THEMATIC RESEARCH · AI INFRASTRUCTURE, POWER & GRID

The Power Wall

The AI Premium Migrated From the Hyperscalers to the Hardware Makers. The Hardware Is a Shortage Rent That Prices Away in Two Years. Power Is Where the Premium Actually Stays — and the Interconnection Queue Is Why

TON618 CAPITAL RESEARCH · AS OF JULY 19, 2026 ·

Every figure is drawn from company filings, grid-operator and government data, or dated public reporting, and is separated throughout into fact ("the data shows…") and estimate ("we infer…"). This is the supply-side note in a five-note series — start with the reader's guide, which maps how the pieces connect. The AI Capex Payback Clock established that ~$700bn of annual AI capex is transferring cash from the hyperscalers to their suppliers; the NVIDIA valuation values the largest single beneficiary. This note asks the question those two leave open — whether the profit that has migrated to the hardware layer stays there — and finds that it does not stay in silicon; it stays in power. The argument turns on a single number: the time it takes to bring new supply to market.

§0 The Argument

~5 years

THE MEDIAN WAIT TO CONNECT NEW POWER TO THE US GRID — UP FROM THREE YEARS A DECADE AGO, AND STILL LENGTHENING EVEN AS THE RAW QUEUE COMES OFF ITS 2023 PEAK. THE WAIT, NOT THE CHIP, IS THE BINDING CONSTRAINT

The claim. For over a decade the technology profit pool sat with the hyperscalers — asset-light platforms earning software margins on other people's compute. The AI wave reversed the polarity: to compete, the hyperscalers must now buy hardware- and power-intensive systems at unprecedented scale, and in doing so they are handing the premium down the stack. But the premium **bifurcates as it lands**. The hardware slice — chips, memory, GPUs — is a **shortage rent with a roughly two-year half-life**: advanced-packaging capacity is doubling annually and H100 rental rates have already fallen **64–75%** from peak. Supply catches demand, and the rent prices away. Power is different. You cannot manufacture a gigawatt on a two-year clock. Just three companies make the large gas turbines, and they are **sold out through 2030**; large transformers run **three-to-five-year** lead times; and the median grid-interconnection wait has stretched to **about five years**, with roughly **three-quarters of projects withdrawing** before they ever connect — while the queue itself, still ~2,000 GW, is being worked down by attrition, not by connection. The chips are the trade. The power is the investment — because it is the one input the buildout cannot bring to market fast enough to compete its own premium away.

A shortage rent is worth exactly as long as the shortage lasts, and a shortage lasts exactly as long as new supply takes to arrive. That single principle sorts the entire AI-infrastructure complex into two piles. Where supply can be added in months, today's fat margins are a temporary scarcity price that competition and capacity will erode — real money, but rented, not owned. Where supply takes years, the scarcity is structural, the pricing power is durable, and the premium stays put. This note applies that test to the physical inputs of the AI buildout and finds a clean split: the semiconductors are in the first pile, and power — generation, turbines, transformers, and above all the grid connection itself — is in the second.

The evidence is already in the cash flows. As our companion notes documented, hyperscaler free cash flow peaked in 2024 and rolled over in 2025 as capital expenditure overran even their surging operating cash. That capital is flowing to suppliers. The question this note settles is *which* suppliers keep it. The answer is the ones whose product cannot be built any faster than it already is.

~62 → 134 GW
US data-center power demand, 2025 → 2030E — toward ~9% of the entire grid

61 months
Median interconnection wait for a 2025 project — up from 36 months in 2015, and lengthening

Sold out to 2030
Large gas-turbine order books across all three global OEMs

~2 yrs
Half-life of the hardware shortage rent, by contrast — the migrating premium's fast-decaying slice

§1 The Migration's Last Stop

The profit moved down the stack. The question is where it stops moving — and that is a question about lead times, not about technology.

The thesis these three notes share is a profit migration. For fifteen years the scarce, value-capturing layer of computing was software: compute itself was abundant and cheap, so the rent accrued to the platforms that sat on top of it — the hyperscalers, earning extraordinary returns on very little capital. Artificial intelligence inverted that structure. It made the *physical* inputs scarce again — leading-edge silicon, high-bandwidth memory, and the electricity to run them — and by the economics of the thing, when a layer becomes scarce and hard to substitute, the premium migrates to it. That is why the picks-and-shovels layer is winning: not sentiment, but the return of a physical bottleneck.

