Columbia Business School
Climate Knowledge Initiative

23 January 2026

Powering Data

Hyae Ryung Kim, Ariela Farchi Behar, Isabel Hoyos, Una Oljaca, Shubhangi Prasad, Yosafat Partogi Simbolon, Clara Zibell, and Gernot Wagner

Executive Summary
Meeting AI Energy Demand

Rising demand: Data centers are on track to triple global electricity use by 2035, with AI the dominant driver. By 2030, they could consume nearly **1,000 TWh per year, ~3% of world demand.**

- Despite surging demand, emissions can plateau if data centers are increasingly powered by renewables. By 2030, over **50% of data center power is projected to be renewable**, but fossil fuels (especially gas and coal in China and the United States) remain a near-term power source. Decarbonizing the energy supply is critical to curbing data center emissions.

Regional trends: The U.S. and China alone will account for 80% of growth, making them the epicenter of both opportunity and grid stress. Europe will grow slower but faces permitting and public pushback. India is rapidly emerging as the next hub.

- **The U.S. grid** faces long interconnection queues, increasing household bills, and local moratoriums. This is pushing data center developers to shift to “power first” siting strategies and behind-the-meter solutions such as onsite renewables, small modular reactors (SMRs) for nuclear, and geothermal.
- **China’s state-led model,** Eastern Data, Western Computing, enables speed and grid decompression by moving data center clusters inland and away from major cities, offering a potential template for the West.

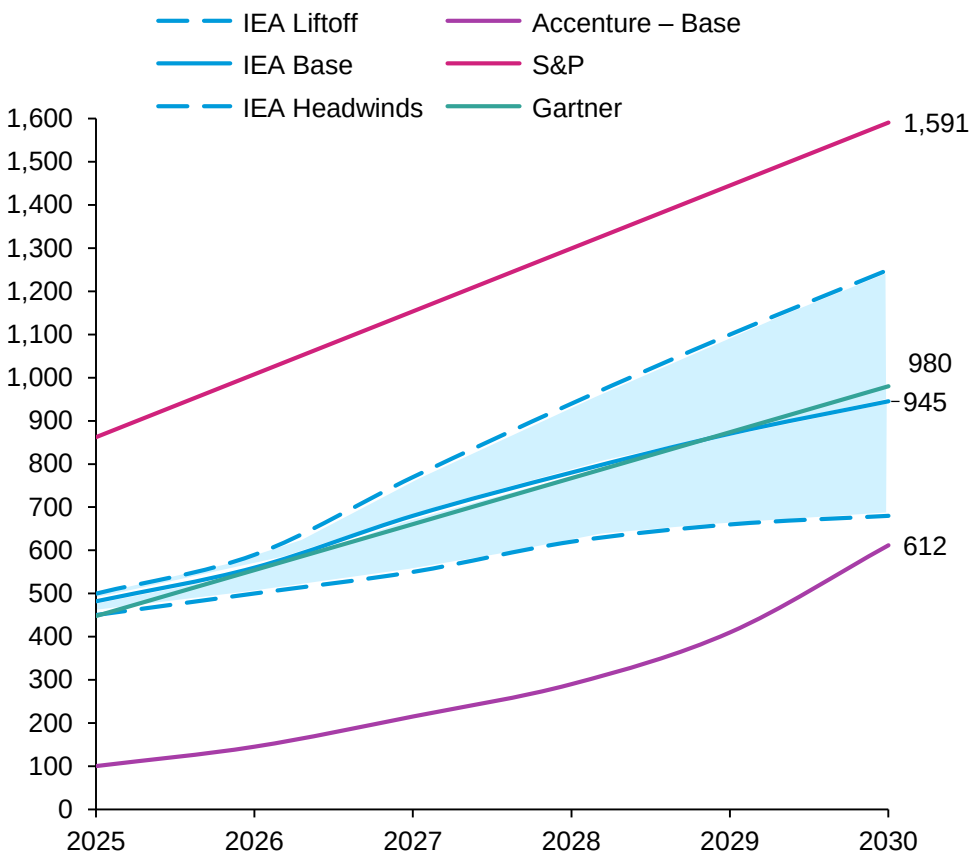
Data center operators: Amazon, Microsoft, Google, Meta, and other companies that operate hyperscale data centers control >50% of capacity and are the biggest buyers of renewables in the U.S. Their procurement decisions effectively shape clean energy investment pipelines.

- Efficiency gains in hardware (GPUs, custom chips), cooling (liquid vs. air), and AI model architectures (e.g., DeepSeek) can help data center operators cut energy use by 10-40x, but it is unlikely to fully offset growth in demand.

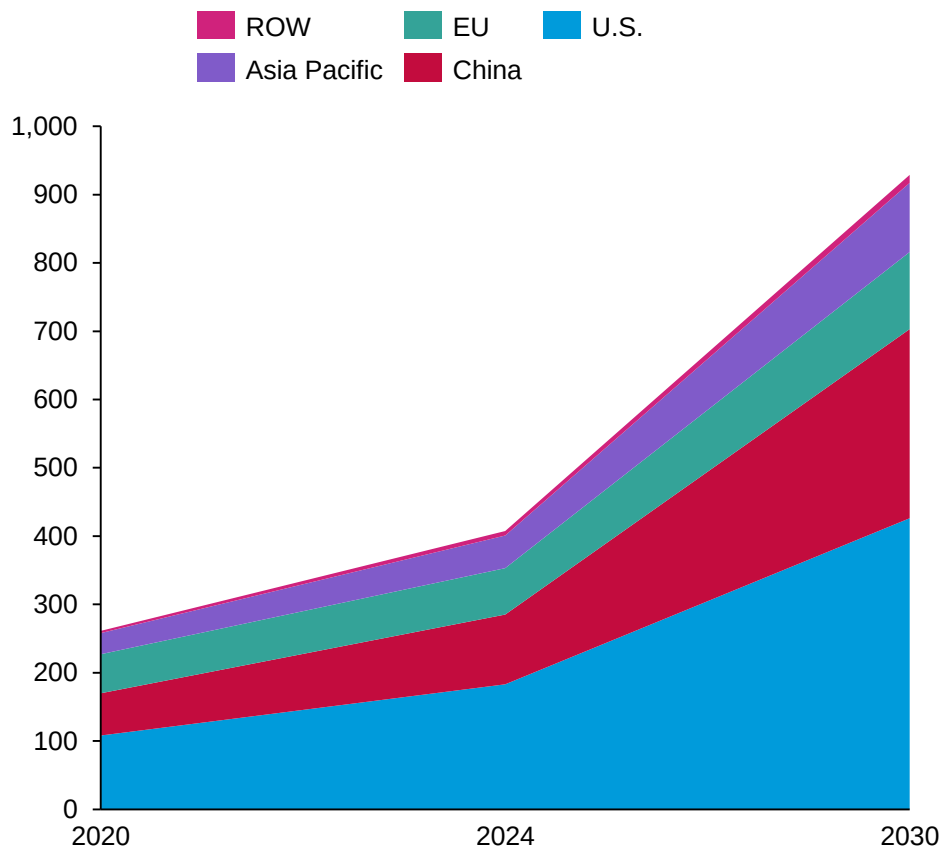
Emerging solutions: AI-driven grid optimization (e.g., virtual power plants, predictive load shifting), innovative energy deals, and geothermal baseload integration could be game changers in efforts to balance reliability, cost, and sustainability.

AI growth driving increased energy demand; U.S. and China ~70% of electricity used by data centers globally and growing

Projected global electricity consumption for data centers[1], TWh



Electricity consumption of data centers by region, IEA base case: 2020-2030



Observations

- Energy demand from data centers is projected to grow in line with adoption of AI.
- U.S and China will lead data center power demand by 2030.
  - U.S. to increase by 240 TWh (+130%)
  - China by 175 TWh (+170%)
  - Europe by 45 TWh (+70%)
- **Electricity demand in advanced economies is projected to rise** for the first time in over a decade.
- Data centers are expected to **make up 30-40% of new energy demand** between 2024 and 2030.
  - Emissions from data centers are projected to plateau and then decrease as the energy supply becomes decarbonized.

