COMPUTE FORECAST

AI INFRASTRUCTURE · DATA CENTERS · CASE STUDY

Speed as a weapon: how power procurement became the AI compute race

The fastest AI infrastructure operators are energizing GPU clusters in months, not years — by owning their power. We map who is winning the velocity war, and why the enabling mechanism is electricity, not silicon.

COMPUTE FORECAST · JULY 2026 · DATA CENTERS / AI INFRASTRUCTURE

IN BRIEF

- The AI compute race is now decided at the substation. Operators that vertically integrated power — on-site gas generation, dedicated substations, front-of-meter nuclear power purchase agreements — are energizing GPU clusters in months. Those dependent on grid interconnection queues face multi-year timelines.
- xAI's Colossus proved raw speed through on-site generation. Microsoft's Fairwater proved fiber-linked, multi-site scale. Stargate's Abilene campus showed how disciplined capital reallocation works in practice. And the four largest US hyperscalers have turned nuclear power into a multi-year compute-enablement strategy.
- Goldman Sachs Research projects US data center power demand rising from 31 GW in 2025 to 66 GW in 2027 — with only around half of near-term scheduled capacity expected to arrive on time. That gap is why power procurement, not GPU allocation, now sets the pace of the AI build-out.

The new bottleneck

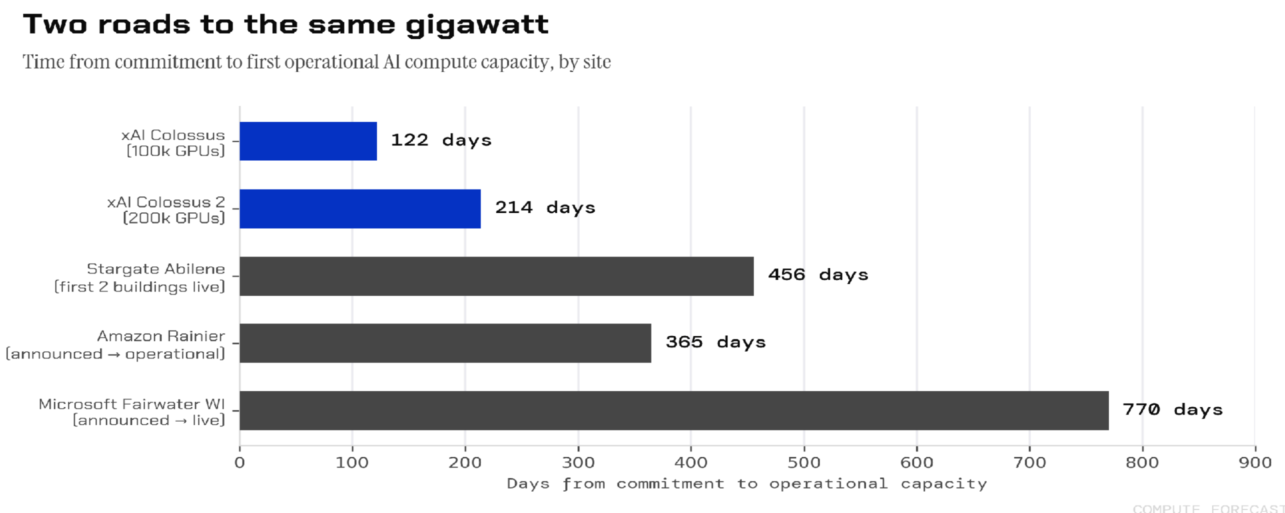
The constraint on how fast an AI lab can train its next model is no longer how many GPUs it can buy. It is how fast it can turn a piece of land into megawatts. Nvidia will sell as many Blackwell racks as anyone can pay for. What separates the operators pulling ahead from the ones falling behind is which side of the meter they are standing on.

Four buildouts make the case with unusual clarity: xAI's Colossus campus in Memphis, Microsoft's Fairwater campus network, OpenAI and Oracle's Stargate site in Abilene, and Amazon's Project Rainier in Indiana. Each took a different route to power. The results are now visible in how quickly — and how reliably — each operator can bring new training capacity online.

xAI Colossus — speed through vertical integration

xAI built its first 100,000-GPU cluster at a converted appliance factory in South Memphis in 122 days, then doubled it to 200,000 GPUs in 92 more days. Nvidia CEO Jensen Huang called the pace "superhuman," noting that a cluster of that scale would typically take three years to plan and another year to commission — roughly four years compressed into months.