But "the hardware layer" is not one thing, and the premium does not stay evenly across it. Sort the inputs by how fast new supply can be brought online and the migrating premium separates cleanly:

- **The silicon is a shortage rent.** The bottleneck in GPUs was never the logic die; it was advanced packaging (TSMC's CoWoS) and high-bandwidth memory. Both are being added fast — CoWoS capacity is roughly *doubling every year*, on its way from ~35,000 wafers a month toward 120,000–140,000 by the end of 2026. The proof that this is a rent, not a moat, is in the price: H100 rental rates have already fallen **64–75%** from their scarcity peak. Each GPU generation goes scarce, then abundant, then cheap, on the same ~24-month clock that supersedes it. NVIDIA sustains its premium by running that treadmill and by software lock-in — *not* because the chips stay scarce.
- **The power is a structural bottleneck.** Everything that turns a data center into a working machine on the grid — firm generation, the turbines that provide it, the transformers and switchgear that step and route the power, and the interconnection that ties it all in — takes *years* to add, is controlled by a handful of suppliers, and is being demanded by everyone at once. This is not a rent that prices away in two years. It is the wall the entire buildout is now running into.

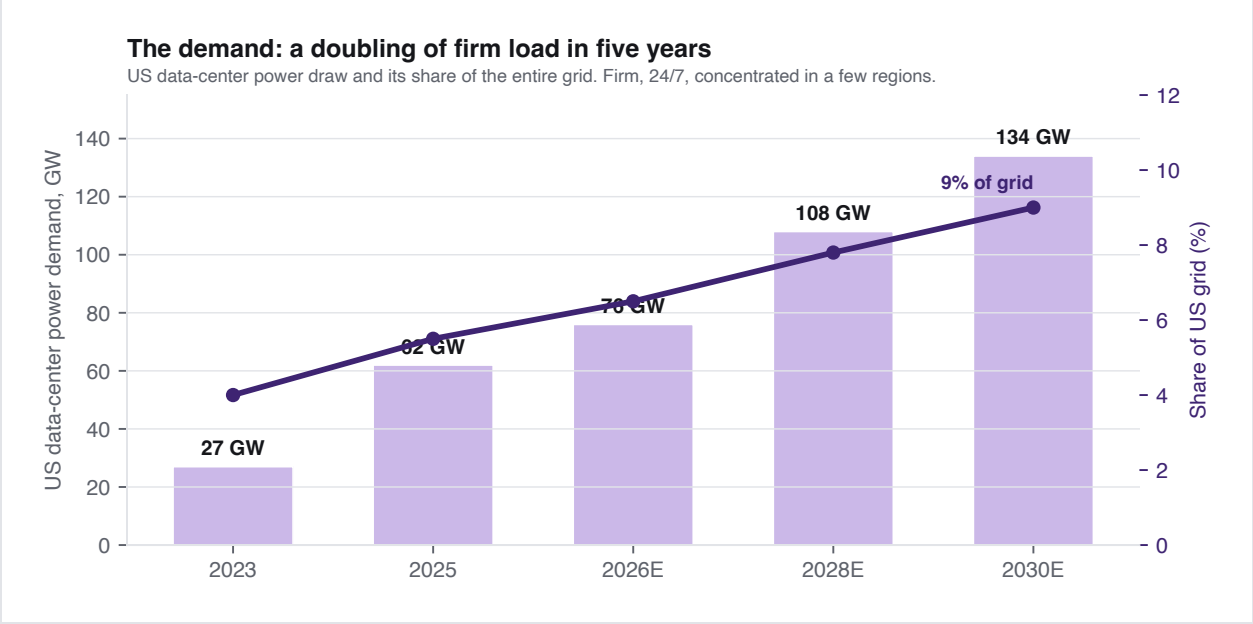
The rest of this note is the second bullet, in numbers.

§2 The Demand

Unprecedented, concentrated, and arriving faster than any grid was built to absorb.

US data-center electricity demand is forecast to run from roughly **62 GW in 2025 to about 76 GW in 2026, ~108 GW in 2028, and ~134 GW by 2030** on S&P Global's series — better than a doubling in five years. Estimates vary widely by scope: Goldman Sachs, on a narrower definition, starts lower (~31 GW in 2025, ~66 GW by 2027); §8 carries the range and the reason the two disagree. As a share of the entire US grid, data centers move from around **4% of load today toward roughly 9% by 2030**. Former Google chief executive Eric Schmidt has testified that data centers will need **29 GW of additional power by 2027 and 67 GW more by 2030**.

