

The AI Data Center

Architectures, Money, Power, Water & Communities

Research cutoff: October 20, 2026.
U.S. focus. Global forecasts are clearly labeled; development, architecture, power, water and community evidence use distinct measures.

THE CENTRAL QUESTION

What is an AI data center? What’s inside, how do hyperscalers and enterprises use them, where do electricity and water come into play, and what are the impacts on local communities?

CONTENTS

Executive Summary	2
Scale and Units	3
Part I — What Is Being Built?	4
Part II — Who Builds and Uses It?	7
Part III — What Happens Outside?	12
Outlook and Methodology	15–17

Executive Summary

Artificial intelligence is driving a major expansion of America's physical infrastructure: data centers, specialized chips, electricity, water and land. This report examines what these facilities look like inside, who builds and uses them, and their economic, environmental and community impacts.

Finding 1 — Density is redefining the facility. Rack power rises from roughly 5–15 kW in conventional rooms to about 135 kW in deployed GB300 racks, with 170+ kW planned, turning AI compute into an electrical-engineering problem. [6,17]

Finding 2 — Expansion is substantial. CBRE counted 10,903 MW operating and 7,481.1 MW under construction across eight primary North American markets in H1 2026, 80.4% preleased. [1]

Finding 3 — Electricity is the binding constraint. EPRI estimates data centers used 4–5% of U.S. electricity in 2024 and models 9–17% in 2030, depending on which projects reach operation. [2]

Finding 4 — New AI loads are driving dispatchable generation. Under reference policies, EPRI sees gas leading near-term incremental supply; Hyperion plans seven gas plants and Abilene has gas-backup turbines. [2,7,8]

Finding 5 — Water accounting is incomplete. On-site cooling is only part of the footprint; power plants serving the grid also withdraw and consume water, and facility-level metered data is rarely disclosed. [4]

Finding 6 — Execution and social license determine returns. Energy connections, permitting, construction financing, rate design and opposition can matter as much as land costs or announced tenant commitments. [1,5]

Finding 7 — Enterprise AI grows beyond mega-campuses. A large share of enterprise AI demand is landing in colocation and cloud capacity rather than in enterprise-owned campuses or corporate server rooms, because most corporate data rooms lack the power and cooling for dense GPU systems. This distributed growth also carries grid and community impacts. [19,23]

Implications for Three Audiences

Investors and suppliers	Utilities and regulators	Host communities
Watch energized MW, credit quality, generation and commissioning timelines.	Require deposits, cost-allocation terms and realistic demand forecasts.	Demand project-specific power, water, jobs, tax and emergency-service reporting.

Bottom line: The most useful leading indicator is no longer an announced campus or dollar amount. It is the intersection of funded project + obtainable electricity + credible offtake + local approval.

Scale and Units: Making Electricity Intuitive

Unit	Technical meaning	Practical comparison
1 MW	1,000 kW	About 800 average U.S. households on an annual-energy-equivalent basis.
1 GW	1,000 MW	Comparable to one large nuclear generating unit; ~800,000 average homes by annual energy equivalent.
1 TW	1,000 GW	1 trillion watts; national-grid-scale instantaneous power.
1 TWh	1 billion kWh	1 GW delivered for 1,000 hours.
8.76 TWh	1 GW × 8,760 hours	A constant 1 GW draw for a non-leap year.

A gigawatt (GW) is a rate of power, like how fast water flows from a pipe. A terawatt-hour (TWh) is total electricity over time, like the number of gallons delivered. Do not confuse a terawatt (TW) with a terawatt-hour (TWh). [3]

Illustrative 1-GW IT campus: At 85% average IT utilization and 1.20 power usage effectiveness (PUE), average facility demand is 1.02 GW; annual consumption is approximately 8.94 TWh. This is a calculation, not a measurement of Prometheus or Hyperion.

Capacity definitions vary: announced campus MW, contracted electrical MW, grid-interconnection MW, commissioned IT MW and measured peak utility draw may all differ. Ratings should never be summed unless their boundary and lifecycle stage match.

U.S. Data Center Electricity Demand: 2024–2030

Scenario	TWh	Share of U.S. Electricity
2024 low estimate	177	~4%
2024 high estimate	192	~5%
2030 low scenario	380	~9%
2030 high scenario	790	~17%

EPRI estimates approximately 177–192 TWh in 2024 and 380–790 TWh in 2030 across scenarios. Includes cryptocurrency mining in its data center scope. [2]

Inside an AI Data Center

An AI data center is a computing facility optimized for training and running AI models, in which accelerators, high-bandwidth networks, power delivery and cooling are designed as one system whose requirements change sharply with scale. Customer size and physical deployment size are different. A small firm may use thousands of GPUs in public cloud, while a large enterprise may run modest local systems. This section compares illustrative deployment architectures—not universal customer tiers. The progression from server to rack to campus raises the stakes for networking, cooling and electricity.

