Michael Olaya

@michaelaolaya

BRING YOUR OWN POWER

The End of the Grid-First Data Center
and Notes on the First 100 GW Supercluster

AUGUST 2026

"Any AI data center even thinking about coming here, they got to bring their own money, bring their own power, reuse their own water."

Gov. Greg Abbott, June 2026

All of this is based on publicly available information: manufacturer disclosures, earnings materials, regulatory filings, EIA data, and trade reporting, current as of August 2026. Figures are directional ranges and will move. Nothing herein is an offer, a price representation, or engineering advice. Comments and corrections: @michaelaolaya on X.

July 2026

Dallas, Texas

This paper makes two claims. **First, the market is underestimating long-term demand for compute, and power will therefore bind harder over the coming decade, not ease.** The market has begun to price the chips. It has not priced the power system required to run them. **Second, as a consequence, by the end of the decade no large AI data center cluster will energize grid-first.** Self-generation first, grid connection later, if at all, will be the default architecture of the buildout. The argument is arithmetic, not ideology: interconnection queues, transmission timelines, and a sold-out turbine supply chain have made time-to-power the binding constraint on time-to-revenue, and no utility can sell time. And where the arithmetic would permit grid service, the politics increasingly will not.

The analysis is grounded in Texas, the center of gravity of the American buildout, where market structure, resource base, and policy are converging on the same answer first. What follows maps the twelve generation options available to a cluster that must power itself, shows that air-permitting geography frequently selects the technology before economics does, demonstrates that the buildout is substantially a natural gas buildout across five machine architectures, and treats superclusters of 10 GW, and eventually 100 GW and beyond, as the base case of the buildout rather than a tail scenario, predicting how each will actually be powered: gas wins the decade, nuclear wins the scale, and the private use network becomes the grid. What Texas is being forced to invent, everyone else will adopt.

A definition, because the term spans several structures. Bring your own power, as used here, covers onsite generation behind the meter (on the customer's side of the utility meter), private use networks, co-located generation with a grid interconnection, dedicated new generation delivered through a utility, bridge power ahead of grid service, and bring-your-own-capacity arrangements deposited into a utility system. These differ legally, financially, and operationally, and the differences matter at the deal level. What unites them, and what this paper predicts, is the architecture: generation capacity is secured by or for the campus before, and independent of, ordinary utility service. Grid-first means the inverse: filing an interconnection request and waiting for the system to deliver. And large, throughout this paper, means clusters of hundreds of megawatts and above.

I. The Prediction

For a hundred years, powering a large facility meant one decision: call the utility. That era is over for AI data centers. Both claims above are falsifiable, and both resolve on public numbers within the decade: watch the load forecasts and watch what the flagship campuses actually energize on.

A. The demand miss

Every institution that forecasts electric load spent the fifteen years before 2023 calibrated to roughly zero growth, because that is what US demand did. Those institutions have now revised their data center projections upward in nearly every forecast cycle since, and the revisions keep arriving in the same direction. The pattern is not an accident of bad modeling; it is structural. Forecasts extrapolate today's workloads, but each capability gain expands the set of tasks worth automating, so demand compounds ahead of any extrapolation: the classic Jevons dynamic, in which efficiency gains grow total consumption rather than shrinking it. Training compute is the visible tip; inference, agents running continuously rather than answering queries, and the migration of ordinary economic work onto models are the mass below the waterline. Federal projections already contemplate AI doubling data center electricity demand by 2030, and those projections have themselves been serially revised upward.

If demand is underestimated, the supply gap is too. The interconnection queue for large loads already runs years past an 18 to 36 month campus construction schedule. The turbines that would relieve the queue are sold out: GE Vernova alone holds 116 GW of gas turbine backlog and slot reservations as of July 2026, up from 83 GW at the start of the year and expected to pass 125 GW under contract by year-end, and pricing headed toward $600/kW, nearly triple 2019 levels. Transmission moves slower than both. A gigawatt of AI compute represents tens of billions of dollars of chips that earn nothing until energized; time to power is time to revenue, and no queue position pays interest. When the binding constraint on a trillion-dollar buildout is a queue, capital routes around the queue. There is a name for what happens next, and it is one of the oldest patterns in industrial economics: when a critical input market cannot clear on the buyer's timeline, the buyer integrates backward into it. Ford built River Rouge; aluminum smelters bought dams; compute is now integrating into generation, because it scales like heavy industry, not like software. Under-forecast demand means the queue lengthens even as the buildout accelerates, which is why self-generation is not a bridge era but the new equilibrium: vertical integration does not un-happen when the shortage eases.

The economics deserve to be explicit, because they explain why expensive power is rational. Paying $20 to $50 per MWh above grid rates costs a gigawatt campus roughly $175 to $440 million a year. Letting the same gigawatt of chips sit dark for a year burns several billion dollars of a hardware fleet that depreciates on a roughly five-year clock whether it runs or not, before counting the revenue and the market position the compute would have earned. Speed-to-power dominates cost-of-power by an order of magnitude, which is why the premium for self-generation is not a cost of the strategy; it is the cheapest insurance in the capital stack. That is the whole prediction; the rest of this paper is its mechanics.

B. Why Texas is the lens

This paper is deliberately written from Texas, because the prediction resolves there first. Texas holds the largest project pipeline in the American buildout, the fastest-growing large-load queue in the country, and the announced home of the flagship campuses. Three structural facts make it the proving ground. The market: ERCOT is an energy-only market whose private use network (PUN) construct already gives a self-supplied campus a legal container, an electrical island that touches the grid on its own terms. The resource: the state sits on the gas belt, with the pipelines, the basins, and the stranded Permian molecules that make self-generation a logistics problem rather than a moonshot. The politics: in June 2026 the Governor said it plainly, that any AI data center even thinking about coming to Texas has to bring its own money, bring its own power, and reuse its own water. Not policy yet. But when the friendliest large-load market in America starts saying bring your own power out loud, the question is settled everywhere else too; other states will simply arrive at it with worse geology and slower paperwork.

