Power Bytes

Monitoring the Situation

If you are a heavy Claude user, you already know the message: "daily limit reached." Lately, this message seems to pop up faster and more often. Frustrated users dismiss this as product friction.

What they're actually bumping into is the ceiling of a constrained supply chain. The AI infrastructure stack, semiconductors, data centers, energy capacity, etc, is physically limited, and those limits ripple forward. By the time they reach the user, it takes the form of a rate limit. Model providers have responded by managing demand: capping access, pushing heavier usage to premium tiers, and optimizing the experience around scarcity. Some users have also noticed a quieter phenomenon, coined as "AI shrinkflation"- models providing less inference depth, less thinking time, or a more constrained version of the product than the headline capability suggests. These frictions are symptoms of a system running into hard upstream limits.

The temptation is to jump straight from AI demand to electricity demand. However, model economics are not that direct. Before demand reaches the grid, it passes through a sequence of bottlenecks, with each one shaping how much latent demand actually becomes grid-visible load.

Model Layer: Inference Demand Explosion

The compute crunch is best understood by starting at the layer closest to the prompt. Upstream bottlenecks shape the rate at which demand reaches the grid, but there has been a notable shift in the character of inference demand itself.

ANALYST CONTACT

Aishwarya Mahesh
Energy & Technology Research Analyst
(281) 217-7675
Aishwarya.Mahesh@admis.com
ADM Investor Services, Inc.

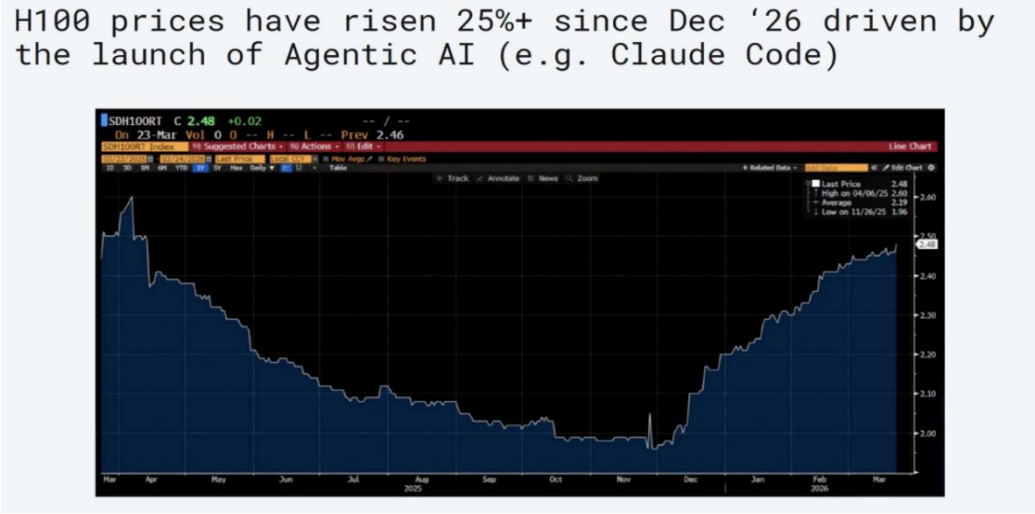
The rise of agentic AI, where models execute multi-step tasks autonomously rather than engaging in tight interaction loops with users, marks a fundamental shift in how tokens are consumed. Before agents, inference demand was relatively segmented. A small class of developers, researchers, and heavy technical users consumed a disproportionate share of frontier-model compute, while most users remained lighter or more occasional.

Agents changed this. A single interaction now triggers a cascade of hidden inference events, *planning, tool calls, iteration*, all behind an in-progress spinner accompanied by a quirky verb. The software running on a user's behalf consumes compute asynchronously, working in the background faster than the user can type. While the user waits for one job to finish, they can start 1, 2, even 10 more. The result is an increased inference intensity per interaction, effectively converting a much larger share of users into power users.

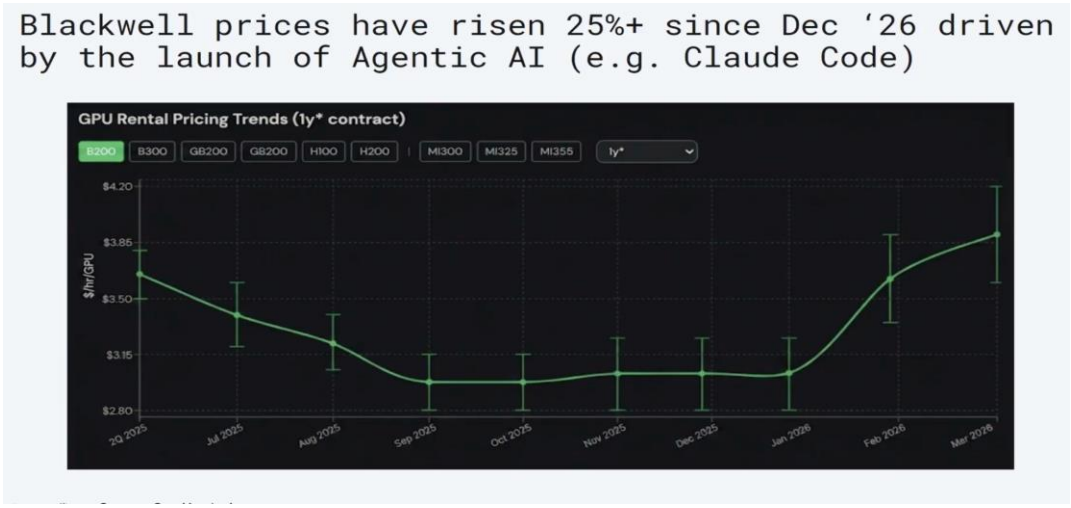
Aishwarya Mahesh
(212) 785-1555
Aishwarya.Mahesh@admis.com

Other than signing new infrastructure agreements to procure more compute, model providers are responding at the product level, where they have the most control. OpenAI explicitly caps usage by tier. Free users are throttled to a lighter model after 10 messages every five hours, while paid tiers carry higher but still hard limits. Google introduced Flex and Priority inference tiers in the Gemini API, separating latency-sensitive requests from those that can wait and pricing scarcity accordingly.

