Everyone's Watching Training Move. Inference Can't Follow.

Second Order Labs™ · June 2026

ABSTRACT

xAI built a 100,000-GPU cluster in Memphis in 122 days because the city had power and could deliver it fast. That choice marks a permanent split in AI compute into two machines that obey different laws of physics.

Keywords: AI Infrastructure, Data Centers, Edge Inference, Power Purchase Agreements, Compute Arbitrage, Nuclear Power

"Because these models run on our edge network, inference happens close to your users, which reduces latency."

— Cloudflare, Workers AI

In 122 days, xAI stood up a 100,000 GPU training cluster in Memphis.[1] Memphis tells you where AI infrastructure is going. It is not a research hub. It is not close to Sand Hill Road. Nobody picked it for prestige. They picked it because power was available and the site could move fast. That choice matters more than most AI infrastructure coverage admits.

The general-purpose data center is fading fast. AI compute is splitting into two different physical systems: a power-bound training core that belongs near huge, cheap energy supplies, and a latency-bound inference edge that has to stay close to users. These halves obey different physics and will be financed and built separately. That boundary already shows up in siting decisions. Call it the **Watt-Latency Line** if you need a label. Most investor notes and trade coverage still focus on training cluster migration, not urban inference capacity. Coverage in the FT, Bloomberg, and Reuters keeps centering on Gulf builds and nuclear-backed training sites. Meanwhile inference, already roughly two-thirds of AI compute, is stuck near cities for a reason no engineer can code around.

I. Training Is Heavy Industry, And Heavy Industry Chases Power

Training a frontier model looks less like software and more like aluminum smelting. It is a continuous industrial process with a savage appetite for electricity. A smelter does not care about customer proximity. It cares about power price and power certainty. AI training now works the same way, which is why clusters are moving toward stranded or underused energy.

The bottleneck is no longer the chip. Epoch AI finds the power needed to train frontier models is now doubling every year, outrunning the grid's ability to deliver it, so power availability is overtaking raw compute as the binding constraint on frontier AI.[2] The gating factor is grid access, especially the interconnection queue.

That delay is the whole problem. In Northern Virginia, the densest data center market in the world, developers can wait up to five years for grid interconnection.[3] So they are looking past the public grid. Behind-the-meter generation, putting compute directly next to a power source, is becoming the obvious move. SemiAnalysis has made the same point: the interconnection queue is the new supply-chain bottleneck, and it pushes builders toward stranded power they can tap without waiting in line.[4]

Figure 1: Training behaves like heavy industry: it migrates to cheap power, not to customers.

Microsoft's 20-year power purchase agreement to restart Three Mile Island, now called the Crane Clean Energy Center, is the clearest signal yet.[5] A dormant nuclear plant does not come back on sentiment. It comes back because a hyperscaler signs a contract long enough to underwrite the restart. AI companies are no longer just buying servers. They are shaping the power sector's biggest capital decisions.

II. Energy as the New Silicon: PPAs Replace GPU Allocation as the Moat

For the last two years, the choke point was NVIDIA allocation. Whoever got the GPUs had the edge. That is changing. The scarcer asset now is a power purchase agreement and a place in the interconnection line. Energy is becoming the harder moat.

That shift redraws the map. Gulf sovereign funds are pouring money into AI infrastructure backed by electricity costs near five cents per kilowatt-hour and deep pools of capital.[6] Microsoft and G42 are building a $1 billion data center in Kenya that runs on geothermal power.[7] Those projects are doing the same thing: moving compute to cheap energy instead of moving energy to demand centers. Power that used to be trapped inside a local market can now be sold globally through compute exports. Financing follows the same logic. A gigawatt-class site can cost more than most venture funds can underwrite, so financing shifts to infrastructure funds and sovereign backers. Microsoft and BlackRock's $30 billion vehicle is aimed straight at data centers and the energy assets behind them.[8] Venture money will chase inference software. Sovereign money will pour concrete around training clusters.

Dimension	Power-Bound Training Core	Latency-Bound Inference Edge
Governing constraint	Gigawatt power availability	Speed-of-light latency floor
Optimal location	Remote energy hubs (Gulf, Kenya, Memphis)	Dense population centers
Competitive moat	PPAs, interconnection queues	Real estate proximity, last-mile networks
Capital source	Sovereign wealth, infrastructure funds	Venture capital, telecom
Industrial analogy	Smelter	Last-mile logistics

Think of it as manufacturing versus delivery. Intelligence gets made in power-rich interiors and served at the edge. The chips may share a vendor. The business does not.

III. Why Can't Inference Just Move to Cheap Power Too?

Because physics says no. Cloudflare, which serves inference from a global edge network, puts the logic plainly: because the models run close to users, latency drops.[9] A round trip from a user to the nearest edge node still runs in tens of milliseconds, and light in fiber will not cross continents inside a responsive budget. Training can sit 1,000 miles away. A live voice interaction cannot.

Most coverage of AI infrastructure spending still ignores the urban siting constraint on inference. Yet inference is already the majority of AI compute. Deloitte estimates it reached roughly two-thirds in 2026, up from half in 2025 and a third in 2023, and the share keeps climbing as models move into production.[10] NVIDIA has said inference now drives a large share of its data center revenue as models are deployed at scale.[11]

> *Intelligence is manufactured in remote, power-dense interiors and delivered at the hyper-local edge, and no facility can be both factory and storefront.*

At that point, inference starts looking less like cloud compute and more like urban property. A node near a major metro is valuable because it can serve millions of users inside the latency budget. A node in a cheap-power interior cannot. Training benefits from distance. Inference pays for proximity.

Figure 2: Inference is priced like urban real estate: proximity to people is the asset.