[1] Projections vary due to uncertain efficiency gains, inconsistent reporting, and different modeling approaches (bottom up, aggregated totals, hybrid, extrapolation).
Sources: IEA, Global data centre electricity consumption by sensitivity case, 2020-2035 (I2025); S&P, Global data center power demand by 2030 (2025); Gartner, Electricity demand for data centers to double by 2030 (2025); Accenture, Powering sustainable AI (2025).
Credit: Yosafat Partogi, Hyae Ryung Kim, and Gernot Wagner. Share with attribution: Kim et al., "Powering Data" (23 January 2026).

Columbia Business School
Climate Knowledge Initiative

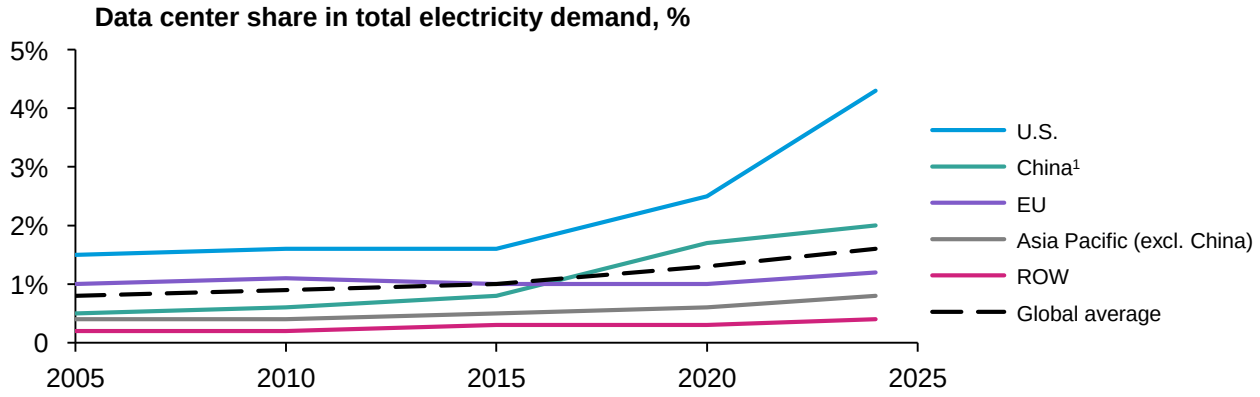
Data centers are concentrated in infrastructure-rich regions, driving gigawatt-scale clusters and local grid congestion risks

The U.S., EU, UK, and China host nine of the 10 largest data center clusters in the world by capacity, GW

Leads in global data center capacity and demand but faces mounting grid stress in key clusters

State-led data center clusters located in renewables-rich grid and moving inland to relieve grid stress

Prioritizes sustainability and regulation, but fragmented grids and public resistance slow data center growth



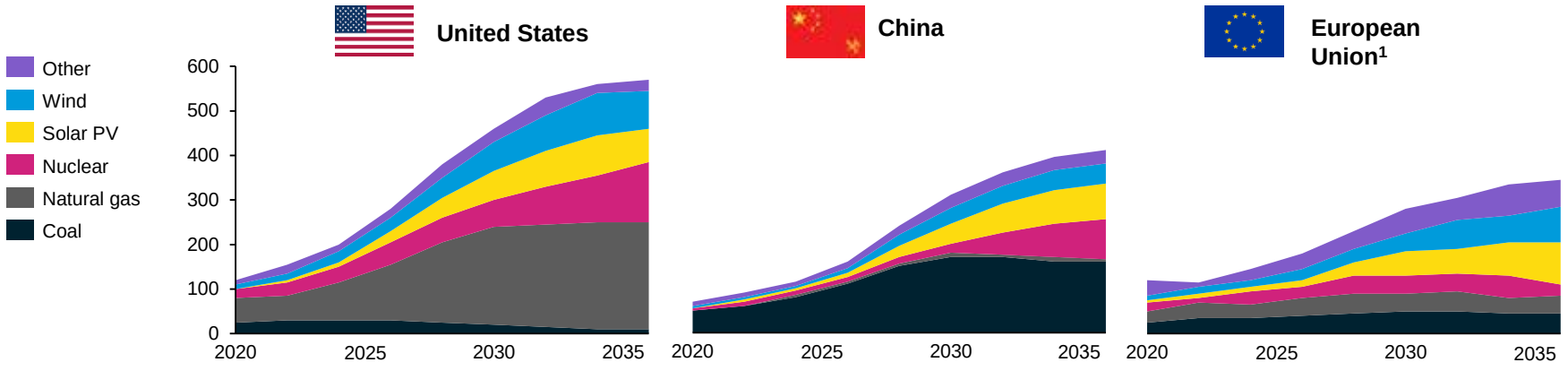
Observations

- Over **15% of global data center capacity under development** remains concentrated in the **top 10 markets** highlighted in the map above, underscoring the need for strategic planning to address growing pressure on local grids.
- In 2024, the U.S., China, and Europe together accounted **for ~85% of global data center electricity consumption,** with 180 TWh, 100 TWh, and 70 TWh, respectively. While the absolute share of national demand may seem modest — just over 4% in the U.S., ~2% in Europe, and ~1.1% in China — the **scale is significant** given these economies' vast power systems and competing demands from industries like manufacturing and transport and residential use.

Note: Singapore ranks among the top clusters but is excluded due to the map's regional scope. [1]Lack of substantial data makes it challenging to accurately estimate China's data center electricity consumption. Source: IEA, Energy and AI (2025).
Credit: Shubhangi Prasad, Isabel Hoyos, Hyae Ryung Kim, and Gernot Wagner. Share with attribution: Kim et al., "Powering Data" (23 January 2026).

Columbia Business School
Climate Knowledge Initiative

Clean energy is expected to grow in share of data center energy mix across geographies, despite different deployment drivers



	United States	China	European Union[1]
Deployment	**Market-driven, cluster-based model**	**State-led, decentralized model**	**Market-driven, urban-centered model**
Power demand growth projection	**+130% (~425 TWh by 2030)**	**+170% (~277 TWh by 2030)**	**+~70% (~45 TWh increase by 2030)**
Policy drivers	• CHIPS Act, AI and clean energy orders • Tax breaks, utility deals, DOE grants • Corporate power purchase agreements (PPAs)	• AI 2030 and Eastern Data, Western Computing • High and New Technology Enterprise (HNTE) 15% tax break, bundled permits, and fast approvals • Co-located with renewables	• Green Deal Digital, Digital Decade • Carbon markets, RRF, and CSRD-driven incentives • **France:** Offers tax breaks for green, efficient data centers • **Germany:** RRF funds support digital and energy upgrades
Permitting	**Fragmented** (1-3 years)	**Centralized and fast-tracked**	**Often slow** (5-7 years), local resistance
Grid flexibility required	**Flexibility urgent** industry piloting batteries, workload shifts, and gas backup	**Minimal flexibility** needed; strong central planning enables coordination but is underutilized	**Regulation-led flexibility;** automation and cooling shifts help, but grid limits slow progress

Observations

- U.S. deployment is **market-driven**, fragmented by state permitting and **dependent on corporate PPAs,** raising risk of grid bottlenecks in major data center clusters.
- **China's state-led model** offers speed and scale with centralized permitting, bundled land-power packages, and national mandates (e.g., Eastern Data, Western Computing), enabling rapid build-out.
- **The EU has strong decarbonization goals and regulatory levers** (CSRD, Green Deal Digital), but **slow permitting and public resistance** (e.g., Dublin moratorium) threaten deployment timelines.
- **Permitting bottlenecks** in the U.S. and EU contrast sharply with **China's fast-tracked national build-out**, giving China a decisive edge in infrastructure readiness.
- **A power mix shift** across regions converges toward more renewables and nuclear by 2035 — but China's top-down mandates may make the transition faster, even if its coal reliance remains a near-term emissions challenge.

[1] Projections based on IEA scenarios. EU data calculated by discounting US and China from global data.

Sources: IEA, Energy and AI (2025); Sinosities, Mapping China's Inland Data Centers (2025); Engineering, The "Eastern Data and Western Computing" Initiative in China Contributes to Its Net-Zero Target (2024); Bruegel, Annual Report (2024); Cushman & Wakefield, Americas Data Center Update (2025).