Anthropic's response is the most indicative. OpenClaw, an open-source agentic framework, became the fastest-growing GitHub project in history, reaching 100,000 stars in under 48 hours with its appeal being that users could run autonomous, compute-intensive workflows on a flat $20 subscription. On April 4, 2026, Anthropic told subscribers they could no longer apply their subscription limits to OpenClaw or any third-party harness, moving that usage to a separate pay-as-you-go billing system. Anthropic's head of Claude Code stated plainly that subscriptions "weren't built for the usage patterns of these third-party tools." The common thread across all three is rationing as frontier labs draw hard lines between casual conversational use and agentic workloads, and pricing the difference.



Source: Crusoe, Bloomberg



Source: Crusoe, SemiAnalysis

Aishwarya Mahesh
(212) 785-1555
Aishwarya.Mahesh@admis.com

This market signal is showing up in chip pricing. According to Epoch AI, the cost to run LLMs at a fixed level of performance has fallen dramatically, in some cases by 40x per year. Under normal conditions, efficiency gains at that scale should push GPU rental prices down. Instead, Blackwell rental rates bottomed around mid-2025 and have since risen over 25%, coinciding almost precisely with the acceleration of agentic adoption. The efficiency gains are being absorbed before they reach the price. Agents generate far more tokens per interaction than conversational use ever did, and each efficiency improvement creates capacity that agentic workloads immediately fill. The H100 and Blackwell charts are an important leading indicator, showing that new demand shows up in the market price of compute before it hits the power meter.

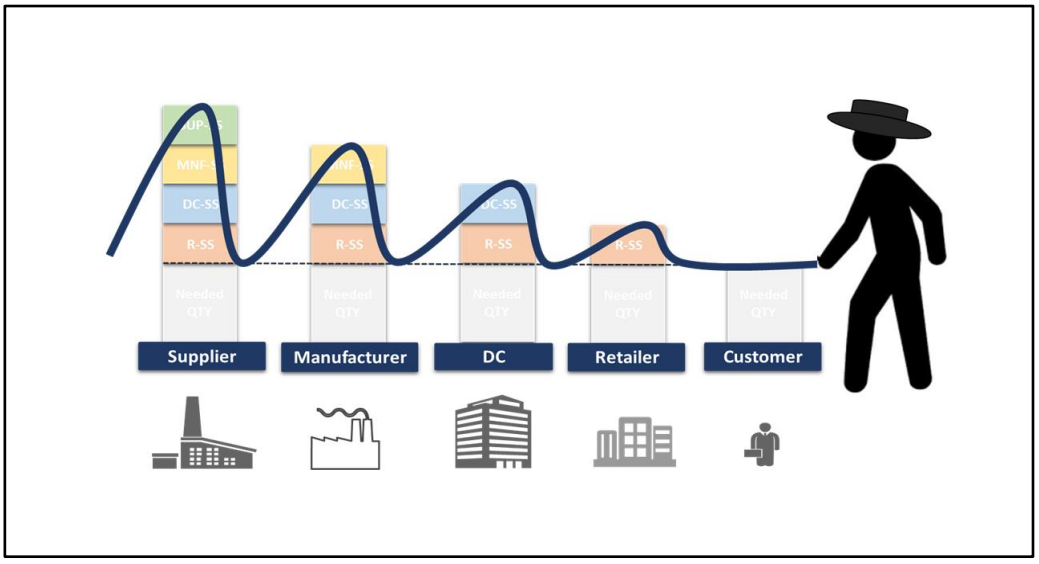
Model providers are not acting as though compute is an abundant on-demand utility. They are reserving supply and signing longer-duration infrastructure agreements. Anthropic's deal to rent SpaceX's Colossus 1 cluster, for instance, runs at $1.25 billion per month through 2029. Yet the same agreement lets either party walk with 90 days' notice, a hedge that reveals the labs' own uncertainty about future demand. They are committing real capital to secure scarce capacity now, while declining to bet the farm on what that capacity will be worth later. The demand is real enough to pull infrastructure behind it, and the caution is real enough to keep the exits open.

For power traders, GPU rental prices and product rationing are the leading indicators for datacenter related load growth. That pricing pressure reflects a surge in inference demand that lands across the entire AI supply chain. Furthermore, power is being contracted aggressively off the back of that demand signal, but whether contracted load actually gets drawn depends on whether the rest of the supply chain resolves in parallel. The components have to come together for power to convert to compute cleanly and that sequencing problem is where the story gets complicated.

The Bottleneck Stack

The Bullwhip Applied to AI Infrastructure

The inference demand surge propagates through every layer of the AI supply chain. In doing so, it triggers a dynamic that supply chain economists have studied for decades: **the bullwhip effect.**



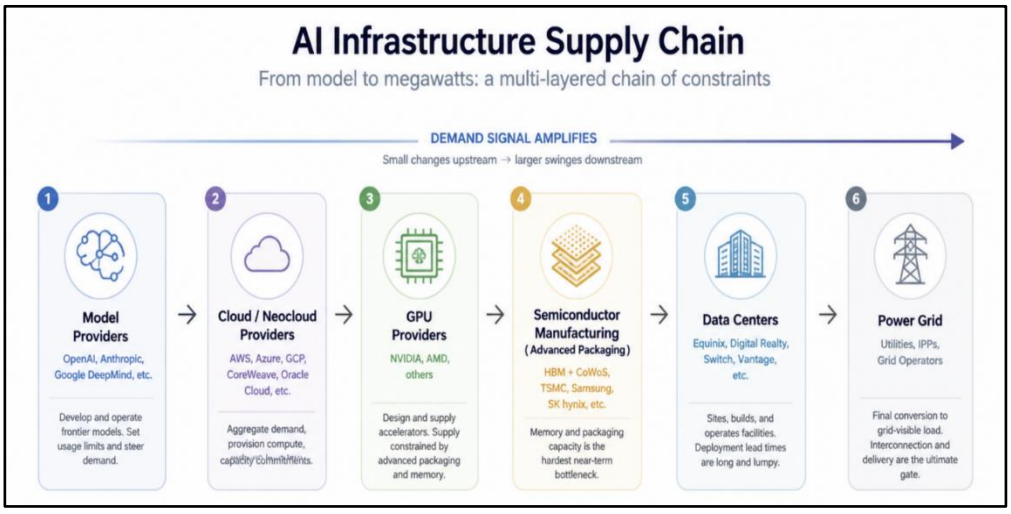
Aishwarya Mahesh
(212) 785-1555
Aishwarya.Mahesh@admis.com