Two features of this demand make it uniquely hard on the grid. First, it is **firm** — an AI training cluster wants power 24 hours a day at high utilization, not the intermittent draw the recent wave of grid additions (solar, wind) is built to supply. Second, it is **concentrated** — it lands in gigawatt-scale bites in a few locations (Northern Virginia, Texas, Ohio, Arizona), overwhelming local transmission that was planned for incremental growth. A grid that adds a few percent of load a year is being asked to absorb a data-center load growing at double digits, in firm blocks, in specific places. The demand is not the question. The question is whether the supply can arrive to meet it — and that is where the wall is.



§3 The Queue

The hero of the story, and the cleanest single measure of why power is the durable bottleneck.

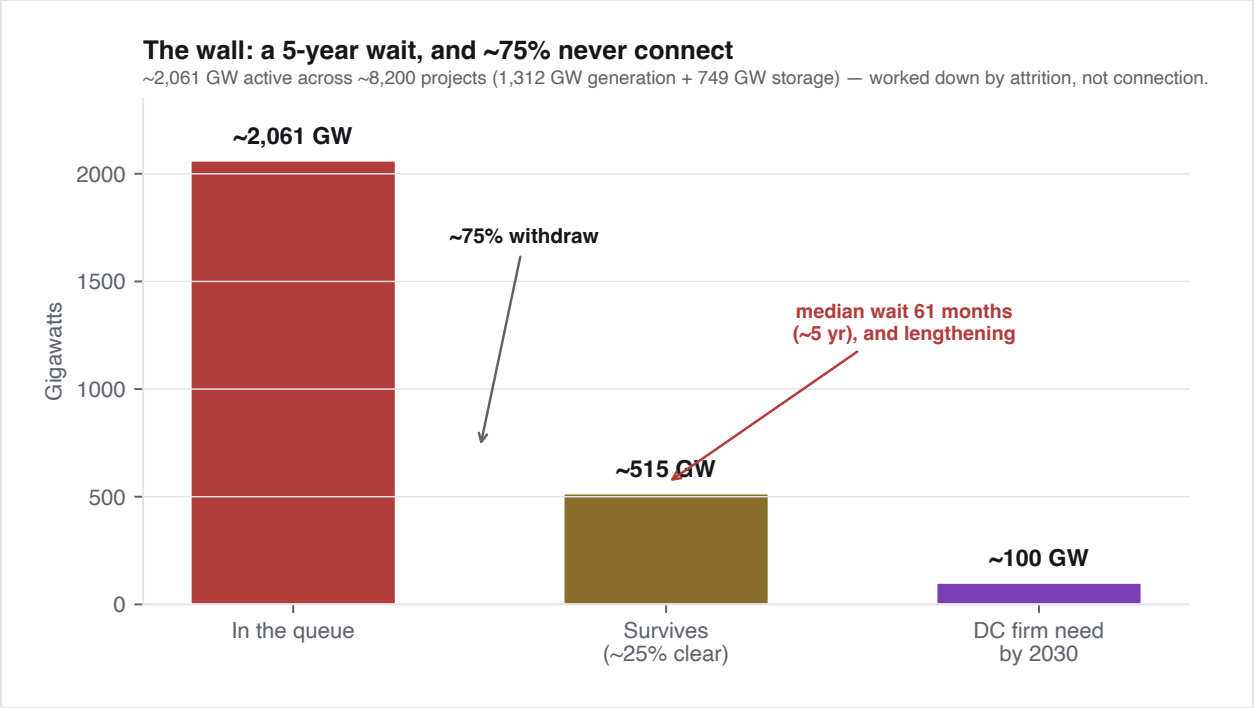
To connect new generation to the US grid, a project must clear the **interconnection queue** — the study-and-approval process run by grid operators. That queue is now the binding constraint on the entire AI buildout, and its numbers are staggering.