That pushes value toward telecom operators that control metro fiber, tower sites, and edge real estate. AWS already sells edge inference on the promise of lower "round-trip times and bandwidth costs."[12] The cell tower is turning into a compute asset, whether carriers fully grasp it yet or not.

IV. Bifurcated Silicon and the Decoupling of Talent From Iron

Once you accept the Watt-Latency Line, chip roadmaps split and talent no longer needs to sit near the machines. One branch of silicon is built for giant training clusters that can tolerate brutal power draw and heat. The other is built for efficient edge inference or on-device use. The dream of one general-purpose accelerator that does both well is fading.

Location splits too. The thermal density of a 100,000 GPU cluster pushes training sites toward places that can handle extreme cooling loads and massive power delivery. Meta says its next-generation AI data center design needs "entirely new approaches to cooling and power distribution."[13] Old industrial sites with heavy grid connections suddenly look valuable again. Former steel mills and aluminum smelters already have the wiring. That matters more than being near a startup district.

Meanwhile the old bond between engineers and servers is breaking. Silicon Valley grew up around keeping people close to machines. That made sense when the machines were in Santa Clara. It makes less sense when the cluster sits next to a geothermal field in Kenya or in a Gulf desert. Researchers can stay in San Francisco. The iron will go where the watts are.

Figure 3: A five-year grid wait is why training is fleeing traditional hubs for behind-the-meter power.

V. The Agentic Time-Bomb That Could Erase the Line

There is one scenario that could scramble this split. The Watt-Latency Line exists because today's high-value inference workloads are interactive and latency-sensitive. A human is waiting. That keeps the compute close. Remove the human and the rule changes.

Autonomous agents running overnight jobs do not care whether a response takes 5 milliseconds or 5 seconds. If the task finishes by morning, latency barely matters. When inference demand tilts toward latency-tolerant agentic work, part of inference moves back to cheap-power interiors. Batch inference starts acting like training. It hunts for watts.

That would hit current infrastructure bets from both sides. Anyone buying edge inference real estate is betting that inference stays interactive and urban. Anyone financing giant training sites is betting that the split holds. Agentic drift will pull some inference workloads back across the Watt-Latency Line if batch demand becomes material. The moment that shift shows up in usage data, a lot of metro-edge assumptions get repriced.

VI. What to Do With the Watt-Latency Line

Training investors should ask one question: how fast can this operator secure power without sitting in a five-year queue. GPU access still matters. Grid access matters more.

A metro inference site should be judged like scarce real estate, not like a generic data center. Distance to users, fiber density, and telecom partnerships will decide who wins. The best edge locations will look overpriced right up until they are not available.

The first real inference real-estate bubble will start when investors realize batch agents do not need to live near users. Watch one metric: the share of inference tokens generated by batch agents. When that rises, edge valuations break.

KEY FINDINGS

- xAI brought a 100,000-GPU training cluster online in Memphis in 122 days, picking the site for power rather than talent or network topology.
- Power availability now beats FLOPs as the frontier AI bottleneck, so Power Purchase Agreements have replaced GPU allocation as the competitive moat.
- Inference already eats roughly two-thirds of AI compute (Deloitte), and a latency floor measured in tens of milliseconds keeps it anchored to cities.
- Microsoft signed a 20-year power agreement to restart Three Mile Island, a contract long enough to justify reviving a dormant nuclear plant.
- Autonomous agents running latency-tolerant batch jobs could pull inference back toward cheap-power interiors and blur the line between training and inference.

REFERENCES

[1] Elon Musk, announcement on xAI Colossus, X. "This weekend, the @xAI team brought our Colossus 100k H100 training cluster online. From start to finish, it was done in 122 days." https://x.com/elonmusk/status/1830650370336473253

[2] Epoch AI, "The power required to train frontier AI models is doubling annually." https://epoch.ai/data-insights/power-usage-trend

[3] Financial Times, "US electricity grid struggles to keep up with AI boom."

[4] SemiAnalysis, "AI Datacenter Energy Dilemma: Race for AI Datacenter Space." https://newsletter.semianalysis.com/p/ai-datacenter-energy-dilemma-race

[5] Constellation Energy, "Constellation to Launch Crane Clean Energy Center." https://www.constellationenergy.com/news/2024/Constellation-to-Launch-Crane-Clean-Energy-Center-Restoring-Jobs-and-Carbon-Free-Power-to-The-Grid.html

[6] Reuters, "Middle East sovereign wealth funds target AI infrastructure."

[7] Bloomberg, "Microsoft, G42 to Build $1 Billion Geothermal Data Center in Kenya." https://www.bloomberg.com/news/articles/2024-05-22/microsoft-g42-announce-1-billion-geothermal-data-center-in-kenya

[8] Wall Street Journal, "Microsoft and BlackRock Form $30 Billion Fund to Invest in AI Infrastructure."

[9] Cloudflare, "Workers AI" (inference runs on the edge network, close to users, to reduce latency). https://blog.cloudflare.com/workers-ai/

[10] Deloitte, "2026 TMT Predictions: More compute for AI, not less" (inference roughly two-thirds of AI compute in 2026, up from half in 2025 and a third in 2023). https://www.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2026/compute-power-ai.html

[11] NVIDIA Q3 FY25 Earnings Call Transcript (inference share of data center revenue). https://s1.q4cdn.com/104539020/files/doc_financials/2025/q3/Transcript-Q3-FY-25-Earnings-Call.pdf

[12] AWS, "Deploying ultra-low latency AI inference at the edge." https://aws.amazon.com/blogs/machine-learning/reduce-conversational-ai-response-time-through-inference-at-the-edge-with-aws-local-zones/

[13] Meta Engineering, "Designing Meta's next-generation AI data center." https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/