The bullwhip effect describes how small demand shocks amplify as they move upstream through a supply chain. Long lead times, poor information sharing between layers, and the rational tendency of each participant to over-order against uncertainty cause orders to become far more volatile than the underlying demand that initiated them. Forecasted demand volatility from model providers compounds as it goes down the supply chain to the power providers, who shoulder relatively enormous amounts of risk for the same end-consumer behavior change.

Applied here, the framework works as follows. The shift to agentic AI as described in the previous section, created a demand signal that was structurally different from anything the supply chain had previously priced. Compute per interaction surged, pushing a far larger share of users into high-intensity usage. Every layer of the supply chain received that signal and responded. But each layer has a different lead time, different capital structure, and different information about what the layers above and below it are doing. The result is the classic bullwhip pattern: orders amplify as they move upstream, supply responses arrive out of phase with actual demand, and the system oscillates between shortage and surplus before it finds equilibrium.

To understand where and when load shows up, it helps to map the AI infrastructure supply chain explicitly. The pipeline runs roughly as follows:

1. Model providers (OpenAI, Anthropic, Google DeepMind) drive inference demand
2. Cloud and neocloud infrastructure provide the compute layer to serve inference (AWS, Azure, CoreWeave)
3. GPU providers supply the hardware required to run inference (NVIDIA, AMD)
    a. Semiconductor Manufacturers provide the components needed to build GPUs (TSMC CoWoS packaging, HBM from SK Hynix, Micron, Samsung)
4. Datacenters house GPUs and keep them running
5. Utilities/IPPs supply power to data centers



Aishwarya Mahesh
(212) 785-1555
Aishwarya.Mahesh@admis.com

Chip Layer

The surge in inference demand flows directly into demand for AI accelerators, but the semiconductor supply chain is not a monolith. Producing a finished, deployable GPU requires specialized components and manufacturing steps, each with their own capacity constraints and lead times that cannot scale in lockstep. Two in particular are binding right now:

HBM

High-bandwidth memory is the specialized memory stacked directly onto AI accelerator chips. It is what allows a GPU to move data fast enough to run large model inference at useful speeds. Unlike standard memory, it is manufactured by only three companies globally: SK Hynix, Micron, and Samsung. It cannot be substituted, stockpiled from alternative suppliers, or replaced with off-the-shelf components. An AI accelerator without sufficient HBM cannot do useful work at rated capacity. It can physically exist; it cannot run. For power markets, this matters because a data center full of non-operational GPU clusters draws no meaningful load. HBM is therefore the first discrete gate in the conversion sequence between announced capacity and grid-visible demand, and it is currently closed.

Micron's entire 2026 HBM output is already covered by price-and-volume agreements, as was its 2025 output. SK Hynix has described customer demand as exceeding its supply capability even as it expands its M15X fab and Cheongju packaging plant. Samsung expects HBM sales to more than triple in 2026, a figure that reflects how far behind supply has run, not how abundant it is about to become. The HBM total addressable market is projected to grow from $35 billion in 2025 to $100 billion by 2028, a trajectory that reflects genuine demand but also that HBM3E and HBM4 stacking requires the same leading-edge infrastructure already running near capacity.[3]

HBM limits how many accelerators can become operational regardless of GPU inventory, data center construction, or grid capacity permitted. Until Micron's Singapore facility, Samsung's HBM4 ramp, and SK Hynix's Cheongju expansion contribute meaningfully at scale, late 2026 into 2027 at the earliest, this gate does not see relief.

Advanced Packaging

Even with HBM available, a GPU die and memory stack are not a deployable accelerator until they are physically integrated. TSMC's CoWoS process is the dominant technology for that integration. Without it, components exist but cannot ship into a complete product.

TSMC's own commentary has been unusually direct. On its Q4 2024 earnings call, it described advanced packaging as "the biggest constraint at the moment." CoWoS capacity more than doubled in 2024 and was targeted to double again in 2025, and still could not fully satisfy demand. By early 2026, TSMC was describing capacity as "very tight" and expanding OSAT coordination. Throughput was approximately 75,000 wafers per month at end-2025, targeting 90,000–130,000 by end-2026 and 150,000-plus in 2027.