At the end of 2025 roughly **8,200 projects were waiting in US interconnection queues, representing about 1,312 GW of generation and 749 GW of storage — some 2,061 GW in total**, roughly 1.7 times the entire installed capacity of the US grid. That total is actually *down* from a peak near 2,600 GW at the end of 2023 —

and the reason it is falling is the tell: not because projects are connecting, but because they are **giving up**. The **median project reaching operation in 2025 waited 61 months — just over five years — up from 36 months a decade earlier**, and load additions at data-center scale sit at the top of that range: Dominion Energy's queue for large commercial-load service stretches **beyond 36 months** for new substation service alone, and JLL's 2026 outlook reports average interconnection timelines now **exceeding four years**. In some regions they run into the next decade.

> *About three-quarters of the projects that enter the US interconnection queue eventually withdraw — defeated by multi-year delays and the unpredictable, often prohibitive cost of the grid upgrades their connection would require.*
>
> ON THE ~75% QUEUE-WITHDRAWAL RATE (2000–2020 REQUEST VINTAGE, BY CAPACITY) — THE ATTRITION THAT SHRINKS THE PILE WITHOUT GROWING THE GRID

The withdrawal rate is the detail that turns a shrinking queue into a hardening wall. Three in four projects never make it through — so the capacity that actually connects is a small fraction of what is queued, and the pile comes down through abandonment, not through hook-ups. That is not a bottleneck easing. In semiconductors, capacity is being added faster than demand and the constraint is genuinely loosening — the price falls, visibly. Here, the raw queue is off its 2023 peak but the metric that decides the thesis, the *time to connect*, keeps lengthening: 36 months a decade ago, 61 months now. A backlog that clears by exhaustion while the wait grows is not a shortage being solved — it is a shortage so severe that most participants quit before they reach the front. Size and duration are moving in opposite directions, and duration is the one that prices power.



Why does this make power a durable premium rather than a temporary one? Because a bottleneck you cannot build your way out of on the buildout's own timescale does not price away. The GPU that rents for a fifth of its 2024 price two years later is a rent competed to the bone. A grid connection whose wait is five years and still lengthening — while demand doubles in five — is a scarcity that persists across the entire investment horizon. Whoever owns firm, connected power in the places the data centers want it holds an asset the buildout cannot reproduce quickly, and that is the definition of durable pricing power.

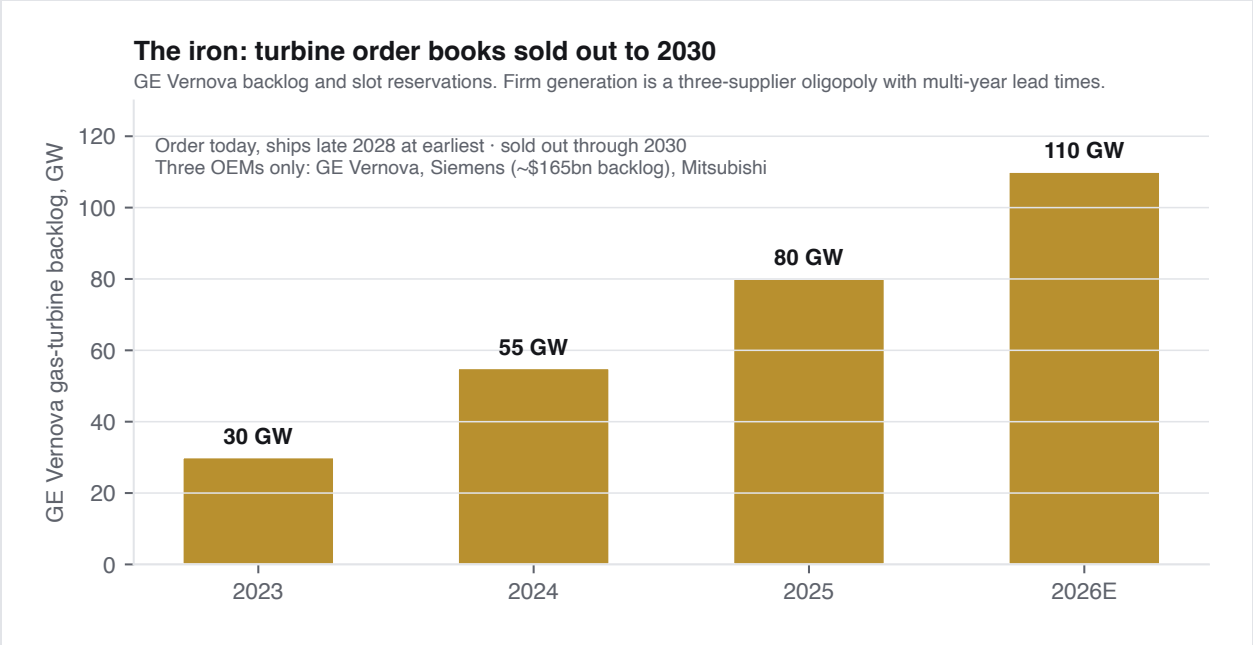
§4 The Iron

Behind the queue stand the machines — and the machines are sold out too.