How the Architecture Evolves

ANATOMY OF AN AI DATA CENTER



Schematic, not to scale. Facility arrangements and responsibility boundaries vary.

Deployment	Compute and fabric	Facility and cooling	Outside impact
Individual server / small room	CPU plus optional 1–8 PCIe GPUs; 1/10/25 GbE	~5–15 kW/rack, air cooling; existing server room	Usually indirect cloud demand; limited local upgrade
Multi-rack / mid-size cluster	PCIe GPUs, 8-GPU servers; 25–100 GbE	~15–40 kW/rack; airflow design, some rear-door cooling	Electrical headroom and retrofit constraints
Dedicated AI hall / enterprise-scale	8-GPU systems, storage and inference fleets; 400–800G fabrics	~40–130 kW dense racks; liquid cooling increasingly useful	Colocation supply, larger utility connections, water choices
Campus / rack-scale AI	GB300 NVL72 / AMD Helios-class racks; high-speed scale-up/out	~135 kW deployed GB300; 170+ kW next-gen goal; liquid cooled	Grid interconnect, generation, water, financing, local approval

These are illustrative deployment patterns, not averages or customer-size thresholds. Rack and product figures are date-specific vendor disclosures. [6,17,18,20]

Prometheus in practice. Meta reports its Ohio supercluster’s first phase operating and training Muse models, with a target of approximately one million interconnected GPUs. GB300 systems draw about 135 kW per rack; Meta reports 30–40% load swings in seconds, illustrating why large AI architecture becomes an electrical-engineering challenge. These are company-reported specifications, not site-level metered power or water use. [6]

Chips, Systems and Networking: The Product Landscape

Compute procurement progresses from PCIe GPUs to eight-GPU systems and rack-scale fabrics, with custom cloud accelerators at the largest operators. Representative products as of October 2026 include Nvidia GB300 NVL72, Grace and the planned Vera Rubin; AMD Instinct MI355X and MI455X/Helios; Intel Xeon 6, Gaudi 3 and the announced Crescent Island; and Google TPU, AWS Trainium and Meta MTIA. Planned products are not treated as deployed systems, and the tables show representative products, not market shares. [17,18,20]

Product layer	Individual server	Multi-rack cluster	Dedicated AI hall	Rack-scale campus
Nvidia	GPU: entry PCIe and workstation-class cards in small servers	GPU: PCIe data center cards (L4, L40S, RTX Pro server)	GPU: HGX 8-GPU systems (Blackwell). CPU: Grace (Arm) in some configurations	Rack-scale: GB300 NVL72, Vera Rubin. CPU: Grace, Vera (Arm). NVLink, InfiniBand, Spectrum-X
AMD	CPU: entry EPYC servers. GPU: Radeon Pro	CPU: EPYC. GPU: Radeon Pro; Instinct PCIe options	CPU: EPYC. GPU: Instinct MI355X 8-GPU systems. Software: ROCm	Rack-scale: Helios with MI455X [20] and EPYC. Pensando NICs; UALink and Ethernet scale-up
Intel	CPU: entry Xeon servers. GPU: Arc Pro	CPU: Xeon 6 (large installed base)	CPU: Xeon 6 and 6+. Accelerators: Gaudi 3 (niche); Crescent Island inference GPU announced	CPU: Xeon as head node; custom x86 with NVLink via Nvidia partnership
Other and custom	Apple Mac Studio and Mac mini clusters (niche)	Few alternatives	Inference ASICs in niches (e.g., Qualcomm, d-Matrix)	Custom: Google TPU, AWS Trainium, Meta MTIA, Microsoft Maia. Arm CPUs: Graviton, Axion, Cobalt
Memory	DDR5; GDDR on entry GPUs	DDR5 plus GDDR on PCIe GPUs	DDR5 plus HBM3e; CXL expansion emerging	HBM3e moving to HBM4 (432 GB per MI455X) [20]; supply from SK hynix, Samsung, Micron
Network silicon	Commodity Ethernet	Broadcom, Intel, Nvidia NICs	Broadcom, Nvidia, AMD Pensando, Marvell	Broadcom, Nvidia, Marvell, AMD Pensando; custom fabrics [17]

Representative products, not complete lineups. Roadmap items (e.g., Intel Crescent Island, Nvidia Vera Rubin) may not be generally available. MI455X specifications per AMD. [20]

System vendors. The system brand matters less as buyers scale. Enterprises buy through Dell, HPE and Lenovo; hyperscalers buy from contract manufacturers and design racks directly with chip vendors. Vendors shown are representative; most sell into more than one tier.