C. The Texas numbers

The scale of what is being requested deserves numbers. ERCOT's all-time peak demand is 85,508 MW, set in August 2023 on a system serving about 90 percent of Texas load. In April 2026 the grid operator filed a preliminary long-term forecast projecting roughly 368,000 MW of demand by 2032, more than four times the record, driven by a large-load interconnection queue that entered 2026 above 233 GW, up nearly 300 percent in a single year, and stood near 474 GW by mid-2026, roughly 90 percent of it data centers. The funnel tells the real story: of all that requested load, only about 9 GW holds approval to energize and roughly 4 GW actually operates today. Against that, the supply side: about 23 GW of new generation was added across 2024 and 2025, and the roughly 460 GW generation queue is about 72 percent solar and storage, with 78 GW of gas. The load being requested is firm and around-the-clock; the generation being built mostly is not.

Is the forecast realistic? No, and ERCOT says so itself: its own summer 2026 operational projection runs 90,500 to 98,000 MW against the preliminary forecast's 112,000 MW for the same summer, and its leadership has told the PUC the long-term numbers likely overstate future load. The queue is inflated by design. A grid-first interconnection request costs little and commits less, so developers file the same project in multiple counties and speculators file projects that will never exist. The honest read is that the 368 GW headline is not a forecast; it is a measurement of how many free options the grid-first model hands out. ERCOT itself filed at the PUC in May 2026 to adjust the forecast.

Neither claim in this paper requires the headline number. Discount the queue by three quarters and the remainder is still roughly 60 GW of new firm load landing on an 85 GW system inside a decade, against 23 GW of recent additions that are mostly intermittent. And the delivered numbers, not the requested ones, are already the story: EIA measured ERCOT demand growing 7 percent in 2025 and projects 14 percent in 2026, on a system that grew roughly 1 percent a year for the prior fifteen. The queue inflation and the demand underestimate coexist because they are the same phenomenon seen from two sides: when interconnection is a free option, requests overstate any single site and the system undersupplies the whole. SB 6's answer is to make large-load requesters put capital at risk, and it went live in July 2026, when ERCOT's first batch study window opened with roughly 205 GW seeking eligibility under the new financial-security rules. Texas is repricing grid access toward the architecture this paper predicts. Bring your own power is what a queue looks like after it stops being free.

D. The political constraint

Everything above is arithmetic. There is a second forcing function arriving at the same architecture from the opposite direction, and it is political. A grid-served data center socializes its costs: transmission upgrades, capacity procurement, and reserve margins are recovered across every ratepayer on the system, so a household in Waxahachie helps pay for a hyperscaler's interconnection whether or not it ever touches the product. That arrangement was invisible while data centers were rounding errors. At gigawatt scale it is a visible transfer from the public to the wealthiest companies on earth, running through the one commodity every voter buys every month. The evidence is already on the tape: capacity prices in the largest US power market jumped nearly tenfold in a single auction cycle with data center growth cited in the record and have cleared at the regulator's price cap in every auction since, federal modeling attributes a large near-term ERCOT price increase to data center demand, electricity costs are a live issue in the 2026 elections, and local opposition has become a leading cause of project denial across the country. The settlement this paper describes is in fact already being codified: the White House now runs a formal Ratepayer Protection Pledge, and the first deal under it, the 10 GW Portsmouth campus in Ohio, requires the developer to build its own generation, pay for its own transmission, carry a dedicated rate structure so families do not fund the buildout, and supply the local grid in emergencies. Bring your own power is becoming the stated price of admission at the federal level, not just in Texas. Water reads worse still, because scarcity is local and visceral: a campus that draws down a town's aquifer has no defensible politics at all.

History is not subtle about what happens when a new and conspicuous class of consumption raises the price of household staples. The French Revolution began, in part, over the price of bread; nobody should want to discover what begins over the price of electricity. It will not matter that AI returns more value than it consumes. The transfer is visible, the benefit is diffuse, and the bill arrives monthly with the utility's logo on it. No economic-development argument survives that arithmetic in a county election. And note carefully that none of this requires the grievance to be true. Even if data centers turn out to raise no one's bill and drain no one's aquifer, the perception is sufficient, because a gleaming, fenced, windowless building consuming a city's worth of power is the most legible villain a populist will ever be handed. Rate cases are complicated; the campus on the hill is not. The industry can win the studies and lose every election, and the elections are what write the ordinances.

Bring your own power is therefore not only how the buildout goes fast; it is how the buildout keeps its license to operate. A campus that generates its own power, brings its own water, and sells surplus capacity into the grid inverts the politics: it stops being a claimant on the commons and becomes a contributor to it. That is exactly the settlement the Governor's formulation sketches: bring your own money, bring your own power, reuse your own water, and lower residential bills. Read as policy, it is a demand. Read as politics, it is an offer, the terms on which the public will let the AI buildout be enormous. Parallel systems are the price of scale, and the buildout should pay it gladly, because the alternative is becoming the face of everyone's utility bill.

Transaction practice already reflects the shift. Bring-your-own-capacity and bring-your-own-generation appear as defined structures in deals now closing; single sites co-developing generation and load at gigawatt scale (hybrid energy campuses) are a recognized project category; and bridge-to-grid, energize behind the meter first and ramp onto long-term supply later, is the standard sequencing in current development agreements. What follows is the part transaction practice has not yet standardized: a map of the machines.

The question is no longer whether to bring your own power. It is which machine, on which dirt, on what schedule. Section II maps the geography; Sections III and IV score twelve technologies on capital cost, efficiency, emissions, speed to power, and the constraint almost everyone underweights, which is where the law of the air will let you build at all. Sections V and VI follow the argument to its two endpoints: the molecule underneath nearly every near-term option, and what bring your own power means when the load is not 500 MW but 10 or 100 GW.

> *Two questions decide everything: when do you need power, and what county is your dirt in?*

II. The Map Decides First

Before any technology comparison matters, federal air law has already made half the decision. The Clean Air Act divides the country into attainment and nonattainment areas for ozone, and the two regimes are different planets for anyone trying to site combustion at scale. In Texas the boundary runs straight through the buildout: the state's two great metros sit in its strictest tier, while nearly everything else in the state is attainment.