Even where a project can secure a grid connection, the physical equipment to generate and route the power carries its own multi-year backlogs, controlled by a very small number of suppliers.

Gas turbines — three makers, sold out to 2030. Firm, dispatchable generation at data-center scale means, in practice, natural-gas turbines, and the large-turbine market is a global oligopoly of exactly three: GE Vernova, Siemens Energy, and Mitsubishi Heavy Industries. All three are effectively sold out for years. GE Vernova ended 2025 with an **~80 GW gas-turbine backlog stretching into 2029**, expects to reach roughly **110 GW of backlog and slot reservations by the end of 2026**, and its chief executive expects reservations to be **sold out through 2030**. A new order for its flagship turbine placed today would ship **late 2028 at the earliest**. Siemens Energy carries a record backlog of roughly **$165bn**; Mitsubishi's slots are largely booked through 2028–29. Only about a fifth of GE Vernova's contracted volume is explicitly data-center load — the rest is utilities and industry competing for the same slots — which means AI must bid against the entire electrification of the economy for a fixed, slowly-growing supply of turbines.

Transformers and switchgear — the quiet chokepoint. The largest custom power transformers now run **three-to-five-year lead times**; standard power-transformer waits sit around **128 weeks** in 2026 and generator step-up units around **144 weeks**, and medium-voltage switchgear carries multi-year backlogs. The demand shock is visible in the order data — demand for generator step-up transformers rose **274% between 2019 and 2025**. The consequence is direct: industry analysis estimates **30–50% of planned data-center sites are at risk of delay or cancellation** because the electrical equipment to energize them cannot be procured on schedule.



Every item in this section shares the property that defines a durable premium: a small supplier base, a multi-year lead time, and demand from the entire economy at once — not just AI. That is why the pricing power here does not decay the way the GPU's does. You can stand up a new advanced-packaging line in a year and a half. You cannot conjure a turbine hall, a transformer factory, or a cleared interconnection in that time.