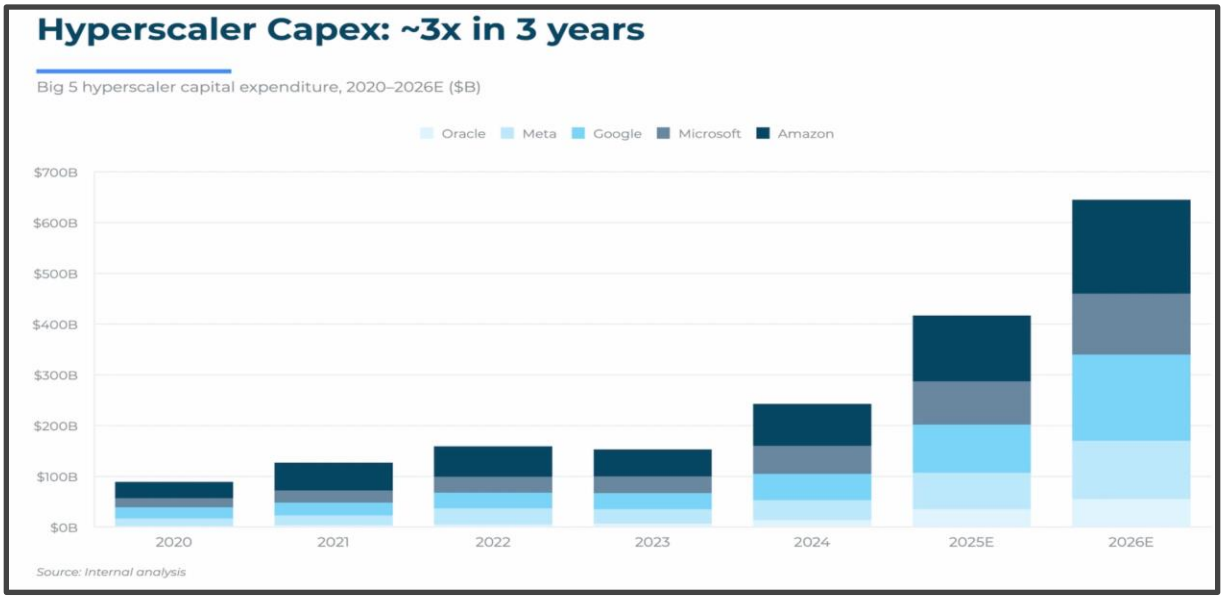
Dell's backlog data confirms the throughput problem directly. Exiting Q4 FY2026, Dell held $43 billion of AI server backlog. It received $34.1 billion of AI orders in the quarter and shipped $9.5 billion. That is not a

Aishwarya Mahesh
(212) 785-1555
Aishwarya.Mahesh@admis.com

demand problem. The upstream supply chain cannot convert committed orders into deliverable systems at market-clearing pace.

Both HBM and CoWoS are prerequisites for the same finished product, so neither constraint eases before the other. The first gate does not open until both resolve, and both are on the same 2027 timeline.

Datacenter Layer



When semiconductor constraints begin to ease, packaged systems start moving downstream into servers and clusters, and clusters into campuses. But even when chips arrive, data hall buildout, cooling, networking, and server integration introduce a further lag. The construction cycle for hyperscale facilities typically runs 12 to 24 months from groundbreaking to energization.

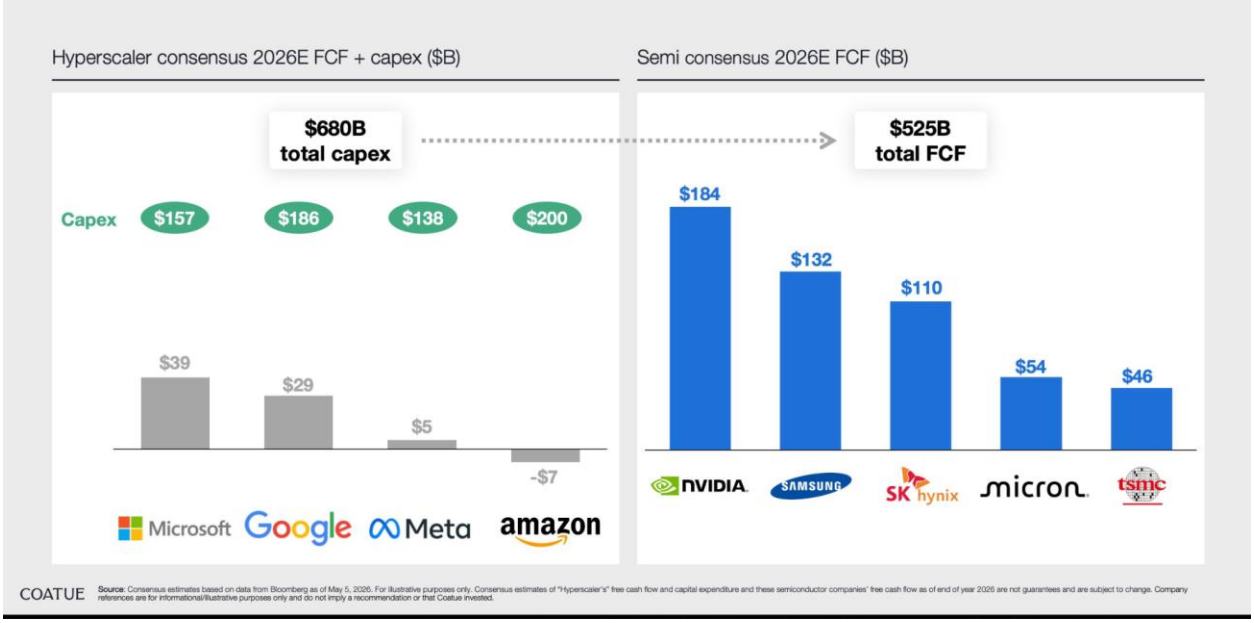
In Q4 FY2026, Dell booked $34.1 billion in AI server orders and exited the quarter with $43 billion in AI server backlog. Over the full fiscal year, Dell secured over $64 billion in AI server orders and shipped more than $25 billion, meaning roughly $39 billion in committed orders did not convert to shipped hardware within the year they were placed. The server backlog data shows the supply chain cannot convert committed orders into deliverable systems at the same pace capital is being committed. This gap between order and shipment is evidence of the bullwhip oscillation at the data center layer.

Tech had a strong quarter with the most recent round of earnings calls. AWS grew 28% year over year in Q1 2026, Google Cloud grew 63%. Despite that, even Amazon's own CFO noted on their earnings call that AWS must lay out cash for land, power, buildings, chips, servers, and networking gear typically six to twenty-four months before it can start billing customers, an unusually direct acknowledgment that committed capital and grid-visible load are separated by a conversion lag.