The site's utility connection started at just 8 MW. Rather than wait 12 to 18 months for a standard grid upgrade, xAI installed on-site natural gas turbines to run the cluster in parallel with the utility build-out, later adding a 150 MW battery system and funding its own substation. That decision — generate first, connect to the grid later — is the single biggest reason Colossus became operational faster than any comparable AI training campus.



TIME FROM COMMITMENT TO FIRST OPERATIONAL CAPACITY, BY SITE

xAI has outlined plans to expand the Memphis campus toward roughly 2 GW and well over half a million GPUs as additional buildings and turbines come online. The confirmed, currently operating cluster stands at roughly 200,000 GPUs — still among the largest single-site AI training deployments anywhere — with further capacity arriving in stages as new power is energized.

> **"Generate first, connect to the grid later" is the decision that separated Colossus's timeline from every grid-dependent campus built in the same period.**

Microsoft Fairwater — scale through a fiber-linked network

Microsoft took the opposite bet. Its Fairwater campus in Mount Pleasant, Wisconsin — a site once earmarked for a failed Foxconn manufacturing plant — went fully operational on June 23, 2026, backed by a $7.3 billion investment by Microsoft in the state and built with nearly 10,000 construction workers over two years. Rather than racing a single site, Microsoft linked Fairwater Wisconsin to its Atlanta campus, live since October 2025, as part of a nationwide AI network: the company deployed more than 120,000 miles of dedicated fiber across the US in a single year — a more than 25% increase in its total fiber footprint — tying every Fairwater site into one distributed system.

Azure general manager Alistair Speirs described the design plainly: "it's not just a single site training an AI model, it's a network of sites supporting that one job." Atlanta, notably, runs on grid power alone — no on-site generation, no backup generators — a bet that grid reliability itself can be engineered for "four nines of availability at three nines of cost." Where xAI chose to own generation, Microsoft chose to network reliable grid capacity across geographies. Both routes produced live, world-class training capacity within roughly two years of groundbreaking.

Stargate Abilene — a disciplined recalibration

OpenAI and Oracle's flagship Stargate campus in Abilene, Texas is now capped at its committed 1.2 GW and roughly 450,000 GB200 Blackwell GPUs across eight buildings, with the first two buildings live and running workloads since September 2025. In March 2026, the partners scrapped a planned expansion beyond that 1.2 GW ceiling — a decision driven by financing terms and OpenAI's preference not to mix current-generation Blackwell hardware with the forthcoming Vera Rubin architecture at the same site.

Microsoft has since stepped in to lease the adjacent capacity Oracle and OpenAI walked away from, committing to build its own facilities and an on-site power plant at the same Abilene campus — turning the site into one shared by two competing cloud providers. The remaining six of Stargate's eight buildings, originally targeted for mid-2026, are now expected to reach full capacity around the turn of 2027.

Oracle has confirmed the two operating buildings publicly and reaffirmed that a larger, 4.5 GW multi-site agreement with OpenAI remains on track. Read plainly, Abilene is a case study in capital discipline with real friction attached: rather than build out additional Blackwell-generation capacity ahead of a site's readiness for next-generation chips, OpenAI and Oracle held that capital back — and a rival cloud provider filled the gap almost immediately.