Product layer	Individual server	Multi-rack cluster	Dedicated AI hall	Rack-scale campus
Server and system vendors	Dell, HP, Lenovo, Apple	Dell, HPE, Lenovo, Supermicro, Cisco	Dell, HPE, Lenovo, Supermicro, Nvidia DGX; packaged "AI factory" offerings	ODMs (Quanta, Wiyynn, Foxconn, Inventec); Supermicro, Dell, HPE for neoclouds (GPU-cloud specialists)
Network vendors	Small-business switching	Cisco, Arista, HPE Aruba, Juniper	Arista, Cisco, Nvidia Spectrum-X	Arista, Nvidia, white-box on Broadcom, in-house designs [17]
Design model	Off the shelf	Validated OEM configurations	OEM reference architectures; ISV-certified	OCP and custom designs; chip vendor often co-designs the rack

Four Transitions That Change the Economics

Air to liquid, as density rises. Around 30–50 kW per rack, traditional air cooling becomes increasingly difficult or expensive depending on hardware and facility design. Liquid cooling is increasingly favored for dense accelerators, but there is no universal threshold.

Server to rack. Smaller buyers purchase servers. Hyperscalers purchase racks such as Nvidia NVL72 or AMD Helios, where the rack functions as a single computer and the network becomes its backplane. [17]

IT problem to utility problem. At the bottom of the market, power is usually an IT budget line. At the top it becomes a grid interconnection, a generation contract and a community issue. [2,7]

Merchant to custom silicon. Only the largest operators can justify designing their own chips; everyone else buys from Nvidia, AMD or Intel.

Where advantage shifts. At small scale, purchasing and integration simplicity matter; at rack scale, accelerator software, memory bandwidth, interconnects, cooling and power delivery must be optimized together. Large operators can justify custom silicon when workload volume and utilization offset design costs. Scarce energized capacity and electrical equipment may have more durable pricing power than hardware-dependent compute rentals, but neither outcome is guaranteed. (Analyst interpretation.)

Implication: Large dedicated AI campuses create highly concentrated demands for electricity and cooling, while much enterprise adoption occurs through existing facilities and colocation. Enterprise AI growth and gigawatt campus construction should not be treated as equivalent. [19]

Enterprise AI Demand

AI demand reaches physical infrastructure through different channels: owned server rooms, private and public cloud, managed AI infrastructure and colocation. Most enterprise AI customers do not build their own campuses. Industry spending forecasts must not be interpreted as deployment-location forecasts.

AI Spending: Training, Inference and the Infrastructure Between

Gartner forecasts global AI infrastructure spending of about \$1.48 trillion in 2026, roughly 56% of all AI spending, and notes that spending to date has been driven primarily by technology companies and hyperscalers. Within AI-optimized cloud infrastructure, inference spending overtakes training in 2026. Gartner attributes the shift to agentic AI, whose multistep, autonomous execution multiplies compute per task. [21,22,24]

Gartner AI-optimized IaaS (global)	2026
Inference spending	\$23.3 billion
Training spending	\$19.0 billion
Total category spending	About \$42 billion (up 96%)
Inference share of category	55% in 2026; 59% in 2027

Global estimate (evidence grade C), limited to AI-optimized infrastructure as a service; excludes on-premises hardware purchases. [21]

IDC reports growing use of non-GPU AI infrastructure and a change in the value share of accelerated-server CPU platforms. Orchestration, data pipelines and CPU-only inference are plausible contributors (analyst interpretation, not IDC findings). IDC reports Arm overtaking x86 by value within accelerated servers in Q1 2026, not across all servers. [18]

Training versus inference — a hardware mix, not a binary choice. Frontier-model training repeatedly updates model weights across large accelerator clusters, prioritizing GPU/TPU throughput, high-bandwidth memory and fast interconnects; CPUs handle data preparation, scheduling and control. Production inference runs an already trained model: large language models still often require GPUs/TPUs for token generation, while CPUs operate request routing, retrieval, application logic, tool execution and smaller models. Latency, throughput, memory capacity and cost per response determine the mix. Therefore, inference growth does not imply a fixed CPU-to-GPU ratio or a one-for-one increase in hyperscale training campuses. This is architectural synthesis, not an IDC estimate. [21,25]

Private and Hybrid, and Increasingly Colocation

Surveys point the same direction on data control but differ on location. In a Cloudian-commissioned survey of 203 enterprise IT decision-makers, 91% said they would choose on-premises, private cloud or hybrid infrastructure over public cloud for AI involving sensitive data, and 40% reported cloud AI spending above initial projections. [23]

Yet “private” increasingly does not mean the corporate building. A Foundry survey backed by colocation provider CoreSite found the share of IT leaders citing on-premises as most suitable for AI fell from 46% in 2025 to 18% in 2026. The same review reports rising power and cooling costs (40%), scaling to meet power and cooling demand (37%) and talent shortages (36%) among operators’ leading concerns. [19]

Analyst interpretation. The surveys are reconcilable: enterprises want control of AI data and costs, but most corporate data rooms cannot host dense GPU racks. Colocation resolves both, which places a large share of enterprise AI demand in established multi-tenant markets rather than in gigawatt campuses. This is the first of the two market types described under U.S. Data Center Capacity by State.