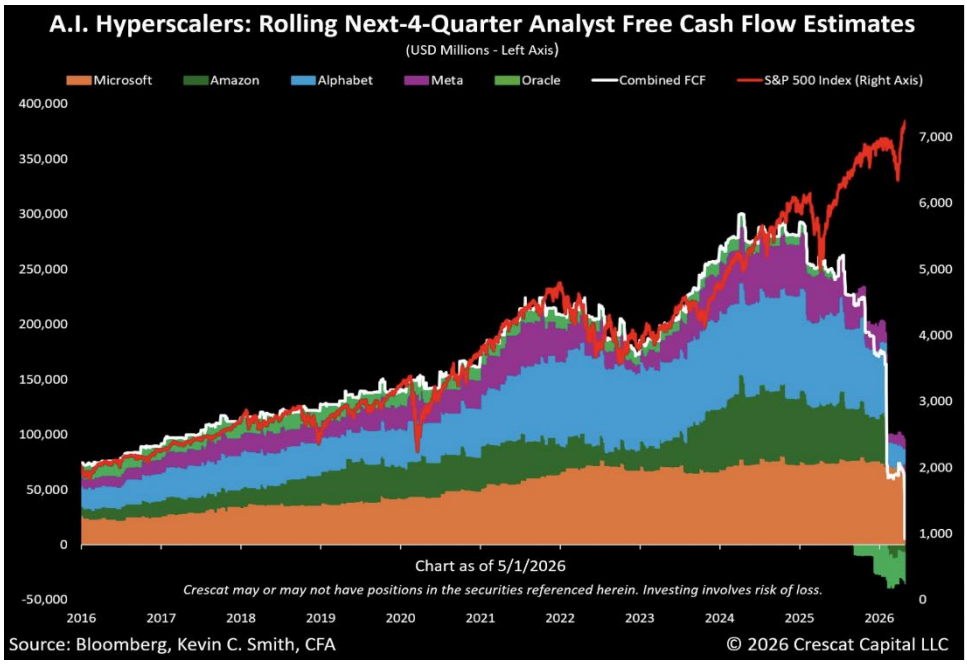
The chip related constraints from Layer 1 are also feeding back into this layer in real time. Meta raised its 2026 capex guidance mid-cycle from $115–135 billion to $125–145 billion, attributing the increase explicitly to higher component costs, particularly memory pricing. HBM scarcity is actively inflating the capital commitments

Aishwarya Mahesh
(212) 785-1555
Aishwarya.Mahesh@admis.com

being made at this layer, which means the oscillation is showing up simultaneously in backlog, deployment lag, and capex inflation.



The financial markets are actively pricing the conversion lag. Rolling next-four-quarter free cash flow estimates for AI hyperscalers grew steadily through 2024, peaked in late 2025, and have since collapsed sharply as capex commitments outpace deployable capacity and near-term revenue generation.



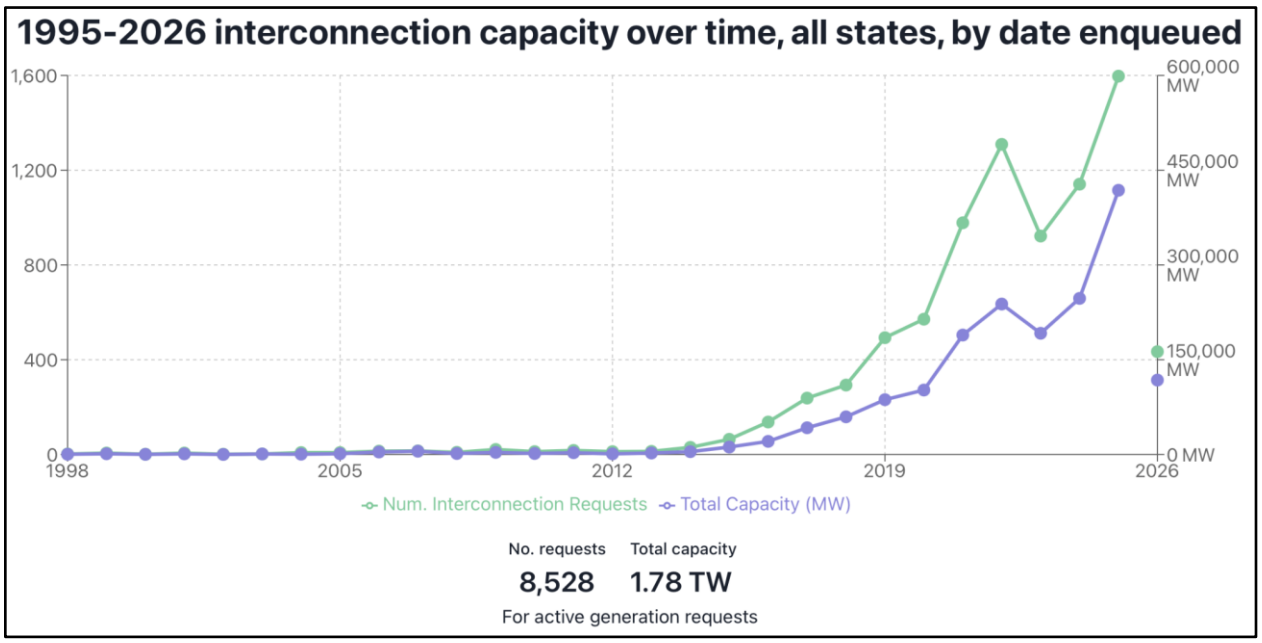
This is analyst consensus marking down future cash flows not because demand has softened, but because capital is being deployed into a constraint stack that has not yet resolved.

Aishwarya Mahesh
(212) 785-1555
Aishwarya.Mahesh@admis.com

Power Layer

When semiconductor constraints ease in 2027 and packaged systems finally accelerate deployment into data centers, electrical service becomes the new binding constraint. Unlike chips, which ship in volume and accumulate, grid capacity resolves as a step function. An AI datacenter campus either clears interconnection and comes online, or it doesn't. This process is one of the longest in the stack.

LBNL's most recent interconnection queue snapshot showed approximately 10,300 active projects representing 1,400 GW of generation and 890 GW of storage requests nationally. Average grid-connection study times run approximately 40 months at major RTOs. The average time from interconnection request to operation for completed projects reached nearly five years in 2023. FERC's Order 2023 reforms govern new entrants; they do not compress the backlog that already exists.