1.2 GW STARGATE ABILENECAPPED BUILD | **4.5 GW** ORACLE-OPENAIMULTI-SITE AGREEMENT

2 BUILDINGS LIVESINCE SEP 2025

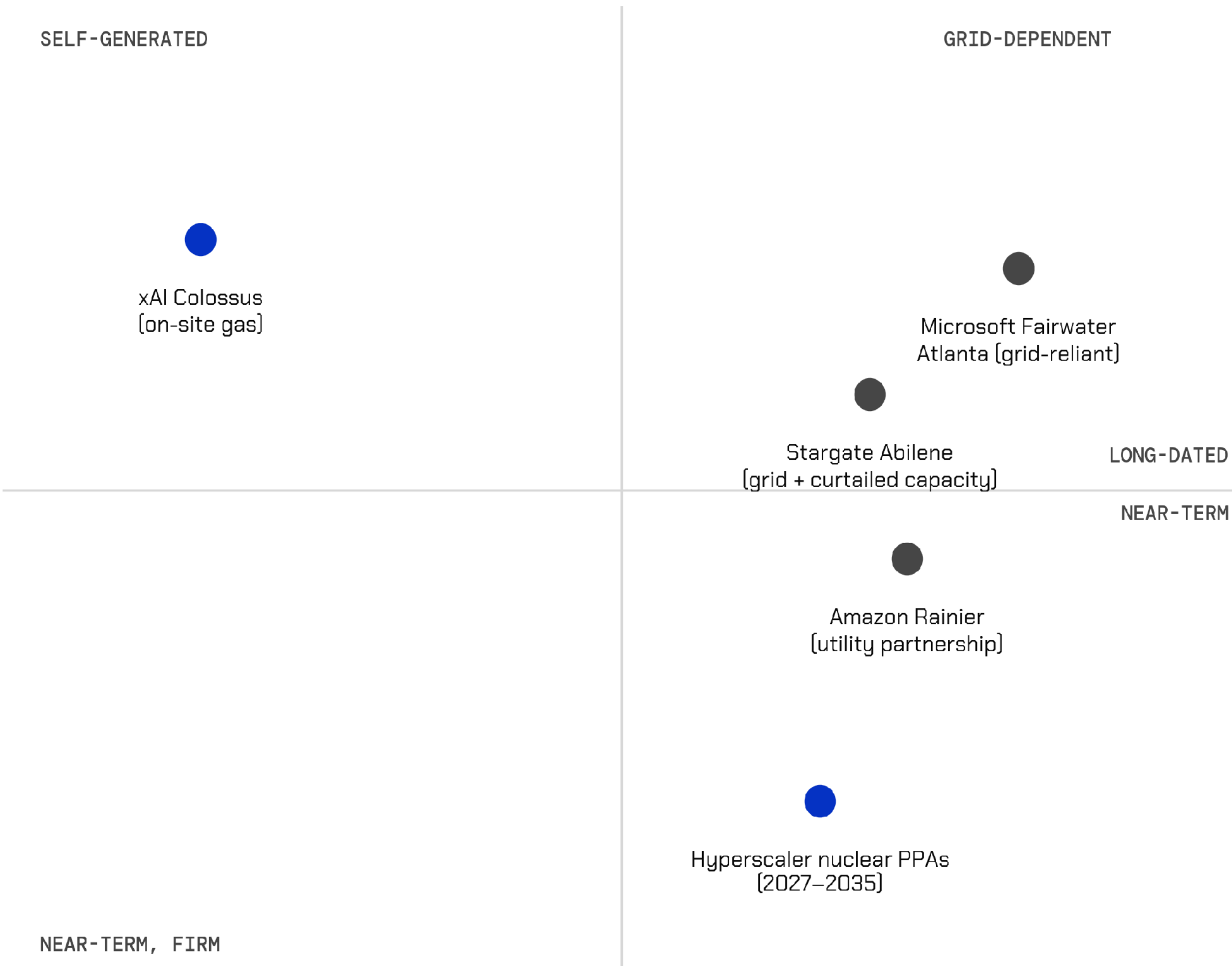
Project Rainier — Amazon's grid-partnership model

Amazon's $11 billion Project Rainier campus in New Carlisle, Indiana became operational in October 2025 and now trains Anthropic's Claude models on roughly 500,000 of Amazon's own Trainium2 chips. AWS chief executive Matt Garman was unambiguous about its status: "This is not some future project… This is running and training their models today." Seven buildings are operational within a campus designed to eventually house around 30, building toward 2.2 GW of capacity, powered through a dedicated partnership with the regional utility.

Anthropic has since committed more than $100 billion over ten years to AWS infrastructure, securing up to 5 GW of future capacity — evidence that a well-structured utility partnership, planned early enough, can deliver speed and scale comparable to full vertical integration.