Source caution: Cloudian sells on-premises storage and CoreSite sells colocation; both surveys are vendor-sponsored (grade C) and indicate direction, not precise shares. Populations, regions and wording differ.

How Enterprise Workloads Are Changing

From pilots to production. Generative AI, chatbots and agents are moving into production, though Gartner sees buyers favoring tactical productivity gains over disruptive change. [19,24]

From chatbots to agents. An agent can issue many model calls, tool calls and data queries per task, multiplying inference demand and the conventional compute around it.

From one model to many. Enterprises increasingly run several smaller, task-specific models rather than a single frontier model, which fits mid-range hardware and existing facilities.

An execution gap remains. Reliability, governance and data readiness, rather than compute availability, often determine whether pilots reach durable production.

Patterns by Vertical

Vertical	Common production workloads	Infrastructure implication	Where compute tends to sit
Financial services	Fraud detection, KYC and AML review, claims, client servicing, back-office agents	Regulated data; latency-sensitive trading and fraud systems; large existing CPU estates	Private cloud and colocation, often near exchanges and existing data centers
Healthcare and life sciences	Prior authorization, clinical documentation, imaging; drug discovery and genomics	Patient-data privacy and residency requirements; research workloads resemble HPC	Private and hybrid; research clusters and HPC centers
Manufacturing	Predictive maintenance, visual inspection, supply chain, engineering design	Real-time inspection needs plant-level edge compute; simulation and digital twins emerging	Edge at the plant plus central data center or cloud
Retail and consumer goods	Customer service, demand forecasting, personalization, pricing, inventory	High-volume, cost-sensitive inference; small infrastructure teams	Public cloud plus modest on-premises or colocation footprints
Telecom	Network operations, customer automation	Distributed, edge-heavy, Ethernet-native operations	Network edge and regional data centers
Public sector	Document processing, citizen services, defense and research	Data residency and supply-chain mandates; slower procurement	On-premises and sovereign or government cloud

Analyst interpretation (grade D); illustrative. Adoption rates vary widely across surveys and are omitted.

U.S. Data Center Capacity by State and Electricity Generation Mix

Established wholesale markets continue to absorb capacity while energy-led mega-campus expand in smaller metros and rural regions. Traditional colocation metrics and AI campus announcements measure different slices of the market. [1,2]

State	2024 GW	2030 GW (proj.)	Gas (2023)	Coal (2023)	Nuclear (2023)	Renewables (2023)
Texas	7.3	14.1	52%	13%	8%	27%
Virginia	6.4	22.3	56%	2%	33%	10%
Oregon	3.0	3.6	38%	0%	0%	62%
Ohio	2.1	5.1	59%	24%	12%	4%
Arizona	1.9	3.9	47%	11%	28%	14%
Georgia	1.8	5.4	47%	13%	29%	11%
Iowa	1.7	2.2	15%	24%	0%	62%
Illinois	1.6	3.3	16%	15%	55%	14%
California	1.5	2.3	44%	<1%	8%	48%
Washington	1.2	1.3	19%	4%	8%	69%
Indiana (emerging)	0.3	5.2	39%	46%	0%	13%
Pennsylvania (emerging)	0.6	2.8	59%	5%	32%	3%
Nevada (emerging)	0.9	2.5	58%	5%	0%	37%
Louisiana (emerging)	<0.1	1.1	76%	5%	13%	3%
Other U.S. states	~9.5	20.8	N/A	N/A	N/A	N/A
U.S. total	39.8	95.9	43%	16%	19%	21%

Capacity is EPRI's estimated nominal IT capacity, Medium scenario; 2030 is a projection of announced and planned projects, not operating supply, and includes cryptocurrency mining. Power mix is each state's 2023 in-state generation (EPA eGRID): the grid data centers draw from, not any campus's contracted or physical supply, and states also import and export power. Renewables = hydro, wind, solar, biomass and geothermal; oil and other fuels (generally under 2%, 3–4% in Indiana and Louisiana) are omitted. Selected emerging states have little operating capacity but large announced pipelines. [2,16]

Growth Is Concentrating

Texas and Virginia hold about a third of U.S. data center capacity today (13.7 of 39.8 GW), and the ten leading states hold about 70%. In EPRI's Medium scenario the two leaders reach 36.4 of 95.9 GW by 2030, so their share rises to 38% rather than falling. Virginia is the only state where data centers already use more than 20% of electricity; EPRI projects 39–57% by 2030, and seven more states (Oregon, Iowa, Nebraska, Nevada, Wyoming, Arizona and Indiana) could pass 20% in the Medium scenario. [2]

The map is also widening. Indiana, Louisiana, Mississippi, Ohio and Pennsylvania have little operating capacity but large announced pipelines, which EPRI reads as developers prioritizing power access and land for large AI training campuses. Indiana alone goes from 0.3 GW to 5.2 GW in the Medium scenario. The new markets bring different grids: Indiana’s generation is 46% coal and Louisiana’s 76% gas, so where a campus lands largely determines what powers it in the near term. [2,16]