A. Severe nonattainment: the 25-ton wall

Dallas-Fort Worth (ten counties: Dallas, Tarrant, Collin, Denton, Ellis, Johnson, Kaufman, Parker, Rockwall, Wise) and Houston-Galveston-Brazoria are classified severe ozone nonattainment under the 2008 national ambient air quality standard (NAAQS) for ozone. In a severe area, any source with potential to emit just 25 tons per year (tpy) of NOx is a major source. Crossing that line triggers Nonattainment New Source Review: LAER, the lowest achievable emission rate with cost explicitly not a defense, plus emission offsets purchased at a 1.3-to-1 ratio in a market where offsets barely trade.

Scale check: one large frame turbine running baseload, even with best-in-class selective catalytic reduction (SCR), emits on the order of 100+ tons of NOx per year. A gigawatt of reciprocating engines, several times that. There is no offset market deep enough. Combustion at scale inside the metroplex is not merely expensive; it is effectively unavailable on ordinary AI-campus timelines.

B. Attainment: the full menu

One county line over, the regime flips. Major-source thresholds rise 4 to 10x, review runs under Prevention of Significant Deterioration (PSD) with best available control technology (BACT) instead of LAER, and no offsets are required. Turbines and engine fleets permit on ordinary timelines. The same machine, moved across a county boundary, goes from impossible to routine.

> *The air basin picks your technology before you do. Every entry in the catalog that follows carries a siting line for exactly this reason.*

III. The Catalog: Twelve Ways to Make Your Own Electrons

Each entry reports capital cost, efficiency, emissions and siting, speed to power, and current field deployment. Figures are installed-cost ranges from public sources as of July 2026, directional by design; Table I summarizes, and Fig. 1 in Sec. VII collapses the catalog onto the two axes that decide. Three exclusions are deliberate: hydrogen, which is a fuel rather than a machine and is treated as such in Sec. V; combined heat and power, which is a multiplier on thermal generation rather than a source; and hydro and coal, which fail the behind-the-meter test on their face. One caution on how to read what follows: a campus does not buy megawatts in the abstract; it buys uptime. Every architecture below ultimately has to be translated into redundancy, maintainability, black-start capability, fuel-security planning, power quality, and failure-mode containment. Batteries, diesel, dual feeds, grid interconnection, and equipment diversity are not separate from the generation decision; they are how the generation decision becomes financeable.

1. Gas turbines

Capex: $1.3–2.2M/MW installed, simple cycle. Equipment pricing headed to ~$600/kW by end-2027, nearly 3x 2019.

Efficiency: 33–40% simple cycle; 60%+ combined cycle. Falls hard at partial load.

Emissions/siting: ~0.5–0.6 t CO_2/MWh. NOx 100+ tpy per baseload frame unit even with SCR; attainment counties only at scale.

Time to power: The binding constraint. ~100 GW under contract at GE Vernova; ~10 GW of open slots across 2029 and 2030 combined. No slot today means 2030+.

In the field: Crusoe's Abilene campus (Stargate). Homer City converting a retired coal site to 4+ GW of gas for data centers.

2. Reciprocating engines (RICE)

Capex: $1.0–1.6M/MW installed. Wärtsilä, Caterpillar, INNIO; 1–20 MW per unit, deployed as fleets.

Efficiency: 42–48%, and it holds at partial load where turbines fall off a cliff.

Emissions/siting: CO_2 similar to turbines; NOx higher per MWh even with SCR. A fleet at scale is a non-starter in severe nonattainment. One county over, standard PSD/BACT.

Time to power: 12–24 months. The fastest meaningful thermal option.

In the field: VoltaGrid alone operates GW-scale fleets on AI loads today.

3. **Microturbines**

   Capex: \$1.5–3M/MW installed. Capstone-class units, 30 kW to ~1 MW each, factory-packaged.

   Efficiency: 26–33% electric simple cycle, mid-30s recuperated; 80%+ total with heat recovery in CHP duty.

   Emissions/siting: Lean-premix combustion at single-digit-ppm NOx; permits where engines cannot, at the scales microturbines can reach. The constraint is the scale, not the air basin.

   Time to power: Months. Ships as an appliance.

   In the field: Roughly a gigawatt cumulative shipped worldwide over two decades, mostly oilfield and commercial CHP. Relevant to edge and sub-10 MW inference sites; not a hyperscale answer.

4. **Diesel gensets**

   Capex: \$0.5–0.9M/MW, the lowest capital cost of any option here.

   Efficiency: 35–45%.

   Emissions/siting: Worst on this list. Permits anywhere precisely because the permit caps runtime at a few hundred hours per year. Constant operation of backup fleets is the fight regulators are watching for.

   Time to power: Months. Nothing energizes faster.

   In the field: Tens of GW of standby capacity already behind US data centers. Some operators run Tier 4 fleets as bridge prime while permanent generation catches up.

5. **Fuel cells (solid oxide)**

   Capex: \$3–4M/MW before the federal investment tax credit (ITC).

   Efficiency: 50–60% electric. Best of anything gas-fed.

   Emissions/siting: Still emits CO2 (it runs on gas) but near-zero NOx: nothing burns. Hundreds of MW stays minor-source even in severe nonattainment; best positioned for the strictest air basins, though gas supply, fire code, noise, and local process still apply.

   Time to power: 12–18 months. Factory-built, ships on trucks.

   In the field: Bloom Energy has 1.5+ GW fielded worldwide and an agreement with AEP for up to 1 GW for AI loads. Equinix runs it across dozens of sites.

6. **Linear generators**

   Capex: ~$3M/MW class before ITC.

   Efficiency: Mid-40s%, flat across the load curve.

   Emissions/siting: Low-temperature flameless reaction; NOx passes California distributed-generation rules with zero aftertreatment. Same siting freedom as fuel cells; gas today, hydrogen or ammonia later on the same hardware.

   Time to power: 12–24 months, factory-built in ~250 kW blocks.

   In the field: Mainspring Energy shipping to utilities and data center campuses now.

7. **Nuclear fission (SMRs)**

   Capex: $8–15M/MW on first-of-a-kind projections; no delivered unit yet exists to anchor the range.