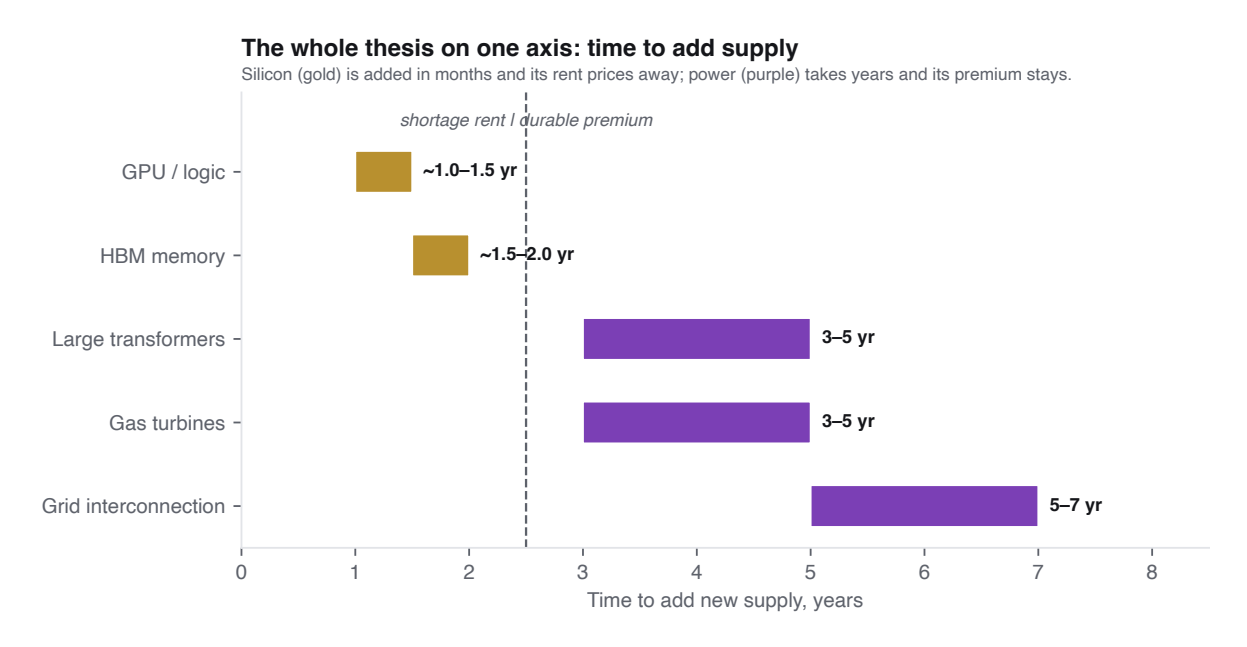
§5 The Lead-Time Race

The whole thesis on one axis: how long it takes to add supply of each scarce input.

Line the bottlenecks up by the one variable that determines whether a premium is rented or owned — the time to bring new supply to market — and the AI-infrastructure complex splits in two.

TIME TO ADD SUPPLY, BY INPUT — THE SORT BETWEEN SHORTAGE RENT AND DURABLE PREMIUM

SCARCE INPUT	TIME TO ADD SUPPLY	SUPPLY RESPONSE NOW	PREMIUM TYPE
GPU / logic silicon	~12–18 mo	Packaging doubling annually	Shortage rent
HBM memory	~18–24 mo	Expanding; historically glut-prone	Shortage rent (cyclical)
Large transformers	~3–5 yr	Backlogs lengthening	Durable
Gas turbines	~3–5 yr	Three OEMs, sold out to 2030	Durable
Grid interconnection	~5–7 yr+	Queue growing faster than it clears	Durable



The top two rows are the semiconductor complex, and their supply response is measured in months and already underway — which is why the H100 rental price has collapsed and why memory, the most cyclical of all, has swung to glut and negative free cash flow inside this very cycle. The bottom three rows are power, and their supply response is measured in *years*, controlled by *handfuls* of suppliers, and *lengthening*. The same AI demand hits both piles; only one pile can answer it quickly. The premium the hyperscalers are handing down the stack is split accordingly: a fast-decaying rent on the silicon, and a durable, multi-year scarcity on the power. **The chips are where the money moves through. The power is where it settles.**

There is a second-order consequence for risk. The silicon premium is exposed to *two* failure modes — supply catching up, or AI demand plateauing (the payback clock). The power premium is exposed to only *one* — a demand plateau — and even that is cushioned, because the multi-year supply lag means power cannot glut quickly the way memory can. A slowdown in AI capex would loosen the GPU market within a quarter or two; it

would take years to unwind a five-year interconnection queue. Power is not only the more durable premium; it is the more defensive one.

§6 The Workaround, and Its Limit

The buildout has a way around the queue. It runs straight back into the turbine backlog.

Faced with five-to-seven-year interconnection waits, the hyperscalers have an answer: **behind-the-meter generation** — build the power on-site, gas turbines or fuel cells, and skip the grid connection entirely. This is the fastest-growing near-term solution, and it is real: it is why so much of the new AI capacity is being sited next to gas supply rather than waiting in a utility queue.

But the workaround does not escape the wall; it relocates to a different part of it. Behind-the-meter gas still needs **the same turbines that are sold out through 2030**, from the same three suppliers. It still needs transformers and switchgear on multi-year backlogs. It trades a grid-interconnection queue for a turbine-delivery queue — faster, perhaps, but gated by the same fixed, oligopolistic supply. The workaround is itself evidence for the thesis: the buildout is so desperate for firm power that it will finance and build its own generation to avoid a seven-year wait, which is precisely the behavior of a bidder facing a scarcity it cannot wait out. That desperation is the premium, made visible.

§7 Winners, Losers, and What Would Change Our Mind

A thesis that cannot name positions is not a thesis. Here are the winners and losers it implies, the time frames over which they should play out, and the signposts that would prove it wrong.

The lead-time test (§5) is also a portfolio map. Sort by time-to-add-supply and the winners are the layers with the longest lead times and fewest suppliers — where the premium is durable; the losers are the short-lead-time layers where the rent prices away, and the players exposed to the buildout *without* owning a durable input. These are directional calls with horizons attached; the specific price targets belong to the valuations that follow, of which the first is named below.