China’s strategy to move its data center clusters inland is enabling grid decompression and offering a model for the West to adopt

National Strategy

- **Eastern Data, Western Computing** aims to establish a network of **10 data center clusters** in Eastern China and **eight computing hubs** in Western China.
- **Computing power corridors** connect remote infrastructure to urban AI users.

Clean Energy Integration

- Data centers are **sited alongside renewable power bases** (Western China has abundant local clean energy for computing hubs to leverage).
- **Grid decompression:** The country is relocating compute away from power-constrained urban hubs.

Utility-Style Compute Access

.

- State-planned **data center hubs** serve as the central hubs for China's unified cloud computing and artificial intelligence infrastructure.
- Companies can remotely access computing power, enabling startups across China to train AI models without owning their own data centers.

Public-Private Governance

- Energy SOEs like State Grid and Nyocor co-invest in data centers, ensuring **integration of power infrastructure with AI compute.**
- Private tech giants (e.g., Alibaba, Tencent, Huawei) are major operators in these hubs but work under central policy guidance and long-term MOUs with local governments.

Case study: Xinjiang and Qinghai, China

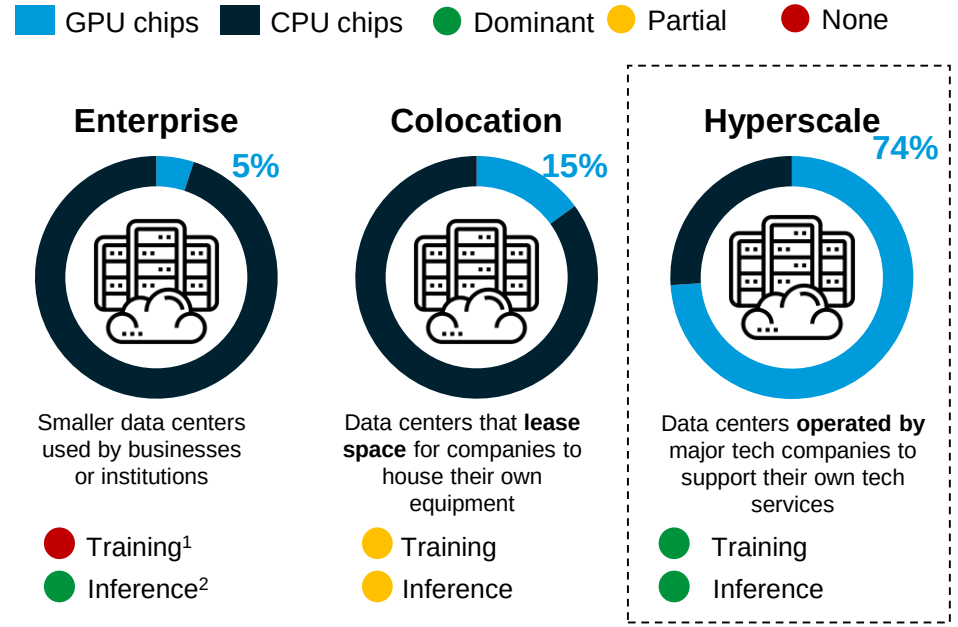
- **Massive build-out:** 39 data center projects have been approved in Xinjiang and neighbouring Qinghai, with plans to use over 115,000 Nvidia chips.
- **Strategic location:** Utilizes remote, low-cost land with clean energy, especially in Qinghai.
- **Deployment model:** Nyocor, a partly state-owned green energy company, is building a large data center with 625 servers and 2,000 Nvidia H100 chips.
- **China’s state-led AI infrastructure model** shows how shifting data centers away from major cities can ease grid stress, offering a **strategic alternative to the West’s market-driven approach.**

Sources: Bloomberg, China’s Got Big Plans for AI — In the Desert (2025); Strider Technologies and SCSP, China’s AI Infrastructure Surge Report (2025).
Credit: Shubhangi Prasad, Isabel Hoyos, Hyae Ryung Kim, and Gernot Wagner. Share with attribution: Kim et al., "Powering Data" (23 January 2026).

Columbia Business School
Climate Knowledge Initiative

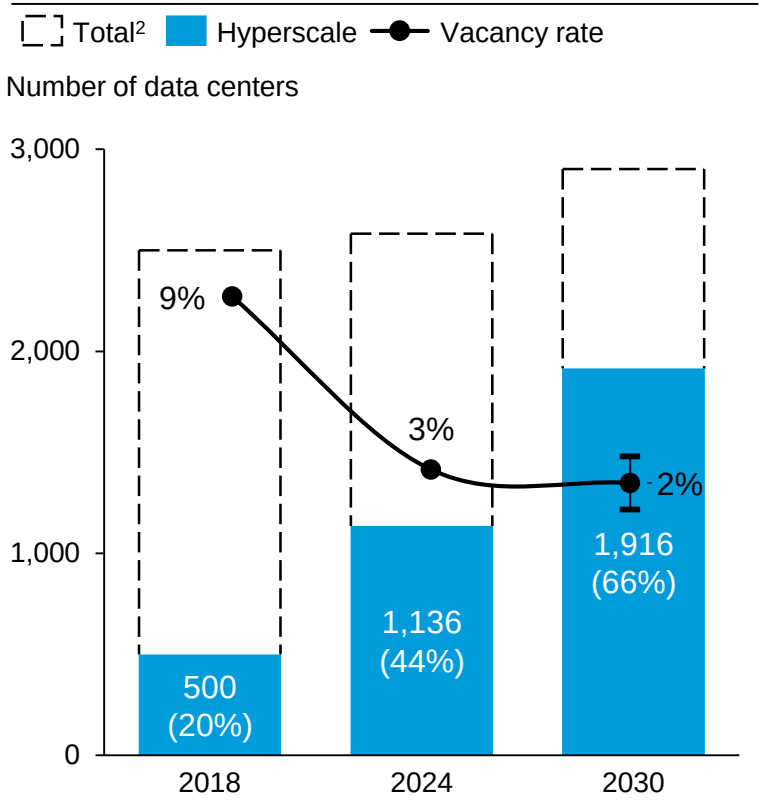
AI workloads fuel demand for high-capacity, energy-intensive data centers; low vacancy rates show demand outpaces supply

Breakdown of AI workloads and hardware[1] in modern data centers, % (2024)



GPU chips are ideal for AI-heavy workloads thanks to their high memory bandwidth and parallel-processing capabilities. The scale of compute required for training makes **hyperscale data centers highly energy-intensive**, despite GPUs being more energy efficient than CPUs for AI.

Global hyperscale data center growth (count) **and U.S. vacancy rates,** % (2018-2030)