Market / region	Evidence as of cutoff	Structural issue
Northern Virginia	4,496.5 MW primary-market operating inventory; 2,420.2 MW under construction [1]	Transmission, land access, regional ratepayer implications
Atlanta	2,882 MW under construction in CBRE primary market [1]	Fast construction, power and permitting
Dallas–Fort Worth	Established major wholesale market [1]	ERCOT supply, connection timing, demand competition
Phoenix	Established primary market [1]	Water-stressed geography; heat rejection
Central Ohio	Prometheus gigawatt-scale rollout, New Albany [6]	Grid coordination, nuclear contracting vs physical supply
Louisiana	Hyperion 5 GW announced compute capacity, Richland Parish [7]	New gas generation and community cost allocation
Abilene, Texas	Crusoe / Lancium 1.2 GW campus target [8]	ERCOT grid integration, backup turbines
Southeast Wisconsin	Fairwater initial building operational; second underway [9]	Supply chains, labor and industrial development

Analyst Interpretation

The industry is bifurcating into: (1) established carrier-rich metros with high operational demand and costly grid expansion; and (2) power-first AI campuses where electrical infrastructure, inexpensive land, long-term offtake and regulatory permission determine feasibility.

Comparability warning: CBRE tracks a defined wholesale market subset; EPRI’s state figures are modeled nominal IT capacity across all data center types. Prometheus/Hyperion/Abilene figures are company-announced campus or IT targets. None of these should be added together or compared as operating supply.

Financing and Ownership

Data center development separates land ownership, developers, facility operators, compute tenants, debt lenders and equity owners. The brand using the AI chips may not own the land or electrical infrastructure. Investment underwriting must map every layer of control and contractual obligation. [1,10] Durable value may attach to scarce grid access, electrical delivery and cooling, while GPU rental economics face utilization and hardware-refresh risk (analyst interpretation).

Layer	How money is made	Critical diligence
Land / developer	Site appreciation, development fees, long-term leases.	Land options, entitlement, power connection, completion risk.
Facility owner / REIT	Rent and contracted power/capacity.	Tenant duration, residual value, escalation, debt.
Compute operator	GPU capacity rental and usage.	Hardware utilization, refresh cycle, credit, concentration.
Generation / utility	Energy sales, capacity contracts, regulated return.	Rates, fuel risk, permitted MW, who carries shortfall.
Equipment / EPC	Transformers, turbines, electrical construction, cooling.	Order conversion, skilled labor, margin and supply chain.
Financial sponsor	Debt interest, equity return, asset appreciation.	Sponsor guarantees, construction draws, take-or-pay terms.

What to Discount

Announced investment is not capital already spent or financing actually closed. Do not treat a headline multi-year compute contract as free cash flow. Reported commitments must be reconciled to debt-service obligations, construction milestones and realized utilization. Project-finance stress can rise sharply when electrical delivery slips. Reuters reported Anthropic's August agreement to rent about 460 MW at Nscale Monarch over six years. Reuters separately described October discussions for a second 460-MW building, not yet an executed contract. Reports place initial capacity in late 2027; Nscale targets the wider 1.35-GW Phase 1 by early 2028. These may reflect different delivery milestones. [11,13,14]

Abilene's financing model. Crusoe's planned 1.2-GW, eight-building Texas campus involves Blue Owl and Primary Digital Infrastructure in financing and ownership. Its grid connection and backup gas turbines illustrate the layered relationship between the developer, infrastructure capital and compute tenants. Energized capacity and reported Oracle/OpenAI tenancy remain unverified in the reviewed primary sources. [8,10,15]

Nscale Monarch: A Finance and Community Watch Case

Nscale describes a self-powered campus independent of the public grid, closed-loop cooling without municipal drinking water, and protection of local electricity rates under state law. It cites over \$20 billion of planned campus investment; larger media cost estimates may include GPUs. These are developer-reported claims—not verified operating outcomes. Confirm generation fuel, water sourcing, legal tariff protections, total financing and eventual energization. [14]

Outside the Building: Electricity

A data center is rarely powered by a single fuel. The grid supplies a time-varying physical mix; operators may separately sign power purchase agreements (PPAs), buy clean-energy certificates, build behind-the-meter systems or pay for generation upgrades. [2,3]

Source	Share of 2025 U.S. utility-scale generation
Natural gas	41%
Renewables	24%
Nuclear	18%
Coal	17%

2025 U.S. utility-scale electricity generation, rounded shares; petroleum/other excluded from simplified comparison. EIA. [3]

Source	Why sought / constraints	Water and community issue
Natural gas	Dispatchable and scalable; turbines, pipelines and fuel exposure limit speed.	Thermal cooling water; CO ₂ emissions; local air permits.
Nuclear	Steady low-carbon generation; existing plants, uprates, restarts; new reactor schedules uncertain.	Thermal withdrawals; spent fuel; long timelines.
Coal	Still part of grid dispatch in relevant regions; load growth may affect retirement schedules.	Cooling withdrawals, emissions, local pollution.
Wind/solar	Fast low-carbon energy additions where interconnection is available; variable output.	Land, storage and transmission requirements; generally lower operational cooling water.
Hydro/storage	Flexible or firming value varies by geography and duration.	River management; hydrology; storage is not primary generation.