Source: Interconnection.fyi

Transformer lead times are the most concrete downstream bottleneck. According to the IEA's survey of leading industry players, it now takes up to four years to secure large power transformers, nearly double the lead times of 2021. Wood Mackenzie data states power transformers averaged 128 weeks and generator step-up units 144 weeks in Q2 2025, with demand for large power transformers up 116% since 2019 against domestic manufacturing capacity that has not kept pace. Wood Mackenzie projects a 30% supply deficit for power transformers in 2025. Equipment needed for 2027 load needed to have been ordered in 2024–2025. Equipment ordered today arrives in 2028 at the earliest.

A campus that has cleared permitting, completed construction, and received interconnection approval cannot energize until the substation hardware is physically installed. That sequencing means the power constraint gates the last mile, after all other capital has been committed.

The repricing in capacity markets has already occurred on the signal of this demand, not on the load itself. PJM capacity auction clearing prices surged from approximately $28/MW-day to $329/MW-day on the initial indication of AI-driven load growth. In ERCOT, 137 new interconnection requests representing approximately

Aishwarya Mahesh
(212) 785-1555
Aishwarya.Mahesh@admis.com

140,000 MW have been submitted and are not yet reflected in current queue charts, with major transmission expansions not expected until after 2030 in a grid with no meaningful import capability to buffer the transition.

Behind the Meter

One response to the interconnection bottleneck is to bypass it entirely. Hyperscaler-scale campuses increasingly route around it with behind-the-meter (BTM) gas generation, self-supplying power on-site rather than waiting in the grid queue. It is important to note, BTM is not a clear resolution to grid constraints. It *substitutes* the transformer/interconnection gate for a turbine-supply gate, and turbines are on a five-year backlog. GE Vernova, Siemens Energy, and Mitsubishi are the only three at-scale suppliers. As of Q1 2026 GE Vernova reported 2026–2027 largely sold out, was booking into 2028–2029, had only ~10 GW of slots left across 2029–2030 combined, and expects reservations sold out through 2030 by year-end 2026. Mitsubishi is sold out into 2028. Only about one-third of GE Vernova's reservations are data-center-aligned.

Labor

Labor does not fit neatly as a fourth discrete layer. Unlike HBM, CoWoS, or a grid interconnection study, skilled labor is a diffuse constraint that applies across all three layers simultaneously. Its importance for the bullwhip argument is how it prevents the cycle from self-correcting.

In a normal bullwhip cycle, oversupply eventually brings the system back into balance: orders get canceled, excess inventory is worked off, and the supply chain right-sizes. Labor is what prevents that correction from happening quickly. Projections suggest more than 300,000 new electricians are needed over the next decade to meet AI-driven demand, with nearly 30% of the current union electrician workforce between the ages of 50 and 70 and approximately 20,000 retiring each year. The consequences are already showing up in project timelines: Oracle shifted data center completion dates from 2027 to 2028 partly due to labor shortages, while Microsoft is employing electricians commuting from as far as 75 miles away to fill roles.

The industry is responding, but on a timeline that further illustrates the problem. Meta and CBRE recently announced LevelUp, a multiyear program to recruit and train thousands of fiber technicians and trade workers specifically for Meta's data center construction sites. CBRE's CEO noted candidly that "it is really, really hard to get those people." The fact that a company of Meta's scale has to stand up its own training pipeline to source construction labor is an indicator of how tight labor has become. Electrician certification and high-voltage expertise develop over multi-year career paths, and they do not respond to capital commitment the way fab capacity does.

The practical consequence is that even as downstream constraints begin to ease in 2027, a parallel surge in construction and commissioning work hits the same workforce already in short supply. The bullwhip cycle takes longer to resolve than the semiconductor timeline alone would suggest, and the floor on how quickly new capacity can be deployed is higher than historical data center buildout cycles would imply.

Aishwarya Mahesh
(212) 785-1555
Aishwarya.Mahesh@admis.com

Quantifying the Overshoot

Translating the constraint sequence into estimated annual U.S. AI data center capacity additions requires treating each layer as a throughput gate. For any given year, the binding constraint (the layer with the lowest effective ceiling) determines how much pipeline capacity actually converts into grid-visible load. Critically, these constraints operate concurrently rather than serially. A data center with grid power but no operational GPUs draws no AI load; a warehouse of GPUs with no energized campus draws no load either. Deployment requires simultaneous clearance across every layer, which means the effective ceiling in any year is set by the lowest throughput.

The Model

The base-case estimate for each year is derived from the following framework:

$$\text{Visible Load}(t) = \min(\text{Demand}(t), C_{CHIP}(t), C_{GRID}(t), C_{BTM}(t), C_{Labor}(t))$$

where each C_i represents the maximum deployable GW that constraint layer *i* permits in year *t*, estimated by converting layer-specific capacity data { HBM production volumes, CoWoS wafer throughput, data center construction completions, grid interconnection clearances and transformer deliveries, labor availability, etc} into equivalent rack-power GW. To estimate the overshoot, we calculate a **conversion rate**, *Visible Load ÷ Pipeline Demand*.

Additional Assumptions

This model is a mid-2026 snapshot of load ceilings, meaning the input figures re-rate as the industry engineers around its constraints, much as the shift to BTM gas emerged to route around grid constraints. Given that, these figures should be read as the reality today, not a ceiling on human ingenuity.

The constraint ceilings are built from observable industry data such as company filings, earnings-call commentary, equipment vendor disclosures, interconnection queue data, transformer lead-time surveys, turbine order books, and data center construction pipeline estimates. Each constraint is then converted into a common GW-equivalent ceiling using a GB300-class rack as the anchor figure. The softer assumptions are on allocation such as how much global chip supply lands in the U.S.-sited capacity, how much U.S. data center growth is AI-driven, and how much of the pipeline can bypass grid constraints through behind-the-meter supply. The last of which is modeled as a range, giving the grid ceiling its lower (grid-only) and upper (BTM-relieved) band. Those assumptions affect the level of the conversion band, but the shape of the finding is still set by the lowest constraint across the layers defined earlier.