Observations

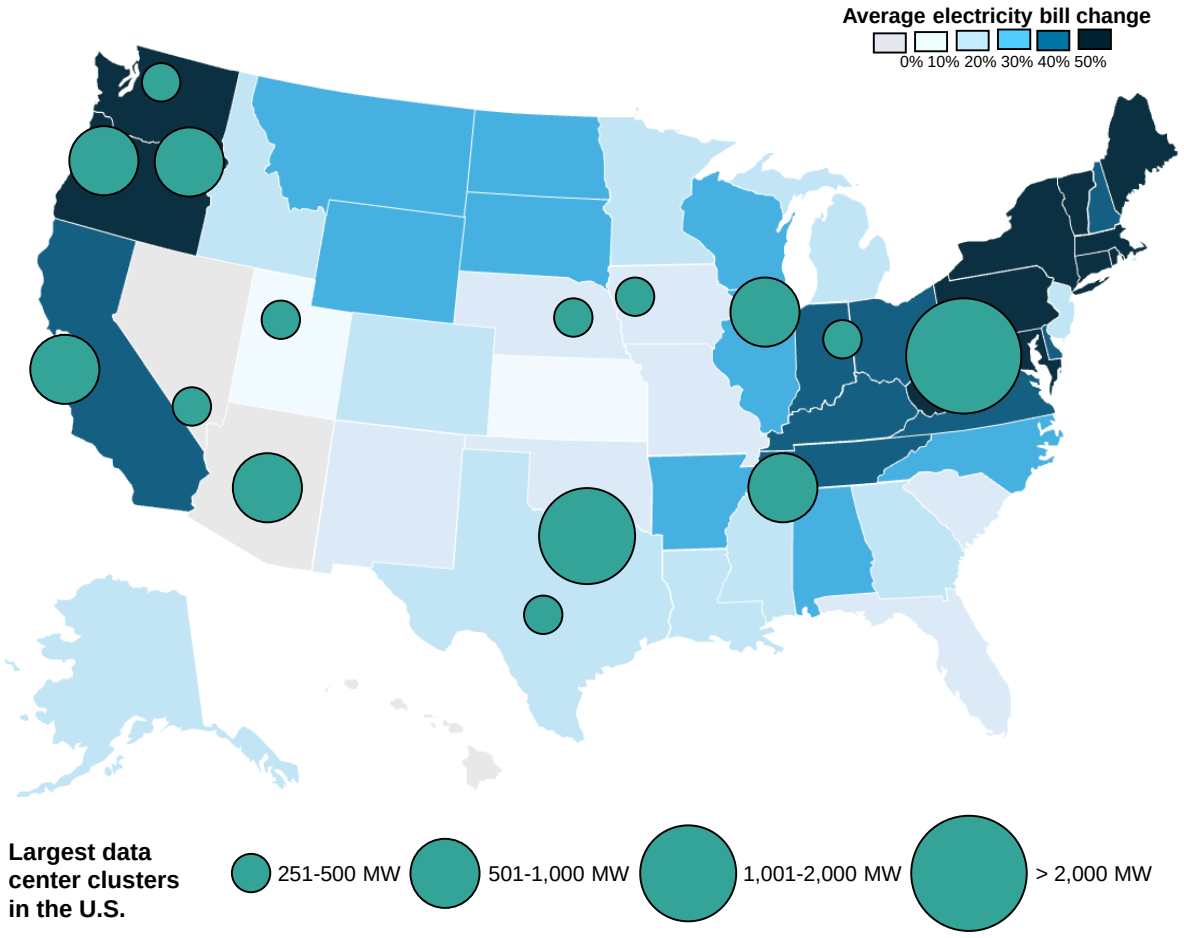
- All three types of data centers perform AI workloads, but **hyperscale data centers are AI-dominant**, which explains higher power and capacity requirements.
- Overall hyperscale capacity growth will be driven by **larger data centers**, rather than facility additions. Median data center size is expected to grow over the next 10 years (up to 375 MW of total capacity).
  - Hyperscale data centers >100 MW consume as much power as 350,000 to 400,000 electric vehicles do.
- Global vacancy rates fell to a record low of 6.6% in Q1 2025. In contrast, average vacancy rates in U.S. primary markets fell to a record low of **1.9% in Q4 2024.**
  - Despite a 43% year-over-year capacity increase in the top four U.S. data center markets (Northern Virginia, Chicago, Atlanta, Phoenix) in Q1 2025, data center availability remains extremely limited.

[1] Central processing unit (CPU) chips are ideal for general-purpose, flexible, and sequential workloads. Graphics processing units (GPU) are optimized for high-speed processing and ideal for compute-intensive workloads like AI training and inference. [2] Total data center projections are highly uncertain, as these include both AI and non-AI data centers.

Sources: IEA, Energy and AI (2025); Synergy Research Group, Hyperscale Operators and Colocation Continue to Drive Huge Changes in Data Center Capacity Trends (2024), McKinsey, How data centers and the energy sector can sate AI's hunger for power (2024); IEA, What the data centre and AI boom could mean for the energy sector (2024); BloomEnergy, Onsite Generation Expected to Fully Power 27% of Data Center Facilities by 2030 (2025), CBRE, Global Data Center Trends 2025 (2025).

Data center growth is adding to U.S. grid strain and pushing up household electricity bills

Household electricity bills are rising near the largest data center clusters



Case study: Dominion Energy, Virginia

- Northern Virginia is the **world's largest data center hub** with a data center capacity of over **5.9 GW,** led by Amazon Web Services (AWS).
- **Dominion Energy,** Virginia's main utility, is grappling with soaring power demand from data centers. By **2030**, an **extra 11 GW of capacity is anticipated**, which would be **more than 40% of Virginia's current peak demand.**
- **A $35 billion AWS expansion** by 2040 is resulting in a massive load increase.
- Dominion must increase its current capacity by nearly 5x by 2038 → **13.3 GW.**
- **+$8.51 per month by 2026:** Dominion has proposed its first residential rate hike since 1992, citing data center demand, to manage grid stress.

Observations

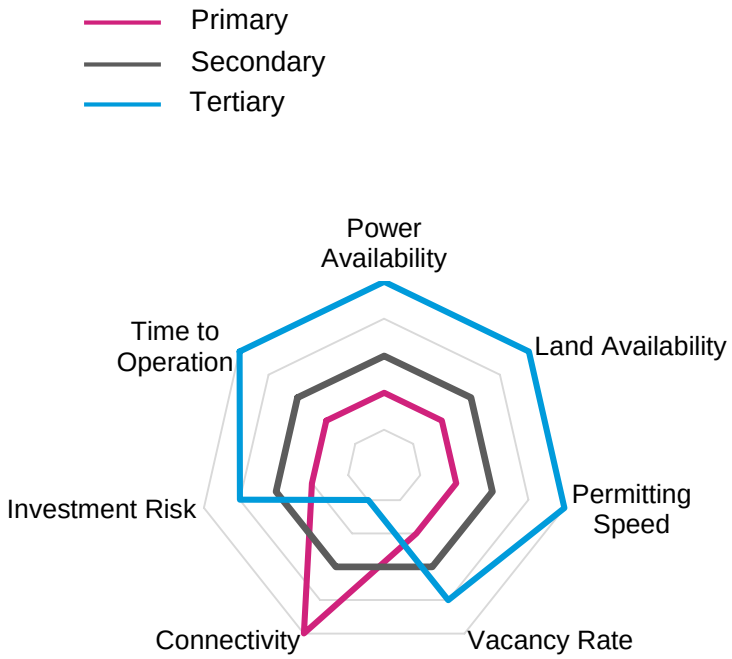
- U.S. power demand is rising sharply, **with data centers and cloud computing emerging as key drivers** of the surge.
- Data centers used **4.4% of U.S. power in 2023,** which could hit **6.7 to 12% by 2030.**
- **By 2028,** U.S. data center growth could **add 150-400 TWh of electricity use, doubling or tripling their 2023 share.**
- The grid needs **18-47 GW more capacity by 2030** to keep up with this increasing demand**.**
- **The PJM grid (13 U.S. states) expects electricity bills to rise 20%+ this summer (2025),** largely due to surging demand from AI and cloud data centers.
- **Virginia's electricity demand could jump 183% by 2040** if unchecked data center growth continues — forcing utilities to build costly infrastructure that pushes up household bills.
- **Grid upgrades and delays in new power projects mean costs get passed to all customers,** not just data centers — prompting states like New Jersey, Georgia, and Oregon to propose separate rate classes to protect residents.

Sources: McKinsey, How data centers and the energy sector can sate AI's hunger for power (2024); The White House Council of Economic Advisers, The Economic Benefits of Unleashing American Energy (2025); Reuters, America's Largest Power Grid Is Struggling to Meet Demand from AI (2025); CBS News, The AI Revolution Is Likely to Drive Up Your Electricity Bill (2025); Power, How Much Power Will Data Centers Consume? (2024); The Washington Post, Amazon launches $35 billion data center expansion in Virginia (2023); Cushman & Wakefield, Americas Data Center Update (2025).

Columbia Business School
Climate Knowledge Initiative

Rising data center demand and saturated primary markets drive ‘power first’ strategy to cut costs and speed deployment

Trade-offs in data center siting across primary, secondary, and tertiary markets[1]



- **Primary markets** remain attractive given established infrastructure and proximity to end users, but limited land and power access leads to **interconnection delays.**
- **Hyperscale data centers** are increasingly being sited in **secondary markets** due to faster permitting and stronger power availability.
- **Tertiary markets** offer abundant land and power but face higher operational risk.
- **Environmental factors** like water availability and environmental justice impacts are **often overlooked in data center siting,** despite growing scrutiny.

Power-first development prioritizes a secure power supply over site selection, incentivizing geographic expansion

The power-first model **prioritizes securing energy capacity** before site selection, permitting, or construction begins, as energy availability becomes the key constraint on scaling AI infrastructure. This can mean siting data centers in **secondary or tertiary markets** with less congestion as primary markets become saturated.