   Efficiency: The relevant spec is 24/7 carbon-free, firm, dense.

   Emissions/siting: Zero operating emissions; air basins do not care about SMRs. Everything else does: NRC licensing, emergency planning zones, HALEU fuel supply, community acceptance.

   Time to power: 2030 at the earliest for new units.

   In the field: Behind the meter today: 0 MW. Susquehanna and the TMI restart are grid-side PPAs, not new steel on a customer site. Amazon/X-energy and Oklo are the programs to watch.

8. **Fusion**

   Capex: Unknowable. Program capital runs in the billions per developer; no delivered $/MW exists because no MW has been delivered.

   Efficiency: The claimed spec is firm, zero-carbon, effectively unlimited fuel.

   Emissions/siting: No air permit issue, and a genuine regulatory advantage over fission: the NRC decided in 2023 to regulate fusion under its byproduct-materials framework rather than the reactor-licensing framework, which removes the emergency-planning-zone and Part 50 burden.

   Time to power: Net electricity delivered to any grid to date: zero. The most aggressive claimed pilots target the late 2020s; grid-meaningful capacity is a 2030s question at the earliest.

   In the field: 0 MW, but the offtake market has arrived ahead of the machines: Helion holds a power purchase agreement with Microsoft, Commonwealth Fusion has sited its first commercial plant, and TAE is selling directly into the data center market.

9. Geothermal

Capex: $4–6M/MW+ for enhanced geothermal.

Efficiency: Firm, clean, always-on.

Emissions/siting: Zero emissions, zero air permit. Conventional geothermal only works where the rock says yes. Enhanced geothermal (drill deep, fracture the rock, make your own reservoir) opens far more of the map, Texas included, but is unproven at scale.

Time to power: 3–5+ years of drilling and development.

In the field: Fervo's Cape Station in Utah, with Google as offtaker, is the proof point. Even that is a grid-side PPA, not behind the meter.

10. Solar

Capex: $0.9–1.1M/MWac. The cheapest MWh ever built.

Efficiency: 25–30% capacity factor. This is the entire catch.

Emissions/siting: Zero marginal emissions, no air permit anywhere. The constraint is land: 5–8 acres per MW, so hundreds of MW of primary solar means thousands of acres. A land-cost problem near metros, not a permitting problem.

Time to power: 12–18 months.

In the field: On nearly every hybrid campus as supplemental energy. Running a 24/7 load on solar alone requires 3–5x overbuild plus storage; nobody does it.

11. Wind

Capex: $1.3–1.7M/MW.

Efficiency: 35–45% capacity factor in West Texas.

Emissions/siting: Zero marginal emissions, no air permit, but the most siting-constrained renewable: resource quality, setbacks, FAA review, and footprint push it to West Texas and the Panhandle, which is rarely where fiber and load want to be.

Time to power: 2–4 years with siting.

In the field: Almost never literally behind the meter. Co-located campuses pair grid wind with flexible load rather than islanded supply.

12. Batteries (BESS)

Capex: $0.6–1.0M/MW for 4-hour lithium-ion, still falling.

Efficiency: 85–90% round trip. Emits whatever charged it.

Emissions/siting: No air emissions; goes anywhere, including the strictest nonattainment counties. Fire code and setbacks are the siting conversation, not the air basin.

Time to power: ~12 months.

In the field: 10+ GW on ERCOT already. Every credible campus includes it: bridging, ramping, power quality, turbine starts. Iron-air and flow chemistries are the next chapter.

IV. The Matrix

TABLE I. Behind-the-meter generation options for large loads, July 2026. "OK in severe nonattainment" means deployable at data-center scale (hundreds of MW), not a demonstration unit. SC = simple cycle; CC = combined cycle; CF = capacity factor; RT = round trip; FOAK = first of a kind; UC = under construction. Figures are indicative installed ranges for screening-level site selection, not EPC budgets.

Technology	Capex $M/MW	Efficiency	Time to power	OK in severe nonattainment?	Deployed BTM today
Gas turbines	1.3–2.2	33–40% SC / 60%+ CC	2030+ w/o slot	No	Multi-GW operating / UC
Recip engines	1.0–1.6	42–48%	12–24 mo	No	GW-scale fleets
Microturbines	1.5–3	26–33%	Months	Yes, at its scale	~1 GW cumulative, edge loads
Diesel gensets	0.5–0.9	35–45%	Months	Backup only (hour-capped)	Tens of GW standby
Fuel cells (SOFC)	3–4 pre-ITC	50–60%	12–18 mo	Yes	1.5+ GW cumulative
Linear generators	~3 pre-ITC	Mid-40s%	12–24 mo	Yes	Hundreds of MW, scaling
Nuclear (SMR)	8–15 FOAK	Firm, zero-C	2030+	Yes (air); NRC governs	0 MW
Fusion	n/a	Firm, zero-C	2030s	Yes	0 MW, offtakes signed
Geothermal	4–6+	Firm, zero-C	3–5+ yrs	Yes (where resource exists)	~0 (grid-side PPAs)
Solar	0.9–1.1	25–30% CF	12–18 mo	Yes	Supplemental everywhere
Wind	1.3–1.7	35–45% CF	2–4 yrs	Yes (siting-bound)	Rarely literal BTM
BESS (4-hr)	0.6–1.0	85–90% RT	~12 mo	Yes	10+ GW on ERCOT

V. The Load-Bearing Molecule

Strip the branding off the catalog and a pattern appears. Of the six technologies that can energize a large load before 2028, five run on natural gas: turbines, reciprocating engines, microturbines, fuel cells, and linear generators. The sixth, batteries, stores whatever the other five make. The behind-the-meter buildout is a natural gas buildout wearing five different machines.

A. Run the numbers

One gigawatt of mid-40s-efficiency gas generation, a heat rate near 7,500 Btu/kWh, running baseload burns roughly 0.18 Bcf/day. Scale that: a 10 GW campus is ~1.8 Bcf/day, the intake of a world-class LNG export terminal, for one customer. A 100 GW program is ~18 Bcf/day, roughly the entire current US LNG export complex, or about one-sixth of total US production. The United States is the only country on earth where those sentences describe an engineering problem rather than a fantasy, which is itself a strategic fact.