The winners — long lead times, few suppliers, durable premium.

WHERE THE MIGRATING PREMIUM SETTLES

BASKET	REPRESENTATIVE NAMES	ROLE IN THE THESIS	HORIZON
Gas-turbine OEMs	GE Vernova, Siemens Energy, Mitsubishi	The scarcest node — three makers, sold out to 2030	3–5 yr
Firm & nuclear generation	Constellation, Vistra, Talen, NRG	Own the dispatchable power data centers will pay a premium to secure	2–5 yr
Electrical equipment	Eaton, Hubbell, Vertiv, nVent, ABB, Schneider	Transformers, switchgear, power management on multi-year backlogs	2–4 yr
Grid build-out	Quanta Services; GE Vernova Electrification	Someone has to build the interconnection the queue is starved of	3–7 yr

The losers — short lead times, or exposure without the durable input.

WHERE THE PREMIUM IS RENTED, NOT OWNED — OR PAID, NOT COLLECTED

BASKET	REPRESENTATIVE NAMES	ROLE IN THE THESIS	HORIZON
Merchant chip / memory pricing power	Micron (most cyclical); the GPU rental market	Rent prices away in ~2 yr; memory is glut-prone and has already round-tripped once this cycle	~2 yr
GPU landlords without power	Neocloud pure-plays	Renting into a falling GPU curve without owning the durable input; the cleanest short-side expression	1–3 yr
Power-less data-center developers	Sites without secured interconnection	30–50% of planned sites at risk of delay or cancellation on equipment and grid access	2–5 yr
Hyperscaler near-term free cash flow	The buyers	Paying the rent; aggregate FCF already peaked (2024) and rolled over (2025)	1–3 yr

An honest qualification: most of the "losers" are losers of *pricing power and premium*, not of business quality. NVIDIA is the finest business in the complex; its chip rent simply decays. The hyperscalers own the customer relationship and can redeploy their cash the moment they choose to stop spending. The one genuinely fragile position is the GPU landlord that has taken on leverage to rent depreciating silicon into a falling price curve while owning none of the durable input — which is the subject of our neocloud credit work.

Why GE Vernova is the single-name expression — and what role it plays. Of the winners, GE Vernova is the purest liquid one, for reasons specific to the thesis rather than to the company's popularity. It sits at the *scarcest* node — the three-supplier turbine oligopoly with the longest visibility (a backlog booked toward 2030) — and it carries a *second* durable leg in its Electrification (grid) segment, so a single security expresses two of the four winner baskets at once. It is a single US-listed pure-play, unlike Siemens (a foreign multi-industry parent) and Mitsubishi (a conglomerate), so the exposure is clean. And critically, it is the thesis's own **falsification instrument**: if the durable-premium call is right, GE Vernova's backlog, pricing, and margins should compound with visibility few other AI-infrastructure names can offer; if the interconnection wait starts *falling* or turbine demand plateaus, GE Vernova is precisely where that would show up first. It is both the recommendation and the test. A standalone valuation of GE Vernova accompanies this note.

Signposts — observable, and pointed. In the spirit of a falsifiable thesis, the markers that would validate or invalidate the call:

- **The interconnection wait, not the queue size.** The core measure — and the subtlety §3 insisted on. The pile is shrinking by attrition; what matters is the *time to connect*. If median waits start *falling* from 61 months, the power scarcity is easing and the durable premium is at risk. As of now the wait is still lengthening.
- **Turbine order books and lead times.** A plateau or decline in the three OEMs' backlogs — or lead times shortening back toward two years — would signal the bottleneck loosening. Currently backlogs are still growing.
- **GPU rental prices, as the control.** The silicon side of the same trade. Continued decline confirms the hardware rent is pricing away on schedule and validates the split at the center of this note. A *reversal* — GPU rents rising durably — would undercut the "silicon is a rent" half of the argument.
- **Behind-the-meter announcements.** More on-site generation deals confirm the queue is binding; a slowdown would suggest grid access is improving.
- **The payback clock.** The one shared risk to *both* legs: if AI capex plateaus (our companion note's central question), demand for power softens too — but with a multi-year lag that makes power the last part of the

complex to feel it, and the reason it is the more defensive leg.