Two axes of power strategy

How AI infrastructure operators are sourcing power — speed vs. durability



COMPUTE FORECAST

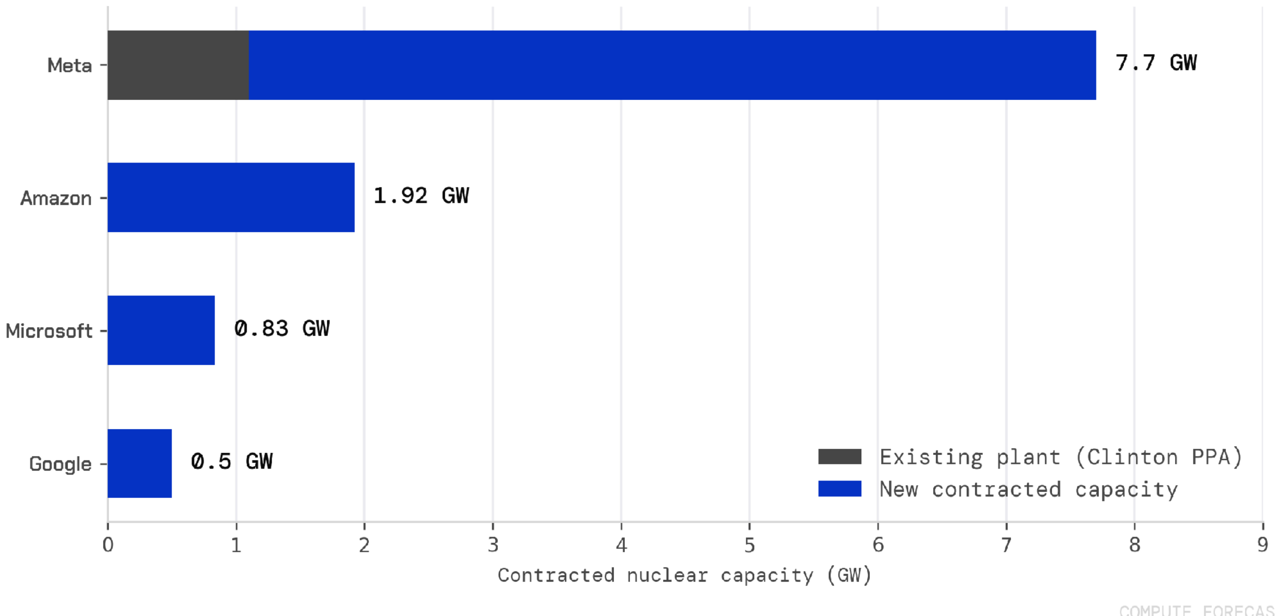
Nuclear power as the durability layer

Beneath these live buildouts sits a longer-dated strategy: all four major US hyperscalers have contracted nuclear capacity to underwrite AI compute growth through the next decade. Microsoft's 20-year agreement with Constellation will bring the restarted Three Mile Island Unit 1 — now the Crane Clean Energy Center — back online with 835 MW, targeted for 2027 or 2028. Amazon's deal with Talen Energy secures up to 1.92 GW from the Susquehanna plant, ramping to full capacity by 2032. Google has committed to 500 MW of small modular reactors with Kairos Power, targeting first delivery by 2030. And Meta has assembled the largest position of all — up to 6.6 GW across agreements with Vistra, TerraPower, and Oklo, on top of an existing 1.1 GW agreement for the Clinton Clean

Energy Center.