B. Molecules are the second queue

Everyone prices the electron queue. Fewer price the molecule queue. A greenfield lateral to a gigawatt-class site is a multi-year, tens-of-millions-of-dollars project with its own contested capacity: firm transport on existing pipes is finite, and interruptible service is a word data centers cannot say. Site assessment now starts with three questions, in exactly this order: which pipelines cross or border the tract, what firm capacity can be reserved and at what contribution-in-aid and reservation cost, and what happens to supply during a freeze event. Gas availability is not a detail of the power plan. It frequently is the power plan.

C. Basis is a site-selection signal

Gas is not one price. Permian gas at Waha has traded near zero and below it because production outruns takeaway. Every stranded-molecule basin is an invitation to move the load to the fuel: co-locating compute with constrained supply converts a producer's disposal problem into the cheapest firm power in the country. The fuel-cost gap between a 60% turbine and a 45% machine is worth about $5/MWh at normal gas prices and almost nothing at Waha prices, which quietly reorders the whole catalog: where gas is cheap, capex and speed to power dominate heat rate.

D. Bridge and destination

The polite framing is that gas is the bridge while clean firm power matures. The honest framing is that gas is the bridge and, for the 2020s, also the destination: nothing else combines dispatchability, speed, and fuel logistics at the required scale. The long-arc hedge is machine selection, not fuel abstinence. Fuel-flexible architectures, hydrogen-capable turbines, and linear generators that take hydrogen or ammonia on the same hardware, let the fleet decarbonize by changing the molecule later without changing the machine.

Whoever controls molecules, machines, and dirt controls the schedule.

VI. The Supercluster Question: How 10 GW and 100 GW Actually Get Powered

The frontier labs' public roadmaps, and essays like Situational Awareness that crystallized them, point at training clusters of 10 GW late this decade and national programs approaching 100 GW in the 2030s. This paper takes those numbers at face value and treats them as the planning case, not the tail: the concentration logic below makes superclusters the equilibrium shape of compute, and the demand miss of Section I argues their arrival is being underestimated along with everything else. The question that matters is not how big these are. It is what gets built to power them, and the answer is an architecture before it is a machine. The architecture climbs a ladder with the load. At hundreds of megawatts, bring your own power is a power plant behind a fence. At 10 GW it is a private utility, a generation portfolio with one customer. At 100 GW it is an energy system with compute as its anchor tenant: multiple campuses, its own transmission, its own fuel contracts, and the public grid as one counterparty among several. The machines are components of that architecture, not the other way around, and the machine prediction follows from it: gas wins the decade, nuclear wins the scale.

A. The case for concentration

First, the objection worth taking seriously: why concentrate at all? The distributed intuition says to sprinkle compute across hundreds of smaller sites where grid slack already exists, dodging the queue instead of routing around it. The math runs the other way, for three reasons. The first is training coherence. A frontier training run synchronizes gradients across every accelerator many times per second; inside one building that rides terabit links across meters, while across a state it rides fiber priced by distance and bounded by the speed of light. The tax compounds with model size. The frontier is defined by the largest coherent cluster; fragments cannot combine into it.

The second reason: it is cheaper to move molecules than gradients. Energy travels by pipe and wire at trivial marginal cost per mile; exabit-scale interconnect does not. The expensive input stays home and the cheap one comes to it. Compute concentrates at the energy. The third is fixed-cost arithmetic, the whole paper in miniature: one gas lateral, one 345 kV yard, one water system, one air permit, one entitlement fight, one operations team, amortized across 10 GW. Distribution does not escape the queues this paper describes; it multiplies them. Fifty sites is fifty interconnection requests, fifty permits, and fifty county elections.

The honest carve-out is latency-sensitive inference, which wants to sit near users. That layer sizes itself, and it is thin: the mass of demand, training, batch inference, and agents running around the clock, is latency-tolerant and migrates to the cheapest coherent megawatt. Even inference consolidates into regional hubs, because utilization rewards batching the way training rewards coherence. The edge is real and small. The center is where the power problem lives, which is why the supercluster is not a maximalist scenario but the equilibrium shape of compute under these cost curves: a city's worth of power inside one perimeter.

B. 10 GW: gas, and the mix is already decided

10 GW is on the order of New York City's peak electric load, delivered to one fence line. No interconnection process in the country is sized for it as a single request, and nobody builds a 10 GW power plant. They build a private utility: an independent power producer with one customer, with the ERCOT private use network as the closest existing blueprint. But the generation mix of the first 10 GW campuses is not an open question. Any campus energizing in 2028 through 2030 runs on machines ordered by 2026, and none of those machines are reactors: no SMR will deliver a commercial electron before 2030, and a 10 GW nuclear campus would require dozens to more than a hundred units of the designs now in licensing. The first 10 GW clusters will therefore be 80 to 90 percent gas by energy: combined-cycle blocks as the baseload spine, reciprocating fleets for ramping and redundancy, batteries for power quality and turbine starts, solar overbuild where the dirt allows trimming fuel burn, fuel cells or linear generators for any tranche sitting in hostile air. The announced flagship campuses that self-generate confirm it; every one of them is gas-led. The strongest case arrived in March 2026, wearing federal letterhead: the Department of Energy's Portsmouth partnership in Ohio pairs 10 GW of data centers with 10 GW of new generation, at least 9.2 GW of it natural gas, on leased federal land, with the developer funding $4.2 billion of transmission so ratepayers do not. That is a 92 percent gas, bring-your-own-everything campus, announced by the government, at exactly the unit size this section describes.

What 10 GW changes is not the technology but the buyer's position in the industrial system. 10 GW of turbines is roughly half a year of GE

Vernova's entire global production. 1.8 Bcf/day makes the campus an anchor shipper that underwrites new pipeline capacity rather than renting old capacity. Transformers carry multi-year queues of their own. At 10 GW you are not procuring equipment; you are reserving industrial output years forward, and the reservations themselves become the asset.