Aishwarya Mahesh
(212) 785-1555
Aishwarya.Mahesh@admis.com

Table 1. Constraint-Implied Deliverability, 2026–2030

Year	Demand (pipeline)	Chip GW	Grid GW (–BTM / +BTM)	Labor GW	Ceiling band	Binding	Conversion band
2026	~18 GW	4.5	3.8 / 4.8	9.5	3.8–4.5 GW	Grid (CoWoS co-binds)	~21–25%
2027	~22 GW	7.0	5.5 / 8.0	10.0	5.5–7.0 GW	Grid; turbine-gated BTM	~25–32%
2028	~26 GW	11.0	8.5 / 13.0	11.0	8.5–11.0 GW	Grid / labor co-bind	~33–42%
2029	~28 GW	15.0	11.5 / 17.0	12.5	11.5–12.5 GW	Grid; labor at top	~41–45%
2030	~30 GW	18.0	14.0 / 20.0	14.5	14.0–14.5 GW	Grid + Labor	~47–48%

Cumulative base-case conversion sums to **~43–50 GW of grid-visible AI load against a ~124 GW pipeline, a through-cycle conversion rate of roughly 35–40%**. The difference is demand staged at various points in the conversion sequence, held in a pre-visible state by the constraint operating at each layer.

For context on what sits in that pipeline: OpenAI has publicly targeted 30 GW of available compute by 2030, up from 1.9 GW at year-end 2025, and NVIDIA's data center revenue in fiscal 2026 reached $193.7 billion.

The Implication

For power market participants, the bullwhip effect is a structural source of pricing risk, and it sits disproportionately with infrastructure providers. The feedback loop works in both directions. On the way up, the demand signal amplifies at every layer: model providers signal aggressive compute needs, hyperscalers place large chip orders, fabs ramp capacity, data center developers break ground, and grid operators receive interconnection request with each layer responding to an already-amplified version of the original signal.

On the way down, the correction can be more sudden. Fransoo et al. describe the coupling risk explicitly. If grid capacity falls short, operators defer turning on new GPU racks; chip suppliers then face order cancellations; upstream fabs see demand evaporate; and the logjam reverberates back through a supply chain that had been building toward a load that never fully arrived. The two supply chains, semiconductor technology and power infrastructure run on different timescales, which amplifies the oscillation

The layers above the grid can serve as leading indicators of where that oscillation currently sits. HBM sold out through 2026 tells you the semiconductor gate is not opening this year, which means data centers under construction will not draw full load on the timelines their interconnection applications assumed. CoWoS throughput tells you how fast packaged systems can reach campuses once HBM resolves. Transformer lead times tell you when energized campuses can begin drawing load. Analysts watching only the interconnection queue are reading the last chapter of a story that starts at the fab.

In a bullwhip cycle, the entity furthest upstream from end demand absorbs the most amplified signal. Power infrastructure providers (utilities, developers, grid operators) sit at the base of the AI supply chain. They are making 20-to-30-year investment commitments based on a demand signal that has been amplified at every layer above them. The hyperscalers at the top of the chain can adjust capex guidance in a quarterly earnings call; a transmission line or substation cannot be unbuilt. The PJM repricing event occurred before the load

Aishwarya Mahesh
(212) 785-1555
Aishwarya.Mahesh@admis.com

materialized. If base-case conversion runs at roughly 40% of the pipeline, the capacity prices that cleared on the full pipeline assumption carries embedded overshoot risk. A risk that sits with infrastructure providers, not the hyperscalers who generated the signal.

What makes this cycle harder to navigate than prior ones is between two supply chains, semiconductor infrastructure and power infrastructure, that are each responding to the same demand spike on irreconcilable timescales. The timing at which it clears each gate, and the rate at which it ultimately becomes grid-visible load, is governed by a sequence that power markets have not historically had to model.

The Bottom Line

What makes the current overshoot structurally worse than normal forecast error is a specific mechanism Lee, Padmanabhan, and Whang identified in their seminal 1997 work known as the rationing game. When a scarce input is being allocated, rational buyers over-order deliberately, knowing that allocations will be cut and that securing excess is the only protection against being caught short.

The result is that the pipeline overstates true need through two independent mechanisms (1) genuine forecast uncertainty about AI adoption and (2) strategic inflation by every layer gaming scarcity. The 124 GW pipeline against a ~50 GW base-case conversion is further evidence of this. This is rational economic behavior under constraint, which means it does not self-correct the way a forecasting error does when better information arrives. That is why structural demand does not immunize a supply chain from bullwhip dynamics.

The core risk (rather than demand materialization) is if AI demand converts into physical load on the timeline the market is underwriting. AI infrastructure is being built through layers that clear on different clocks: chips, packaging, servers, data centers, transformers, interconnection, and labor. Some degree of overbuild or underbuild is therefore not a forecasting accident. Rather, it is the natural result of trying to finance a continuous demand curve through a discontinuous and discrete physical constraint stack. In other words, it is what the system produces when *fast money is used to finance slow steel*.

The practical implication is to watch downstream signals with greater discipline. The recent move by both CME and ICE to launch GPU compute futures markets is worth noting as a leading indicator. Compute pricing volatility now has a forward curve, and that curve will increasingly telegraph where AI infrastructure demand is heading. In addition, hyperscaler capex guidance, co-location lease absorption rates, semiconductor supply chain, and GPU utilization disclosures are the canary. By the time the overbuild is visible at the grid level, the capital allocation decision will already have been made.

The data, comments and/or opinions contained herein are provided solely for informational purposes by ADM Investor Services, Inc. ("ADMIS") and in no way should be construed to be data, comments or opinions of the Archer Daniels Midland Company. This report includes information from sources believed to be reliable and accurate as of the date of this publication, but no independent verification has been made and we do not guarantee its accuracy or completeness. Opinions expressed are subject to change without notice. This report should not be construed as a request to engage in any transaction involving the purchase or sale of a futures contract and/or commodity option thereon. The risk of loss in trading futures contracts or commodity options can be substantial, and investors should carefully consider the inherent risks of such an investment in light of their financial condition. Any reproduction or

Aishwarya Mahesh
(212) 785-1555
Aishwarya.Mahesh@admis.com

retransmission of this report without the express written consent of ADMIS is strictly prohibited. Again, the data, comments and/or opinions contained herein are provided by ADMIS and NOT the Archer Daniels Midland Company. Copyright (c) ADM Investor Services, Inc.