Advantages	Key priorities
✚ Eases grid burden	• Resolve regulatory uncertainty
✚ Reduces transmission distance	• Build new network infrastructure
✚ Improves reliability and resilience	• Develop labor force

Example: Google + TPG Rise Climate + Intersect Power (2024)

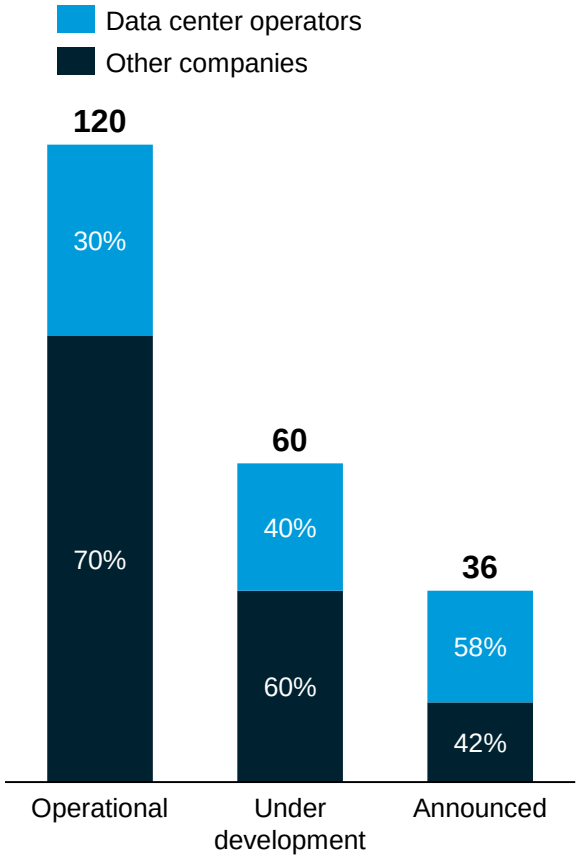
- The partnership between **Intersect Power, Google, and TPG Rise Climate** will deliver renewable power and storage to new data centers.
- Intersect Power plans to catalyze **$20 billion in clean energy and storage infrastructure** by 2030.
- Google will act as an anchor tenant in colocated industrial parks, pairing data centers with clean power.
- Phase 1 is expected online in 2026; full completion is targeted for 2027.

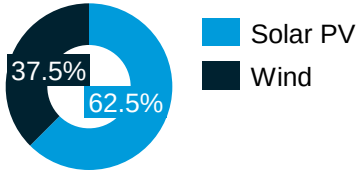
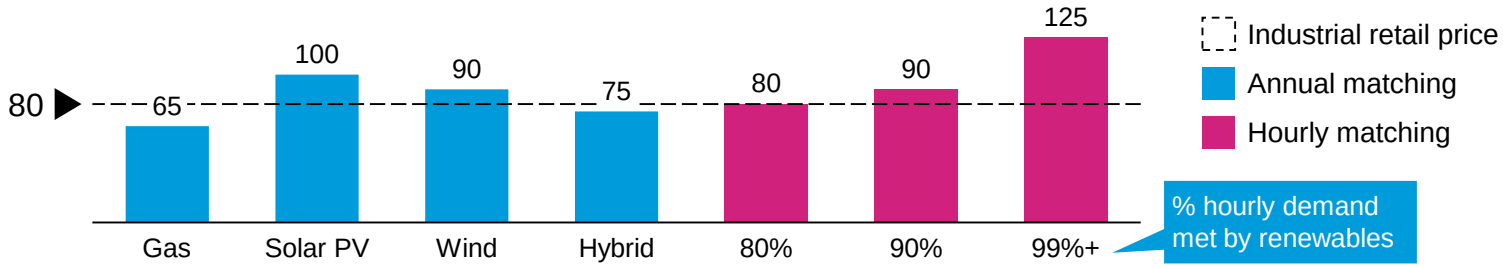
[1] Primary markets are established hubs with dense infrastructure; secondary markets are emerging cities with growing demand and lower costs; tertiary markets are smaller or rural areas just beginning to attract data center investment.

Sources: CKI Analysis; Intersect Power, Intersect Power Forms Strategic Partnership with Google and TPG Rise Climate (2024); Latitude Media, Google’s new data center model signals a massive market shift (2024).

Data center operators account for a growing share of renewable PPAs, mostly annual, with increasing hourly matching

Corporate renewable PPAs by project stage, GW



	Annual matching PPAs	Hourly matching PPAs
Description	Enough capacity procured to meet 100% of annual power demand, independent of hourly demand.	Some or all electricity consumption matched hourly by low-emissions sources located in the same region.
Advantages	Help meet renewable energy goals while retaining flexibility in electricity sources and lower costs.	Higher guarantee of covering demand, decreasing emissions, and mitigating volatility risk of power costs.
Disadvantages	Independent of real-time energy use, limiting value for grid reliability.	Higher costs, more operationally complex, and regionally limited due to baseload needs.
Primary power sources	 Data center PPAs in the U.S. by electricity source	Google and Microsoft developing hybrid portfolios that harness baseload technologies such as geothermal and nuclear to ensure continuous renewable power supply. Google achieved an average of 66% carbon-free energy in 2024.
Average cost of power[1] (US$/MWh)		

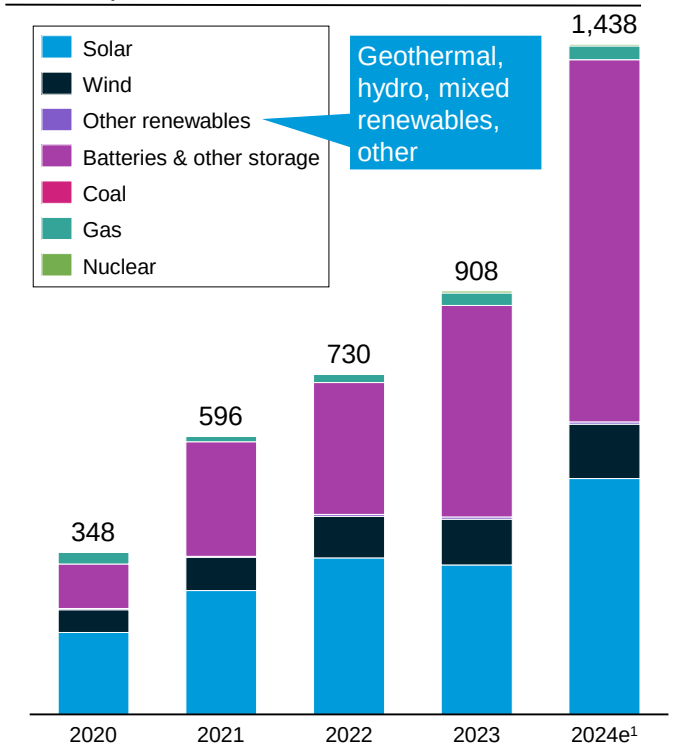
[1] Excluding grid fixed costs.

Sources: IEA, Energy and AI (2025); McKinsey, How hyperscalers are fueling the race for 24/7 clean power (2024); Google, Environmental report (2025).

With U.S. interconnection queues growing >3x, operators are expanding to onsite generation to meet growing power demand

With interconnection delays and long construction timelines increasing costs, wait times, and risks for project developers…

U.S. interconnection queue by energy source, GW



Long wait times

Five-year average wait time from request to commercial operation

70% increase in time required to secure a connection since 2014

Low success rates

10% success rate for projects expected to come online by 2028

20-80% Interconnection agreement suspension rates across the U.S.[1]

NYISO, SPP, PJM, and ISO-NE have higher suspension rates (46-79%); ERCOT, CAISO, and MISO have lower rates (~20%)

…data center operators are moving beyond the utility grid to secure a stable power supply

Grid-connected behind-the-meter generation

Primarily serves local load, with limited or no export.

- ✚ Reduces grid dependence
- ✚ Improves resilience
- ✚ Often avoids interconnection queue
- ▬ Adds CapEx and OpEx; still reliant on grid for full operation

In 2025, **APR Energy** announced the deployment of four mobile gas turbines providing **>100** MW of BTM power to an unnamed U.S. hyperscale data center operator.