Three Checks for Every Campus

1. **Physical:** Which balancing authority and utility actually supply the site, and what is the hourly grid emissions/fuel mix?
2. **Contractual:** Which PPAs, nuclear offtakes and renewable certificates are associated with the operator?
3. **Incremental:** Which generating plants, storage and transmission assets would not exist without the project?

Hyperion illustrates incremental supply. Meta’s announced 5-GW Louisiana campus is backed by an agreement financing Entergy’s seven new natural-gas plants, three batteries, nuclear uprates and other purchases, plus support for up to 2.5 GW of clean generation. Those are proposed arrangements, not delivered MW or verified hour-by-hour campus fuel supply. [7]

Outside the Building: Water and Cooling

Cooling design does not settle water impact. Separate withdrawal (volume taken from a source), consumption (volume not immediately returned, often evaporated), discharge (volume returned), and indirect power-sector water (associated with generating electricity). [4]

Ceres estimates that data centers across seven analyzed states depend on approximately 3.4 trillion gallons of freshwater annually for electricity generation. This is Ceres's attributed water-dependency estimate, not metered on-site data center water consumption; its attribution method and hydropower treatment merit independent scrutiny. Ceres finds 66% of water-using generating plants in the study exposed to medium-high through extremely high water stress. [4]

Design choice	What it changes	Important limitation
Evaporative cooling	Often lowers electricity for heat rejection.	Consumes water; drought peaks matter.
Dry / air cooling	Can sharply reduce on-site evaporation.	May require more electricity, especially in heat.
Direct-to-chip closed loop	Efficient heat removal inside racks.	Final heat rejection can still use water.
Reclaimed / recycled water	Reduces reliance on potable freshwater.	Requires local supply, treatment and monitoring.
Cooling plus grid power	Combines direct site use and remote plant use.	Water burden may appear in a different watershed.

Site-Level Disclosure Standard

- Metered annual withdrawal, consumption and discharge in gallons or cubic meters.
- Source: municipal potable, groundwater, river, reclaimed or other; permitted peak day.
- Cooling system description plus on-site water usage effectiveness (WUE) where measurable.
- Indirect generation water by power plant or regional grid, keeping attribution caveats explicit.
- Drought or basin stress, competing users, mitigation commitments and compliance records.

Do not convert a claimed “zero-water evaporative” design into zero total water footprint; operational data and indirect generating impacts remain relevant. [4,8]

Water disclosures in practice. Crusoe reports roughly one million gallons of municipal water to initially fill each Abilene building’s closed-loop cooling system; that is not an annual-consumption figure. Nscale says Monarch’s proposed 1.35-GW Phase 1 will pair on-site power with non-potable closed-loop cooling that avoids municipal drinking water; delivered generating capacity and fuel, verified withdrawals, source-water data and ratepayer safeguards are not yet established. [14,15]

Outside the Building: Communities

Construction activity can be substantial, but permanent jobs are much smaller relative to capital invested: Meta forecasts 7,500 peak construction roles but about 1,000 operational roles at Hyperion. Such figures are project-specific and should not be generalized as a national per-MW employment ratio. [7]

Fairwater and redevelopment. Microsoft says its Mount Pleasant, Wisconsin, facility became operational in June 2026 with nearly 550 on-site workers, and describes \$4.7 billion of local investment through 2028. These reported benefits should be evaluated alongside public infrastructure costs, tax treatment and actual water and electricity consumption. [9]

Claim from developer	Independent verification needed
Local jobs / local hiring	Payroll location; employee vs contractor; permanent vs peak construction; wages and training.
Tax benefit	Abatements and exemptions; assessed values; school and county shares; long-term infrastructure spending.
Ratepayer protection	Utility tariff; generation cost recovery; collateral; minimum-payment and exit clauses.
Water responsibility	Permitted withdrawal and actual metering; aquifer or watershed context; competing demand.
Low-carbon operations	Hourly grid mix; contracted matched clean energy; gas generators and emissions permits.
Land stewardship	Acreage, previous land use, habitat, traffic, lighting, sound and land restoration.

The power-plant water stress described in the water section is also a host-region issue. [4] Financial Times and local reporting also document permitting opposition and construction execution risk. [5]

Who Pays for Grid Expansion?