Model Data Sources

Dell Technologies, Inc. *Fourth Quarter and Fiscal Year 2026 Financial Results* (AI server backlog ~$43B; orders vs. shipments). https://investors.delltechnologies.com

EPRI. *Powering Intelligence: Data Center Load Scenarios (2026 Update)* (U.S. data-center capacity scenarios, 2024–2030, used as the demand-case band). https://www.epri.com

GE Vernova Inc. *Form 8-K, First Quarter 2026 Results*, April 22, 2026 (gas turbine backlog and slot-reservation agreements; ~10 GW of 2029–2030 slots remaining; reservations expected sold out through 2030 by year-end 2026). https://www.sec.gov (EDGAR CIK 0001996810). Slot-tightening detail via Power Engineering, April 23, 2026, https://www.power-eng.com; sold-out-through-2030 guidance via Utility Dive, December 11, 2025, https://www.utilitydive.com/news/ge-vernova-gas-turbine-investor/807662/; "2026–2027 largely sold out" and "one-third data-center-aligned" via Axios, May 23, 2025, https://www.axios.com/2025/05/23/ge-vernova-ai-gas-turbine-demand.

Lawrence Berkeley National Laboratory (LBNL). *Queued Up: 2025 Edition* (interconnection queue volumes; median time to commercial operation). https://emp.lbl.gov/queues

Micron Technology, Inc. *Fiscal Q1 2026 Earnings Call and Investor Presentation* (calendar-2026 HBM supply committed on price and volume; HBM TAM $35B in 2025 to $100B by 2028), December 17, 2025. https://investors.micron.com

Morgan Stanley Research. *AI Accelerator and GPU Shipment Estimates, 2026* (NVIDIA unit-shipment forecast; CoWoS wafer allocation). Cited via industry coverage, 2026.

NVIDIA Corporation. *Financial Results for Fourth Quarter and Fiscal 2026* (data center revenue; Blackwell/GB300 ramp), February 25, 2026. https://www.sec.gov/Archives/edgar/data/0001045810/000104581026000019/q4fy26pr.htm

Primary Venture Partners. *The Gas Turbine Bottleneck Reshaping Energy Infrastructure*, April 1, 2026 (three at-scale turbine suppliers; ~5-year backlogs; Mitsubishi sold out into 2028). https://www.primary.vc/articles/the-gas-turbine-bottleneck-reshaping-energy-infrastructure-ex8qe

SK Hynix, Micron, and Samsung. *Earnings Calls and Investor Presentations, 2025–2026* (HBM output and shortage commentary through 2027; HBM4 ramp). HBM TAM projections from Yole Group and TrendForce.

TSMC. *Q4 2024 Earnings Call Transcript* (advanced packaging as binding constraint). CoWoS wafer-throughput estimates from TechInsights, Bernstein Research, and TrendForce, 2025–2026.

Wood Mackenzie. *Untangling the US Transformer Supply Chain Crisis*, August 2025 (transformer lead times: 128-week power, 144-week GSU; 30% supply deficit in 2025; turbine pricing to ~$600/kW by 2027). https://www.woodmac.com/press-releases/power-transformers-and-distribution-transformers-will-face-supply-deficits-of-30-and-10-in-2025/

Aishwarya Mahesh
(212) 785-1555
Aishwarya.Mahesh@admis.com

Works Cited

Associated Builders and Contractors. *Construction Workforce Shortage Tops 439,000 in 2025*. https://www.abc.org

CBRE. *2025 Global Data Center Trends Report*; Uptime Institute. Global Data Center Survey 2025.

CBRE Group, Inc. *Meta and CBRE Announce LevelUp, a Multiyear Program*. https://ir.cbre.com; CoStar, https://www.costar.com

CME Group and Silicon Data. *CME Group and Silicon Data Partner to Launch First Compute Futures*, May 12, 2026. https://www.cmegroup.com

DataCenterDynamics. *SK Hynix to Invest $12.85bn in Advanced Packaging Plant in Cheongju, South Korea*. https://www.datacenterdynamics.com

Epoch AI. *LLM Inference Price Trends*. https://epoch.ai

ERCOT Large Load Integration Team. *Large Load Interconnection Status Update*, March 13, 2026. https://www.ercot.com; Zero Emission Grid, March 19, 2026, https://www.zeroemissiongrid.com.

Fortune. *AI Data Centers Face an Electrician Shortage*, March 2, 2026. https://fortune.com

Fransoo, J. et al. (2025). *The AI Infrastructure Boom and Supply Chain Dynamics.* Cited in Gad Allon, "AI's Great Infrastructure Boom: Bullwhip or Building the Future?" Substack, December 1, 2025. https://gadallon.substack.com/p/ais-great-infrastructure-boom-bullwhip

Intercontinental Exchange (ICE) and Ornn. *ICE and Ornn to Launch GPU Compute Futures Contracts*, May 19, 2026. https://www.businesswire.com

International Energy Agency (IEA). *Building the Future Transmission Grid*, 2025. https://www.iea.org/reports/building-the-future-transmission-grid

OpenAI Newsroom. *OpenAI Targets 30GW of Compute by 2030*, April 22, 2026. https://www.androidheadlines.com

SpaceXAI. *New Compute Partnership with Anthropic.* https://x.ai/news/anthropic-compute-partnership

Tom's Hardware. *Skyrocketing Component Prices Push Big Tech Capex to Record $725 Billion*, May 1, 2026 (Meta capex guidance; AWS and Google Cloud growth). https://www.tomshardware.com

Utility Dive. *PJM Capacity Prices Set Another Record with 22% Jump*, July 23, 2025. https://www.utilitydive.com