Grid-connected with microgrid capabilities

Able to disconnect from the grid and operate independently.

- ✚ High resilience
- ✚ Operational during outages
- ✚ Optimizes use of onsite renewables
- ▬ Requires advanced controls and storage → complex, expensive

Snohomish PUD's microgrid in Arlington, WA, integrates solar, BESS, and EV charging with intelligent controls to support a data center and enable seamless islanding during outages.

Fully off-grid generation

Off-grid sources operate entirely independent from the grid.

- ✚ No exposure to interconnection delays or grid failures
- ✚ Enables remote or modular deployment (e.g., military purposes)
- ▬ High upfront costs, risk of supply shortfalls

ECL announced a **1 GW** off-grid, modular, hydrogen-powered colocation data center in Texas with Phase 1 completed in summer 2025.

[1] Interconnection agreement suspension rates refer to the frequency at which approved projects pause or delay their grid connection timelines.

Sources: Berkeley Lab, Grid connection barriers to renewable energy deployment in the United States (2024); Berkeley Lab, Generation, Storage, and Hybrid Capacity in Interconnection Queues (2025); Enverus, 2025 Interconnection Queue Outlook (2025); Utility Dive, Fortress backs behind-the-meter gas turbines (2025); Business Wire, ECL Announces World's First 1 Gigawatt Off-Grid, Hydrogen-Powered AI Factory Data Center (2024); Tencent, How a Microgrid Is Solving a Key Energy Challenge (2025), OffgridAi, How off-grid solar microgrids can power the AI race (2024).

Columbia Business School
Climate Knowledge Initiative

Nuclear and geothermal can deliver clean and firm electricity for data centers, strengthening grid reliability for 24/7 operations

Nuclear and geothermal can reduce reliance on fossil-fuel peaker plants as a growing share of data center electricity demand is met by renewables…

Firm power

Power source that **provide continuous, uninterrupted energy** regardless of season or weather conditions.

e.g. CAISO snapshot: **nuclear, geothermal, biogas,** and **biomass**.

Peaker plants

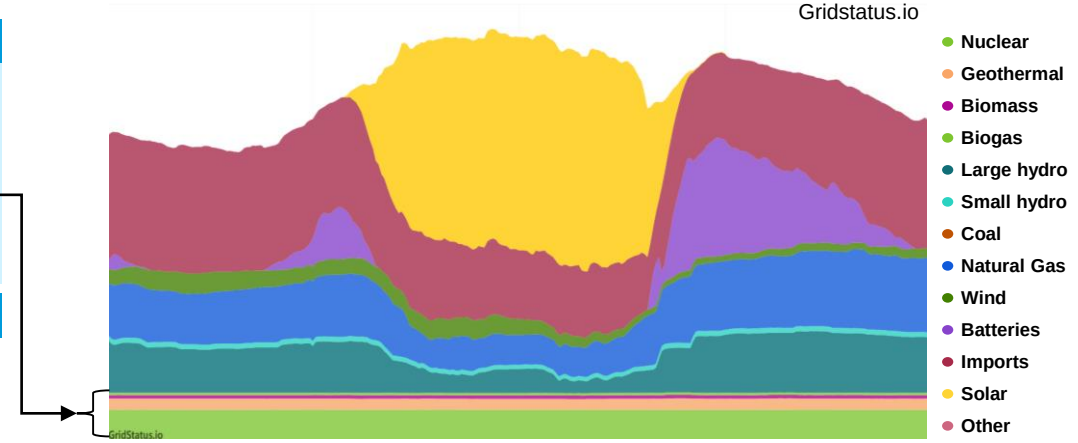
Power facilities that **meet grid needs during periods of highest electricity demand** (peak load), helping maintain grid reliability. They operate flexibly and can be turned on quickly to respond to surges in demand.

Clean energy transition

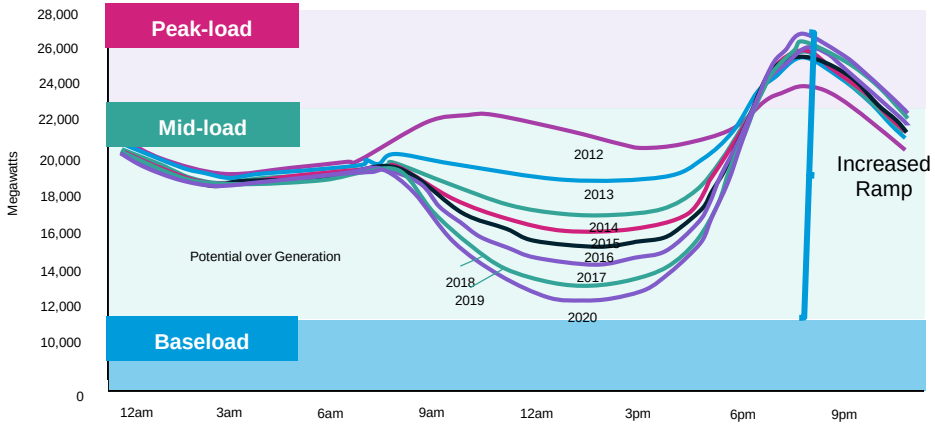
As aging Peaker plants retire, there is a **need to bring new firm power sources online** to continue ensuring grid stability.

Nuclear and geothermal emerge as sources of clean, firm energy that can **compliment intermittent renewable sources by providing continuous clean power – helping create a reliable grid.**

Fuel mix for CAISO, 19 January 2026
Gridstatus.io



Hourly energy demand (year average) in California



… and satisfy data centers' energy needs

1

High-capacity factor

~70-90% with geothermal nearer to 70 and nuclear to 90, ensures stable and continuous electricity supply

2

Carbon free energy

~0g CO_2e/kWh of energy produced enables hyperscalers to work towards hourly clean energy targets

3

Longer service lifetime

Lifespan of 30-50+ yrs for geothermal and **60-80 yrs** for nuclear

Sources: IEA, Energy and AI (2025); McKinsey, 2025 Global Energy Perspective, (2025); Utility Drive, Clean firm generation in the energy transition (2024); Gridstatus.io, Fuel mix – CAISO (2026);
Credit: Ariela Farchi, Isabel Hoyos, and Gernot Wagner. Share with attribution: Kim et al., "Powering Data" (23 January 2026).

Columbia Business School
Climate Knowledge Initiative

Big tech hyperscalers account for >50% of new U.S. clean energy deals, leveraging agreements to meet growing AI demand

Notable energy investments announced by leading data center operators in the U.S., 2020-2025[1]

Microsoft

- **835 MW, Nuclear:** Enabled the restart of Constellation Energy's nuclear facility in Pennsylvania, which had been retired in 2019
- **500 MW, Solar:** In line with its data center community pledge, signed an agreement with Pivot Energy to develop over 100 community-scale solar projects to the ease data center energy burden
- **250 MW, Solar:** Built a new solar project with National Grid Renewables as part of a $3.3 billion investment in AI infrastructure in Wisconsin

Meta

- **1.1 GW, Nuclear**: Partnered with Constellation Energy to power an existing high-performing nuclear facility in Illinois, allowing the plant to relicense and continue its operations for 20 years
- **790 MW, Solar and Wind:** Signed four energy agreements to procure solar and wind power, delivered through the local grid in Ohio, Arizona, and Texas
- **150 MW Geothermal:** Partnered with XGS to deploy a next-generation geothermal project in New Mexico, leveraging XGS's proprietary technology

Amazon

- **2.5 GW, Nuclear:** Invested more than $1 billion in nuclear energy in the U.S. across developing SMRs and existing plants with X-Energy, Dominion, and Talen Energy
- **650 MW, Wind:** Invested in the first utility-scale wind farm in Mississippi in collaboration with AES and agreed to fund future upgrades, positioning it to enable 1.3 GW of renewable energy in the area
- **200 MW, Solar:** Repurposed an old coal mine brownfield into the largest expected solar farm in Maryland with CPV Backbone