C. 100 GW: why gas hands off to nuclear

100 GW is not a data center with a power problem; it is an energy system whose primary customer is compute. It is roughly a fifth of average US electric load, arriving as a portfolio of 10-to-25 GW hybrid energy campuses. The lazy answer is that gas simply cannot scale to it. That is wrong, and the history matters: the United States built roughly 200 GW of gas capacity in about six years around the turn of the millennium. The manufacturing base expanded once and can expand again. If gas will not carry 100 GW alone, it is not because it physically cannot.

It is because of a problem the gas answer creates for itself. At 100 GW the buildout burns roughly 18 Bcf/day, a sixth of current US production, and demand at that scale is not a price taker: it bids up its own fuel. Every dollar per MMBtu the buildout adds to the gas market costs the entire gas fleet about $7.50/MWh at typical heat rates, forever, across every campus. The buildout eats its own operating margin, and the marginal gas megawatt gets worse as the program grows. Nuclear inverts every term of that equation: near-zero fuel-price exposure, no air-basin constraint, energy density that fits a campus, and, decisively, the only generation technology that manufactures the way compute does, in factory series with a learning curve. The economics of the handoff are mechanical: nuclear takes the marginal build at the point where a settled-design, Nth-of-a-kind SMR beats a gas megawatt whose fuel the buildout itself has made expensive. The prediction has named failure modes, and they should be watched rather than assumed away: it breaks if NRC licensing cadence does not accelerate, if factory learning curves fail to materialize the way they failed the last American nuclear buildout, if HALEU fuel supply stays constrained, or if communities refuse reactor campuses. The metric that settles it is delivered units, not announcements.

The financing side is the leading indicator, and it has already moved. Reactor series have historically failed for want of a counterparty large enough to underwrite the first units through to the cheap ones. Hyperscaler balance sheets are the first buyers in history big enough to do it, and they are doing it now: Amazon is funding X-energy toward a multi-gigawatt program, Google has contracted a Kairos reactor series, and Meta has run a gigawatt-scale nuclear procurement. Those offtakes are not electricity purchases; they are payments to stand up a production line. Gas carries the first half of the 2030s because it is the only thing that exists. Nuclear takes the growth share in the back half because at 100 GW it is the only thing that keeps working.

D. The geothermal question

Nuclear is not the only technology that clears the demand-side bar, and a paper that takes 100 GW seriously owes the strongest alternative a hearing. Enhanced geothermal shares every virtue that makes the scale handoff work: firm, zero-carbon, no fuel-price endogeneity, no air permit, and none of nuclear's licensing apparatus, because a geothermal well is regulated like a well, not like a reactor. And it holds the one card nuclear lacks: a learning curve that is already running on American iron. Fervo took the shale playbook, horizontal drilling and multistage completions, and applied it to hot rock; its drilling times have fallen from months toward weeks, 70 percent in four years, to 21 days on its deepest and hottest design as of mid-2026, with first commercial power at Cape Station scheduled for October 2026. The industrial base is not hypothetical. The rigs, crews, and services that drill the Permian redeploy to geothermal without retraining a workforce or rebuilding a supply chain, and Texas law treats the resource like oil and gas, which is to say favorably.

The constraint is thermodynamic, not industrial. Rock at accessible depth runs 150 to 200 degrees Celsius, and heat that cool converts to power at roughly 10 to 15 percent, which is why an enhanced-geothermal well delivers single-digit megawatts today. Run the arithmetic at campus scale: 10 GW is on the order of two thousand wells, a program the shale industry runs in a season, but the result is a wellfield, not a power block, with spacing set by thermal drawdown of the rock itself. Deliverable, and sprawling. Nuclear puts the same firm gigawatt on a fraction of the land with no reservoir to manage. The variable that flips the arithmetic is temperature: superhot drilling, 300 degrees and beyond, multiplies per-well output several-fold and closes most of the density gap, and it is exactly the kind of problem drilling learning curves have solved before.

Strip the catalog to the technologies with the demand-side physics to anchor a 100 GW system, firm output with no fuel-price feedback and no air-basin ceiling, and two remain. So the scale call stands, and it gains a hedge. Nuclear wins the scale if its failure modes hold;

geothermal is what the handoff falls back to if they fire, because it is the only other candidate with the right demand-side physics and a cost curve already bending. The two are different American industrial muscle groups: nuclear manufactures like compute, in factory series; geothermal drills like shale, in field iteration. A serious 100 GW program runs both and lets delivered cost per firm megawatt decide the mix. Watch the drilling curves the same way this paper watches reactor deliveries.

The ladder does not stop at 100 GW. Gas carries the 10 GW campus; fission and geothermal carry the 100 GW program; and fusion, the only entry in the catalog with effectively unlimited fuel, no price feedback at any scale, and reactor-free permitting, is the candidate for whatever comes after.* That is the correct reading of the fusion offtakes already signed: hyperscalers are not buying electricity a decade early, they are underwriting the option on the technology that makes the next order of magnitude feasible, the same first-unit financing they are extending to fission, applied one rung further out.

* A terawatt cluster is not absurd on this paper's own logic; it is what Claim One compounds into if the demand miss continues through the 2040s. One terawatt is roughly twice the average electric load of the United States, delivered to a program rather than a country. Served by gas it would burn roughly 180 Bcf/day, more than one and a half times total US production, which is to say it will not be served by gas; by fission, it is on the order of a thousand gigawatt-class reactors or their geothermal-field equivalent; by fusion, it is the one number in this paper at which no fuel supply blinks. And the logic that concentrates compute at the energy has one rung past even that: if terrestrial energy binds, the last stranded resource is the sun itself, and orbit is the ultimate attainment county, no air basin, no aquifer, no ratepayers. Delivered orbital compute today rounds to zero, and radiator mass and launch economics are the named failure modes; watch launched megawatts the way this paper watches reactor deliveries and drilling curves. Anyone underwriting the 2040s should hold both options.

And at 100 GW the behind-the-meter framing dissolves into something larger. The meter was the boundary between a customer and the grid. A hundred-gigawatt program is not behind anyone's meter; it is a second grid, purpose-built for compute, interoperating with the first. Bridge-to-grid runs in reverse: the utility interconnection becomes the backup, and the private use network becomes the network.