The migration is real; the cash flows already show it. What this note adds is where it comes to rest, and what to own to hold it. The premium that left the hyperscalers does not stop at the chip — the chip hands most of it back within two years. It stops at the wall the buildout cannot climb on its own schedule: the grid, and the handful of suppliers who build it.

§8 Sourcing and Method

Data-center demand forecasts are from Goldman Sachs, S&P Global, Gartner, EPRI and government/DOE sources current to 2026. The forecasts differ materially by scope — Goldman's narrower definition puts 2025 demand near 31 GW, S&P Global's broader one near 62 GW for the same year — so §2 anchors the trajectory on S&P Global's internally consistent series (~62 GW in 2025 to ~134 GW in 2030) and flags Goldman's lower path rather than splicing the two; the 4%→9%-of-grid share blends consumption-share and peak-demand-share metrics and is stated as directional. Interconnection figures are from Lawrence Berkeley National Laboratory's *Queued Up: 2026 Edition* (June 2026, data as of year-end 2025): ~2,061 GW active (1,312 GW generation + 749 GW storage) across ~8,200 projects, a 61-month median time-to-operation for 2025 projects, and ~75% capacity-weighted withdrawal for the 2000–2020 request vintage — the queue itself is *down* ~10% in 2025 and ~12% in 2024 from its ~2,600 GW end-2023 peak, which §3 treats as attrition, not connection. Data-center-specific waits are from Dominion Energy disclosures and JLL's 2026 outlook (which reports average interconnection timelines exceeding four years). Turbine backlogs and lead times are from GE Vernova and Siemens Energy investor disclosures and trade reporting; transformer and switchgear lead times from PV Magazine, POWER, WoodMac survey data, and industry procurement analysis. The hardware-side comparators (CoWoS capacity, H100 rental decline) carry over from our companion notes, sourced there.

This note makes no price target and initiates no position; like the other thematic notes in the series it is deliberately kept off our signal ledger. It is the structural foundation for a forthcoming valuation of the power-and-grid beneficiaries. The method throughout has been to sort the AI-infrastructure complex by the one variable that decides whether a premium is rented or owned — the time to bring new supply to market — and to follow that sort to where the durable value sits.

Reference data. All filing, grid-operator, and market data as of July 19, 2026; sources are named in the paragraph above and are dated where they are cited. This note carries no target price and takes no position in any name.

§9 Disclosures

Information only. TON618 Capital. This report is for information purposes only. Nothing here is an offer to sell or a solicitation of an offer to buy any security, fund interest, or digital asset, and nothing here is personalized investment advice or a recommendation regarding any instrument.

Publisher's exclusion. All research is published solely as general, impersonal information of regular circulation. It is not tailored to the objectives or circumstances of any individual and is not issued in connection with compensation from any client. The Fund has no clients and distributes all research free of charge. On that basis it publishes in reliance on the publisher's exclusion from the definition of "investment adviser" under the Investment Advisers Act of 1940 (§202(a)(11)(D); cf. *Lowe v. SEC*, 472 U.S. 181 (1985)).

Registration & conflicts. TON618 Capital is not registered as an investment adviser or broker-dealer in any capacity. The Fund is a Bitcoin fund and may hold or transact in the securities or digital assets it discusses. This note discusses, among others, GE Vernova (GEV), Siemens Energy, Mitsubishi Heavy Industries, NVIDIA (NVDA), Micron (MU), and various utilities, independent power producers and electrical-equipment makers; the Fund

holds no position, long or short, in any of them, and has no economic interest in the price of any security named here. The Fund receives no compensation from any party in connection with its research.

Use of AI. Artificial intelligence is used in the creation of this research. All methodology and data integrity are reviewed and approved before publication by TON618 Capital's Chief Investment Officer, **Keyth Beck**; errors may nonetheless occur, and readers should verify independently.

CFA. This report was prepared to align with CFA Institute analytical standards (methodology only). CFA® and Chartered Financial Analyst® are registered trademarks owned by CFA Institute. That reference describes the analytical framework applied; it does **not** imply the report was prepared, reviewed, or authored by a CFA charterholder, and the report is **not** issued, reviewed, endorsed, certified, or approved by — nor affiliated with — CFA Institute.

Risk & feedback. Past performance is not indicative of future results. Digital assets and equities are volatile and may result in total loss of capital. Corrections and feedback are welcome — please direct them to CIO **Keyth Beck** at **keyth@ton618capital.com**.