Google

- **3 GW, Hydro:** Signed largest hydroelectricity agreement to date with Brookfield
- **890 MW, Solar:** Invested in the largest solar project in MISO with Swift Current and network updates in PJM with Energix
- **500 MW, Nuclear:** Partnered with Kairos power to lower costs and scale access to power in communities with SMRs
- **120 MW, Geothermal:** Collaborated with Fervo Energy to secure reliable carbon-free energy, enhancing and scaling its geothermal technologies in Nevada

Observations

- By 2030, data centers are expected to source upwards of 1000[2] TWh annually.
  - Amazon, Microsoft, and Google account for 59% of all hyperscale data center capacity.
- Amazon, Microsoft, Meta, and Google all have renewable energy targets, with some going as far as having 24/7 carbon-free energy goals. **Procuring energy from renewable sources is critical to achieving established goals while meeting increasing energy demand** from AI training and inference data centers.
  - Tech companies have seen an **increase in energy demand and scope 2 emissions** as a result of increasing data center capacity.
  - Renewable PPAs and offtake agreements allow companies to curb scope 2 emissions while meeting increasing energy requirements.
- For energy developers, an agreement with a hyperscale operator indicates project security and **creates attractive conditions for future investors.**

[1] Investments have been announced; this does not imply that the projects are in operation with the MW online yet. [2] Based on IEA projections – number varies based on projection source.
Sources: Google, Environmental Report (2025); Microsoft, Environmental Sustainability Report (2025); Meta, Sustainability report (2024); Amazon, Carbon-Free Energy (2024); IEA, 5 ways Big Tech could have big impacts on clean energy transitions (2021); Tech Drives Power Contracts, World Economic Forum (2025); Utility Dive, Big tech is upending the energy landscape (2024); CKI Analysis.

As U.S. energy demand outpaces grid capacity, hyperscalers are investing in solutions to boost efficiency

Hyperscalers face energy consumption issues

Google Google's data center electricity consumption surged by **27% in 2024,** reaching 30.8M MWh, or a **7x increase over the past decade.**

Meta Meta's planned data center in Louisiana is projected to consume **2.3x the electricity consumed by the entire city of New Orleans.**

amazon AWS rapid credit program is unable to secure guaranteed power for new data centers in Northern Virginia due to **local grid strain, causing a bottleneck and delayed expansion plan.**

Efficiency solutions

High Medium Low

AI-driven optimization

Investment level: → **Efficiency gain:** ↑

- **Intelligent cooling control** – **Google** leverages DeepMind to create autonomous, cloud-based control systems for cooling its data centers, to reach a 40% reduction in energy used for cooling.
- **Predictive maintenance** – **AWS** can forecast component failures and automate the process of moving workloads to healthy hardware without customer impact using telemetry data in real time.
- **AI-powered load shifting** – **Microsoft** uses ML models trained on historical workload patterns (specifically CPU usage) to forecast future demand within Azure.

Advanced cooling systems

Investment level: ↑ **Efficiency gain:** →

- Hyperscalers are **shifting from traditional air cooling to more efficient liquid cooling.**
- Amazon is deploying custom **cold-plate liquid cooling systems** to better manage the heat generated by powerful AI chips.
- Recent Microsoft research shows that **switching cooling** can reduce:
  - **15 to 20%** of energy demand
  - **30 to 50%** of water consumption

Energy-efficient hardware

Investment level: ↑ **Efficiency gain:** ↑

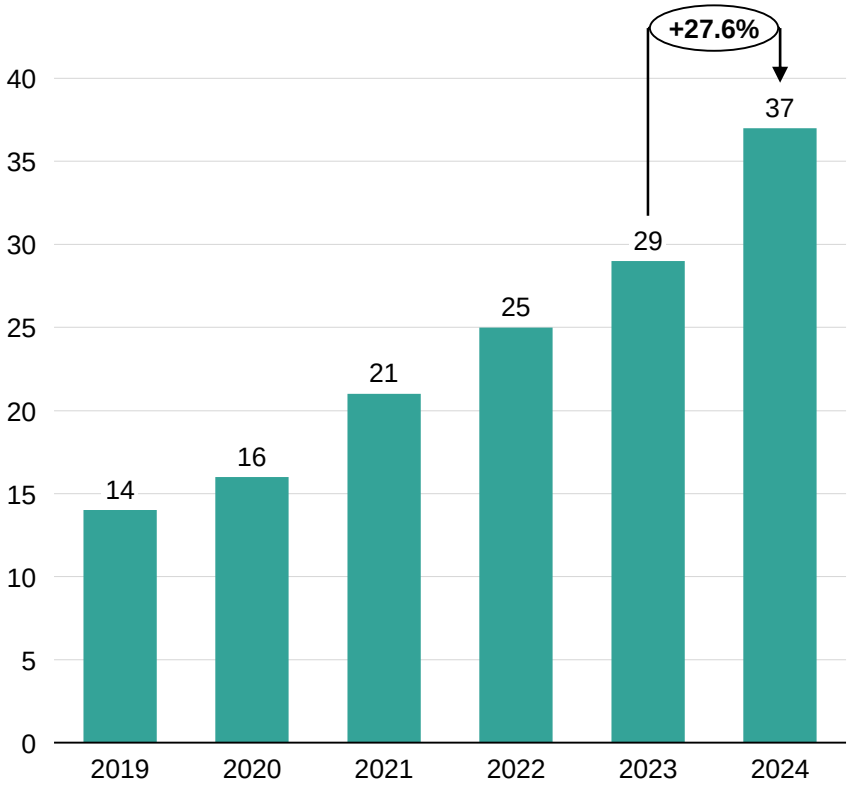
- **Lower power processor –** CPUs and GPUs are being engineered for superior performance per watt. Nvidia and Intel are developing specialized chips for specific workloads.
- **Efficient memory (RAM)** – Industry is shifting toward lower voltage memory modules, which will lead to substantial energy savings at the massive scale of hyperscale data centers.
- **Replacing hard disk drives with solid-state drives** will consume less power.

Source: Financial Times, How more efficient data centres could unlock the AI boom (2025).
Credit: Yosafat Partogi, Hyae Ryung Kim, and Gernot Wagner. Share with attribution: Kim et al., "Powering Data" (23 January 2026).

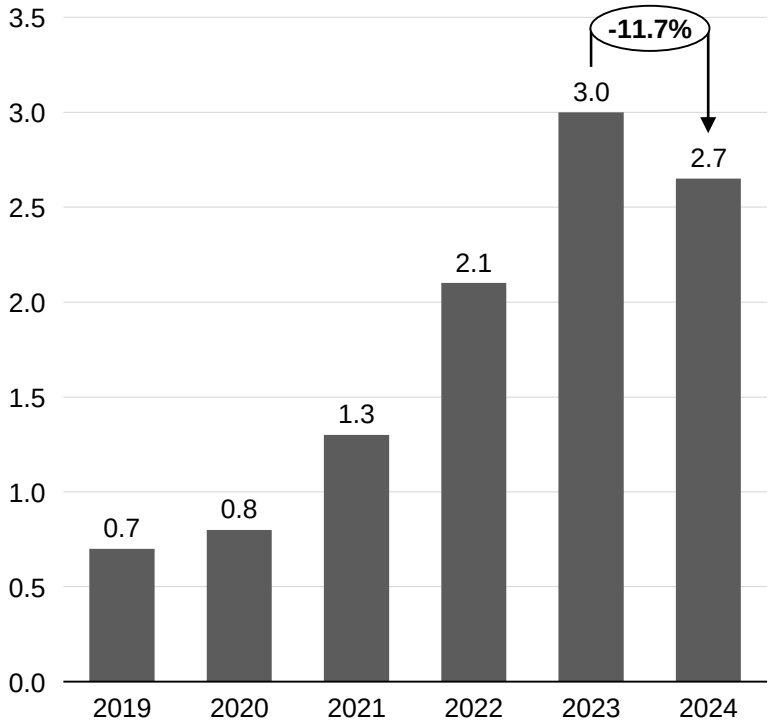
Google's data centers showed a decrease in emissions of ~12% in 2024, despite a ~28% increase in energy demand

With data centers nearly doubling energy consumption since 2019, Google has focused on clean energy procurement and efficiency to control emissions

Data center electricity consumption, million MWh



Data center energy emissions, $MtCO_2e$



Efficiency

- Google deployed highly efficient computer chips, TPUs, which has led to a 6x improvement in computing power per unit of electricity compared to that in 2020.
- In collaboration with local utilities, Google has implemented a demand response initiative, aiming to make its machine learning capabilities more flexible and shifting power demand from its data centers during certain times of the day or year.