> *The superclusters will be won by whoever aggregated dirt, molecules, and machine reservations years before the load arrived. That work is being done now, mostly quietly, mostly in Texas.*

VII. The Decision

Collapse the catalog onto the two axes and the answer usually writes itself (Fig. 1).

WHEN DO YOU NEED POWER

	2026–27	**2028–29**	**2030+**
Severe nonattainment *DFW · Houston*	Fuel cells Linear gens · BESS	Scale the same fleet	SMRs Fusion, if delivered
Attainment *gas belt*	Recips · Microturbines Diesel bridge · BESS	Turbines, slot holders Solar hybrid · Wind	SMRs · Geothermal Fusion

FIG. 1. The decision matrix. The clock and the county select the machine before economics does. Placement is indicative; Table I carries the underlying figures, and diesel appears only as hour-capped bridge and backup.

A. Axis one: when do you need power?

2026–27: reciprocating engines, microturbines, fuel cells, linear generators, batteries, diesel bridge. These are the only machines that exist on that clock. **2028–29:** gas turbines, if and only if you already hold a production slot; solar additions to hybrid campuses. **2030+:** SMRs, geothermal, fusion if it delivers, and whatever your development agreements grew into.

B. Axis two: what county is your dirt in?

Attainment: the full menu. Turbines and engine fleets at scale, on ordinary permitting timelines. **Severe nonattainment:** combustion is off the table at scale; fuel cells and linear generators are the dispatchable options (microturbines below utility scale), and batteries and solar ride along freely. The county line is worth more than most technology analyses.

C. What follows from this

Three consequences fall out of the two axes. First, sites become the scarce asset. A tract with gas supply, air-permit headroom, water, fiber, and an interconnection position is worth a multiple of the identical dirt without them, because it compresses the only variable that cannot be bought later: time. Second, queues become tradeable value, and not just the electron queue: turbine slots, fuel-cell reservations, transformer delivery positions, and firm gas transport are all appreciating assets in exactly the way land options are. Third, the technologies best positioned for strict air basins (fuel cells, linear generators, batteries) capture the metros, while combustion consolidates in attainment counties along the gas belt, and the two geographies develop different price signatures. The supercluster era only amplifies all three: at 10 GW the campus is a utility, at 100 GW the private use network is the grid, and in both cases the winners assembled molecules, machines, and dirt before the load showed up. And because demand is under-forecast, all three consequences intensify with time rather than decay: every upward revision to the load forecast lengthens the queues, deepens the equipment shortage, and raises the value of a site that can power itself.

Step back and the whole argument is one reversal. For a hundred years, building something big meant calling the utility; the factory connected to the grid. The AI buildout runs that sentence backward. Compute secures its power first, builds its own network, and treats the public grid as one interconnection among several: useful, optional, and no longer in the critical path. The private use network becomes the grid. That is the architecture every section of this paper converges on, from the county line to the supercluster.

VIII. Notes on the First 100 GW Supercluster

Follow the ladder to its top rung and the energy analysis runs out before the argument does. An energy system with compute as its anchor tenant still sits inside somebody's jurisdiction: county permits, municipal water authorities, road districts, appraisal rolls, school boards, and the ordinary police powers of whichever locality the fence happens to cross. Section VI counted fifty county elections as the price of distribution; perfect concentration still leaves the program governed by governments that never planned for it. Governance is the last queue, and it is the only one this paper has not yet routed around. The conclusion sits one step past the meter inversion: the first 100 GW cluster will not be built in a jurisdiction. It will be one.

A. Why a hundred gigawatts wants one address

Section VI established that 100 GW arrives as a portfolio of 10-to-25 GW campuses. The question this section answers is why the portfolio should share an address, because the alternative, a dozen campuses scattered across the gas belt, looks safer and is not. The first reason is the machine itself. The concentration logic of Section VI does not stop at the fence line: a jurisdiction whose diameter light crosses in under a millisecond can still train as one machine, with campuses far enough apart to contain fire, weather, and failure, and close enough that the program behaves as a single cluster. The jurisdiction is the largest geography that is still a computer.

The second reason is the oldest one in utility economics: reserve sharing. Ten isolated islands each carry their own spare machines, their own maintenance margins, their own black-start capability; one interconnected system pools them. At program scale the difference is measured in gigawatts of generation that never has to be built. The same pooling logic runs through fuel (one storage hub instead of ten stranded laterals), water (one reuse loop), operations (one control room, one trained workforce), and politics (one settlement instead of a dozen county-by-county negotiations that each reset the precedent). And the program-scale load does something no single campus can: it becomes an asset to the public grid. A system that can shed ten percent of its load on command is a 10 GW peaker ERCOT never had to build, which converts the political offer of Section I from a promise into hardware. The campus that would have been the face of everyone's utility bill becomes the reason the lights stayed on.

B. The system

Now the spec, stated plainly, because the numbers are the argument. A 100 GW private system is larger than the peak demand of all but a handful of nations. It does not have a system operator; it is one. The functions every national grid takes for granted get built inside the fence: frequency regulation on an islanded system dominated by large thermal units, grid-forming battery fleets for stability and turbine starts, pooled operating reserves, staggered maintenance seasons, black start, and a control room that dispatches generation, compute, and the grid interface as one optimization.

The physical layers follow. Fuel: half a dozen or more large-diameter laterals from at least two basins, contracted as an anchor shipper, plus on-system storage in the salt formations Texas conveniently has, sized to ride through a Uri-length freeze, the February 2021 standard every

Texas power plan is now measured against, without shedding a training run. Electricity: a private high-voltage ring connecting generation campuses to compute campuses, with the public interconnections hung off it as market interfaces rather than lifelines. Heat: 100 GW of electricity in is roughly 100 GW of heat out, which makes heat rejection a first-order land and water input and makes the closed-loop water commitment an engineering requirement as much as a political one. Fiber: more long-haul capacity between campuses than most countries own, because the gradient traffic of Section VI never leaves the district. And the city: construction labor in the tens of thousands at peak, a permanent operating population that with families runs to six figures, and the schools, hospitals, and training academies that keep it. The jurisdiction is not an accessory to the energy system. It is a component of it.