CBRE reports growing use of take-or-pay floors at 60–85% of allocated electricity capacity. These contracts can require tenants to pay for reserved capacity even when utilization is lower. Large-load tariffs, minimum contract terms and exit fees can protect existing ratepayers from stranded infrastructure costs. The outcome depends on filed tariffs, forecast accuracy and enforceable cost-recovery provisions—not merely company assurances. [1]

Recommended Local Impact Scorecard

Rate each project on five transparent categories: economic additionality (net local value), energy affordability, water security, environmental burden and governance/transparency. Compare forward-looking commitments (announced investment, construction jobs, water-saving designs and ratepayer protections) with audited spending, permanent payroll, metered water and approved tariffs after commissioning. Use explicit evidence grades and avoid an ordinal national ranking when disclosures are incomparable.

2027–2030 Outlook and Watch Indicators

SemiAnalysis projects over 50 GW of data center IT additions in 2027 outside China, a global ex-China forecast rather than a U.S.-only number. It should not be silently substituted for EPRI's annual U.S. TWh demand range or CBRE's U.S. wholesale construction inventory. [12]

Scenario	What would have to happen	Key signal
Accelerated buildout	Advanced projects meet interconnection, construction and commissioning deadlines; tenants monetize compute.	Monthly energized MW and delivered GPUs rise; strong leasing utilization.
Power-limited base case	The grid, turbines, substations and permits govern actual commercial operation.	Large announced-to-energized gap persists.
Demand/finance correction	AI revenue, margins or credit conditions fall short of contracted capacity assumptions.	Lease cancellations, repricing, financing gaps, paused sites.

Quarterly Indicator Dashboard

- GW by lifecycle: announced / permitted / funded / physically built / energized / operational.
- Fuel supply: new gas MW, nuclear uprates/restarts, new renewables, storage and transmission completions.
- Water: permits, metered withdrawals, WUE, source basin stress and electricity-sector indirect footprint.
- Capital: closed equity/debt, offtake commitments, capex paid, counterparty concentration.
- Community: zoning hearings, tariff settlements, lawsuits, amendments and local benefit realization.

Editorial Interpretation

The architecture establishes the constraints: denser accelerator systems require new rack designs, higher-bandwidth fabrics, cooling systems and electrical infrastructure. Cloud providers, colocation operators, developers and financiers turn that demand into facilities; utilities and equipment manufacturers determine how quickly projects can operate. The costs and benefits extend to host communities. [1,2,4,7]

Answer to the central question: America's AI infrastructure is being built by hyperscalers, neoclouds, data center developers, utilities and financial sponsors; most enterprises use it through cloud and colocation rather than owning campuses. Inside, compute has shifted from individual servers toward densely networked accelerator racks and multi-building systems. Outside, that shift reshapes generation, transmission, water demand and local economics. Avoiding disproportionate community costs requires enforceable tariffs, transparent resource accounting and credible permitting—not just company assurances. [1,2,4,7]

If capacity grows as projected, power and cooling supply chains may have strong demand. If capacity does not materialize, stranded land, generation investment, debt and local infrastructure can create meaningful costs. The central opportunity is to measure the transition from announced ambition to energized economic value.

Q1 2027 watchpoints: Executed electricity agreements, actual energization, water permits and measured withdrawals, closed financing and utility rate decisions.

Methodology, Caveats and Source Register

Evidence hierarchy. (A) Government, regulatory and system-operator records; (B) primary company disclosure, labeled company-reported; (C) third-party market research, labeled estimate; (D) derived calculation or analyst interpretation, with assumptions shown. No undocumented campus-level fuel or water figures are imputed.

Project examples. Prometheus, Hyperion, Abilene, Fairwater and Monarch illustrate contrasting hyperscale engineering, financing, power and local-impact arrangements—not five different deployment-size categories. Examples are distributed through the analysis rather than ranked by GW. “Not disclosed” means unavailable in the reviewed evidence, not zero.

Comparability. CBRE wholesale MW, EPRI modeled IT MW/TWh, claimed campus GW and power procurement MW each have distinct scope. Direct cooling water is not interchangeable with power-plant withdrawal. A PPA does not prove hour-by-hour physical delivery.

Research and editorial disclosure. This report was developed and governed by 7-Seat LLC. Research objectives, analytical framework, data selection, interpretation and conclusions were determined by the author. OpenAI ChatGPT and Anthropic Claude were used to assist with research, drafting, source cross-checking, editing and production quality assurance. Responsibility for the report’s accuracy, analysis and conclusions remains with the author.

