Insights – *State of Venture Q3 2025* (2,324 deals +8 % QoQ; 51 % AI funding share; d-Matrix $300 M C rep.; Modular AI $250 M C)
- CBRE – *Global Data Center Trends 2025* (cost, power density, vacancy metrics)
- JLL – Financing the Future: Trends in 2025 Data Centre Investment (power-density 130 → 250 kW; site-selection commentary)

Inference Cost and API Pricing

- OpenAI (2025) GPT-4o mini pricing update – $0.15 / $0.60 per 1 M tokens.
- Anthropic (2025) Claude Sonnet 4.5 – $3 / $15 per 1 M tokens.
- Google AI Studio (2025) Gemini 1.5 Pro – $0.30 / $2.50 per 1 M tokens.
- Cohere (2025) Command R & R+ – $0.15 / $0.60 and $2.50 / $10.00 per 1 M tokens.

Startup and Funding References

EdgeCortix (Aug 18 2025) Series B close (~$100 M); Axelera AI (Mar 6 2025) €61.6 M EU grant; Hailo (Apr 2 2024) $120 M Growth round; DeepX (Aug 9 2025) $79 M C + IPO prep; d-Matrix (Nov 19 2024) launch + $300 M C rep.

Supporting and Analytical Inputs

Everest Group & Greyhound Research via CIO 2025 (commentary on AI workload economics); Synergy Research (2025) Global DC M&A totals ($73 B in 2024 vs $26 B in 2023); Americans for Financial Reform (2025) PE Data Centers Report (ownership concentration).

Internal Workbook

Mavka_AI_Inference_Data_2025-10-19.xlsx
— Sheets: Inference Pricing, Data-Center Metrics, Rack Power Density, Funding by Stage, Startups.

Prepared by Mavka Capital — Frontier Infrastructure & Compute Markets, October 2025. Confidential analytical brief. Distribution restricted to clients and partners under NDA.