C. The machinery and the sequence

A jurisdiction of one's own sounds radical and is not, least of all in Texas, where the machinery already exists at retail scale. Municipal utility districts and public improvement districts, the workhorse financing tools of master-planned development, fund water, roads, and utilities against future assessed value every day. General-law incorporation lets a workforce settlement become a city; SpaceX incorporated Starbase in 2025 with a few hundred voters. The master-planned community is the quiet precedent for the whole program: assemble the land, secure the entitlements, form the districts, deliver in phases against contracted takedowns. A supercluster is that model at industrial scale, with substations where the amenity centers would be.

The sequence is the master-planned playbook at industrial scale. Assemble the contiguous position in attainment gas country where pipelines cross or border the tract. Form the districts and pass the charter, the political settlement of Section I written down: surplus power exported, water closed-loop, the neighbors' bills lower with the program than without it. Master-plan the hundred and energize the first ten: Phase 1 is a 10 GW campus, the unit Portsmouth just proved financeable, and the jurisdiction is what lets the next nine repeat it without renegotiating the world each time. Energize that first phase on the 2026–27 machines, reciprocating fleets, batteries, and fuel cells, while the turbine slots, reactor programs, and wellfields mature behind them, the ladder of Fig. 1 climbed in place. Finance each phase the way master-planned communities finance streets, against the assessed value the previous phase created, and deliver capacity against contracted takedowns the way lots are delivered to builders. And build the city deliberately, because a decade of construction at this scale is a city whether anyone plans it or not, and the ones that are planned keep their license to operate. Nothing in the sequence is novel except the load.

Call it Atlas: a special-jurisdiction supercluster city, sized to hold up the compute economy the way its namesake held the sky. Nothing in it requires new physics or new law, only the assembly, years early, of dirt, molecules, machines, districts, and consent. Like every piece of infrastructure that ends up looking inevitable, it will belong to whoever read the map first. And the tell, when it comes, will be mundane: district formation filings and quiet land assembly in attainment gas country, years before the first chip arrives.

> *Both claims, and the supercluster call that follows from them, are public and gradeable. Watch the load forecasts, watch the reactor delivery schedules and the drilling curves, and watch what every flagship campus announced from here forward actually does in its first 24 months: it will energize on its own generation. The grid is not coming to save anyone's energization date, and the voters are not going to underwrite it. Bring your own power is not a constraint. It is the design, and Texas is drawing it first.*

NOTES ON FIGURES AND SOURCES

All capital-cost, efficiency, fuel, and deployment figures are directional ranges assembled from public sources (manufacturer disclosures, earnings materials, regulatory filings, EIA data, and trade reporting) as of July 2026, and will move. Demand outlook: IEA and EIA data-center electricity projections, both revised upward across recent cycles. Turbine backlog and pricing: GE Vernova Q2 2026 disclosures (gas equipment backlog and slot reservations at 116 GW, at least 125 GW expected under contract by year-end, output ramping to 20 GW annually); Wood Mackenzie on turbine price inflation. Air-permitting geography: EPA Green Book and TCEQ classifications. The DFW and HGB areas are severe nonattainment under the 2008 ozone NAAQS, which carries the binding requirements used here (25 tpy NOx major-source threshold, LAER, 1.3:1 offsets); the DFW ten-county list is the 2008-standard area, and the separate 2015-standard designation covers a smaller area at a lower classification. Gas arithmetic assumes ~7,500 Btu/kWh heat rates: 0.18 Bcf/day per baseload GW; ~18 Bcf/day per 100 GW, against US production of roughly 105+ Bcf/day and LNG exports in the mid-teens Bcf/day. Supercluster scale references draw on public frontier-lab disclosures and L. Aschenbrenner, Situational Awareness (2024). The circa-2000 gas buildout figure (roughly 200 GW of capacity additions over about six years) is from EIA generator-additions data. Hyperscaler nuclear commitments (Amazon and X-energy, Google and Kairos Power, Meta's nuclear procurement) are from company announcements. Gas-led flagship self-generation campuses referenced in Sec. VI include Abilene (Stargate), the xAI Memphis buildout, Homer City, the announced Louisiana hyperscale campus, and the DOE Portsmouth partnership

(SB Energy, Pike County, Ohio; DOE fact sheet, March 2026: 10 GW of data centers, 9.2 GW of new gas, $4.2 billion of developer-funded transmission, federal land lease, Ratepayer Protection Pledge); grid-side nuclear PPAs at existing plants (Susquehanna, Three Mile Island) are categorized separately per Sec. III. Texas load data: ERCOT's preliminary Long-Term Load Forecast for 2026–2032 (April 2026, PUCT Project 58777), ERCOT board and system-planning reports (December 2025) on the large-load interconnection queue, the fastest-growing reported by any US market operator, and EIA Short-Term Energy Outlook ERCOT growth figures. Political-constraint evidence: PJM Base Residual Auction clearing prices from the 2024/25 through 2028/29 delivery years (roughly $29 to $270 to $329 per MW-day across consecutive auctions, then two straight years pinned at the FERC price cap, $333 and $325, with PJM's own simulation showing an uncapped 2028/29 price above $550 and a special reliability backstop auction called for September 2026, large-load growth cited throughout the record); EIA Short-Term Energy Outlook modeling of data-center-driven ERCOT price impacts (2026); and municipal records and press reporting on data center moratoria and project denials. The Starbase municipal precedent is the May 2025 Cameron County incorporation election. Geothermal scale figures draw on Fervo Energy drilling-performance and Cape Station disclosures (July 2026: ninth 3.0-design well drilled in 21 days; first 100 MW scheduled October 2026; NASDAQ listing 2026), DOE Enhanced Geothermal Shot materials, and standard binary-cycle heat-to-power efficiencies at stated resource temperatures. Fusion regulatory treatment reflects the NRC's 2023 decision to license fusion under the byproduct-materials framework. Nothing herein is an offer, a price representation, or engineering advice; consult qualified air-quality, gas-supply, and engineering counsel before siting decisions.