[1] CBRE, H1 2026 North America Data Center Trends

<https://www.cbre.com/insights/books/north-america-data-center-trends-h1-2026>

[2] EPRI, Powering Intelligence 2026

<https://powering-intelligence.epri.com/executive-summary.html>

[3] U.S. EIA, Electricity in the United States (2025 data)

<https://www.eia.gov/energyexplained/electricity/electricity-in-the-us.php>

[4] Ceres, Water Behind the Watts (Aug. 2026)

<https://www.ceres.org/resources/reports/water-behind-the-watts-the-hidden-risk-of-powering-data-centers>

[5] Financial Times, America’s AI infrastructure boom (Oct. 18 2026)

<https://www.ft.com/content/9ea5a37d-15c9-4d5b-94d6-719048c9cd3c>

[6] Meta Engineering, Inside Prometheus (Oct. 19 2026)

<https://engineering.fb.com/2026/10/19/data-center-engineering/inside-prometheus-meta-gigawatt-scale-supercluster/>

[7] Meta, Richland Parish expansion (July 2026)

<https://datacenters.atmeta.com/2026/07/deepening-our-investment-in-richland-parish-louisiana/>

[8] Crusoe, Abilene expansion to 1.2 GW (accessed Oct. 2026)

<https://crusoe.ai/newsroom/crusoe-expands-ai-data-center-campus-in-abilene-to-1-2-gigawatts>

[9] Microsoft, Fairwater first facility operational (June 2026)

<https://news.microsoft.com/source/2026/06/23/microsoft-completes-construction-on-first-datacenter-facility-in-mount-pleasant-wisconsin/>

[10] Crusoe / Blue Owl / Primary Digital Abilene joint venture (accessed Oct. 2026)

<https://www.crusoe.ai/resources/newsroom/crusoe-blue-owl-capital-and-primary-digital-infrastructure-enter-joint-venture>

[11] Reuters, talks for a second Nscale–Anthropic deal (Oct. 19 2026)

<https://www.reuters.com/technology/nscale-anthropic-talks-45-billion-cloud-computing-deal-bloomberg-news-reports-2026-10-19/>

[12] SemiAnalysis, Datacenters 2027 (Oct. 19 2026)

<https://newsletter.semianalysis.com/p/datacenters-2027-the-year-of-acceleration>

[13] Reuters, Anthropic 460 MW Nscale rental agreement (Aug. 26 2026)

<https://www.reuters.com/technology/anthropic-pay-nscale-45-billion-rent-ai-computing-power-bloomberg-news-reports-2026-08-26/>

[14] Nscale, Monarch Compute Campus, project description (accessed Oct. 2026)

<https://www.nscale.com/ai-infrastructure/monarch>

[15] Crusoe, Abilene site and cooling reference (accessed Oct. 2026)

<https://www.crusoe.ai/resources/blog/an-inside-look-at-the-abilene-ai-data-center>

[16] U.S. EPA, eGRID2023 Summary Tables, State Resource Mix (rev. 2, 2025)

https://www.epa.gov/system/files/documents/2025-06/summary_tables_rev2.pdf

- [17] Arista Networks, Rack-Scale Architectures for Scale-Up and Scale-Out AI Fabrics (Oct. 7 2026)
<https://www.arista.com/en/company/news/press-release/24818-pr-20261007>
- [18] IDC, Worldwide Quarterly AI Infrastructure Tracker, Q1 2026 (July 2026)
<https://www.idc.com/?p=117709>
- [19] Informa Connect, Where will enterprise data center operators run their AI? (July 2026)
<https://informaconnect.com/where-will-enterprise-data-center-operators-run-their-ai/>
- [20] AMD, Instinct MI455X GPU product specifications (launched July 23 2026)
<https://www.amd.com/en/products/accelerators/instinct/mi400/mi455x.html>
- [21] Gartner, AI-Optimized IaaS Spending to Grow 96% in 2026 (Aug. 10 2026)
<https://www.gartner.com/en/newsroom/press-releases/2026-08-10-gartner-forecasts-worldwide-artificial-intelligence-optimized-iaas-spending-to-grow-96-percent-in-2026>
- [22] Gartner, Forecasts Worldwide AI Spending to Grow 49.5% in 2026 (Sept. 16 2026)
<https://gartner.com/en/newsroom/press-releases/2026-09-16-gartner-forecasts-worldwide-ai-spending-to-grow-49-point-5-percent-in-2026>
- [23] Cloudian, Enterprise AI Infrastructure Survey 2026 (Mar. 3 2026; vendor-commissioned)
<https://cloudian.com/press/enterprise-repatriating-ai-workloads/>
- [24] Gartner, Forecasts Worldwide AI Spending to Grow 47% in 2026 (May 19 2026)
<https://gartner.com/en/newsroom/press-releases/2026-05-19-gartner-forecasts-worldwide-ai-spending-to-grow-47-percent-in-2026>
- [25] Google Cloud, What is an AI workload? (updated Oct. 15 2026)
<https://cloud.google.com/discover/what-is-an-ai-workload>
-

ABOUT 7-SEAT

7-Seat LLC is a brand infrastructure and go-to-market strategy consultancy working at the intersection of enterprise technology, positioning and AI. The AI Data Center is part of 7-Seat's Insights series of independent research on the technologies and infrastructure reshaping markets.