Meta leads the hyperscaler nuclear build-out

Contracted nuclear capacity by hyperscaler, in-service 2027–2035



CONTRACTED NUCLEAR CAPACITY BY HYPERSCALER, IN-SERVICE 2027-2035

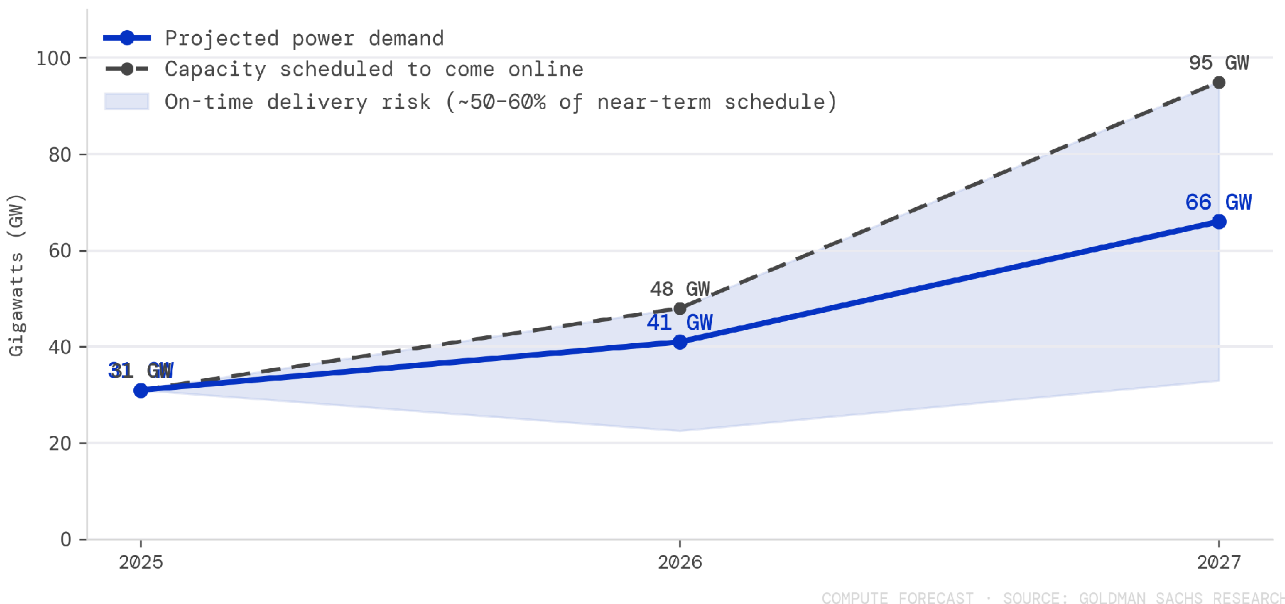
None of this nuclear capacity is powering training runs today — these are multi-year commitments designed to secure firm, carbon-free baseload for the back half of the decade. They function as the durability layer beneath the faster, nearer-term strategies described above: gas, grid partnerships, and fiber-linked networks solve for 2025 and 2026; nuclear solves for 2030 and beyond.

Why this is now a board-level decision

The market context explains why. Goldman Sachs Research projects that US data center power demand will climb from 31 GW in 2025 to 41 GW in 2026 and 66 GW in 2027, alongside data center capacity scaling toward roughly 95 GW by the end of 2027 — which would lift data centers to 8.5% of US peak summer power demand, up from 4.1% today.

Half of scheduled capacity may not land on time

US data center power demand vs. scheduled capacity, 2025–2027



US DATA CENTER POWER DEMAND VS. SCHEDULED CAPACITY, 2025-2027 · SOURCE: GOLDMAN SACHS RESEARCH

The number that matters most for capital planning is the delivery gap: Goldman Sachs Research estimates that only around 60% of capacity scheduled for the coming year will materialize on time, falling to roughly 50% over a two-year horizon. Reliability risk is concentrated in the Mid-Atlantic, Mid-Continent, and Northwest power markets, while Texas and Georgia are comparatively well positioned. When half of an industry's scheduled capacity is at risk of slipping, the operators who control their own energization timeline — through on-site generation, direct utility partnerships, or long-dated nuclear contracts — turn a shared industry constraint into a private advantage.

What this means for capital allocation

Five staged actions for boards and investors allocating capital into AI infrastructure through 2027–2030:

01 Make power procurement a board-level decision. The binding constraint on training throughput is megawatts energized on schedule, not GPUs allocated. Every large compute commitment should carry an explicit, dated energization plan before capital is released.

02 Underwrite on secured power, not announced power. Favor partners with on-site generation, direct utility agreements, or funded substations over those waiting on standard grid interconnection queues.

03 Build hardware-generation timing into contracts. Stargate Abilene shows that pausing a Blackwell-era expansion in favor of a Rubin-era deployment can be the more disciplined move. Structure capacity commitments with generation-transition flexibility built in.

04 Distinguish operating capacity from roadmap capacity in every diligence process. A live 200,000-GPU cluster and an announced 555,000-GPU target are different assets with different risk profiles — price them accordingly.

05 Treat nuclear as the long-term hedge, not the near-term solution. Nuclear PPAs deliver firm power from 2027 onward. Near-term compute needs are being met by gas, grid partnerships, and fiber-linked multi-site networks — plan capital across both horizons.

Compute Forecast tracks the infrastructure, capital, and power decisions shaping the AI build-out across data centers, neoclouds, and the global energy grid.