Clean energy procurement

- Google signed its first PPA in 2010 and has been pioneering power purchasing models, with more than 2.5 GW of clean energy online to serve its operations.
- To procure clean energy faster, Google reimagined its PPA process, cutting negotiation and execution time by 80%.
- Colocation deals with TPG Rise and Intersect Power bring a new way to power AI development.
- In 2024, clean energy purchases resulted in an estimated 8.2 $MtCO_2e$ avoided.

Sources: Google, Environmental Report (2025); Google, A Smarter Way to Buy Clean Energy (2023); Google, How We're Making Data Centers More Flexible (2025); Google, A New Approach to Data Center and Clean Energy Growth (2024).

DeepSeek disrupts AI industry with key architecture and model developments; efficiency improvements target training phase

Key architecture and model innovations in DeepSeek V3

Targeted Improvements: Speed Cost Performance

Multi-head Latent Attention

Reduces key-value (KV[4]) cache by rearranging the traditional matrix into a smaller latent structure.

- Reduces required memory (13-5%) and compute operations per token, increases inference speed

Mixture of Experts[3]

Only relevant parameters are activated per input (<6% of parameters per token).

- Results in 2-4x lower FLOPS, lowers memory use, speeds training, reduces number of chips needed by 87% (DeepSeek uses 2,048 vs. industry average of 16,000+)

All four innovations are concerned with the AI training[1] phase, but only Mixture of Experts also relates to the AI inference[2] phase.

Multi-token Prediction

Tokens are predicted in parallel rather than sequentially.

- Requires fewer training tokens, enhances model performance

Mixed Precision Framework

Reserves higher precision (FP32) for key operations and uses a low-precision format (FP8) for compute-density operations.

- Reduces training time, reduces GPU memory usage, enables faster computation

~91% fewer **computing hours** needed and **10-40x** lower **total energy consumption** than other AI models

Observations

- Improvements in speed, cost, and performance reduce overall energy usage through efficiency.
- Speed, cost, and performance improvements made during the training phase are reflected in the inference phase upon model deployment.
- While model and architecture design improves energy efficiency, choice of hardware, cooling, and the energy source in data centers also affects total energy consumption (in both training and inference phases).

[1] AI training is the initial model development phase where it learns to recognize patterns through large datasets and intensive computation. [2] AI inference is the deployment phase in which a model makes real-time predictions or decisions. Lower compute power is needed, but user proximity (low latency) is required. [3] MOE minimizes performance degradation. [4] KV cache is the stored token information that models use to respond to queries (greater cache = greater memory use).

Sources: DeepSeek, DeepSeek V3 Technical Report (2024); Bain & Company, DeepSeek: A Game Changer in AI Efficiency? (2025); Ji et al., Towards Economical Inference: Enabling DeepSeek's Multi-Head Latent Attention in Any Transformer-based LLMs (2025).

AI solutions for the energy grid are moving into the mainstream, with a growing number of startups gaining significant traction

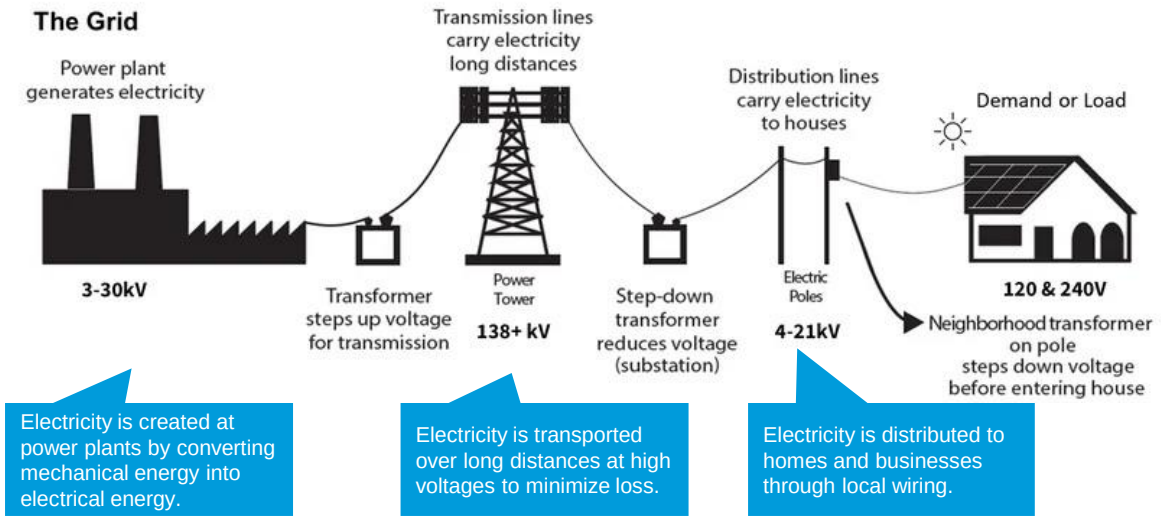
NON-EXHAUSTIVE

Category	Subcategory	Specific AI-powered technology	Description	Market adoption (*niche, growing, mainstream*)	Relative investment needs (*low, med., high*)	Comparative quadrant[1]	Example
Renewable Energy Generation & Storage	**Renewables Optimization & Management**	AI-powered solar and wind forecasting	AI models process hyperlocal weather data, satellite imagery, and historical plant performance to generate **highly accurate forecasts of energy production.**	Mainstream	Low to med.	**Scalable solutions**	tomorrow.io
	BESS Optimization & Trading	AI-based battery dispatch algorithm	The AI analyzes real-time electricity market prices, grid demand forecasts, and the battery's state of charge to **execute the most profitable charge/discharge strategy.**	Growing to Mainstream	Variable/ Scalable [2]	**Scalable solutions**	TESLA
Grid Operations, Transmission & Distribution	**Grid Enhancement & Infrastructure Build-out**	AI-assisted grid planning and digital twins	AI creates digital models of the grid to **simulate the impact of new assets** (like EV charging hubs) before they are built.	Growing	High	Strategic bets	SIEMENS *Gridscale X*
	Advanced Weather Intelligence	Hyperlocal threat forecasting	**AI provides granular weather forecasts** specifically for utility service territories, predicting the risk of ice storms on lines, high winds, or wildfire conditions.	Mainstream	Low to med.	**Scalable aolutions**	Climavision
	Fault Detection & Response	AI-powered fault analysis (including FLISR)	AI analyzes data from smart meters and line sensors to instantly **pinpoint the location and type of a fault** and can automate grid switching to restore power.	Growing to mainstream	High	Foundational platforms	Schneider Electric *EcoStruxure Grid*
Energy Efficiency	**Commercial & Industrial Energy Management**	AI-optimized building energy management systems (BEMS)	The AI learns a building's unique thermal properties and occupancy patterns to **intelligently control HVAC and lighting, minimizing energy use.**	Mainstream	Medium	**Scalable solutions**	GridPoint
Energy Consumption & Management	**Demand Response & Load Flexibility**	AI-aggregated demand response (VPPs)	AI platforms aggregate thousands of consumer devices into a virtual power plant. When the grid is stressed, the AI sends signals to these devices to slightly **reduce their load in a coordinated way.**	Growing to Mainstream	Variable / Scalable [2]	**Scalable Solutions**	voltus
	Energy Trading & Forecasting	AI for price & load forecasting	ML models analyze vast historical datasets, weather patterns, and economic indicators to **predict wholesale electricity prices and regional demand.**	Mainstream	High	Foundational Platforms	Gridmatic

[1] Strategic Bet = High Investment, Low/Growing Adoption; Foundational Platform = High Investment, Mainstream Adoption; Scalable Solutions = Low/Medium Investment, Mainstream Adoption.
[2] Investment depends on project scale. | Sources: CBInsights, The grid tech market map (2024); EnergyEvolution, Challenges in Implementing Smart Grid Tech (2024).
Credit: Yosafat Partogi, Hyae Ryung Kim, and Gernot Wagner. Share with attribution: Kim et al., "Powering Data" (23 January 2026).

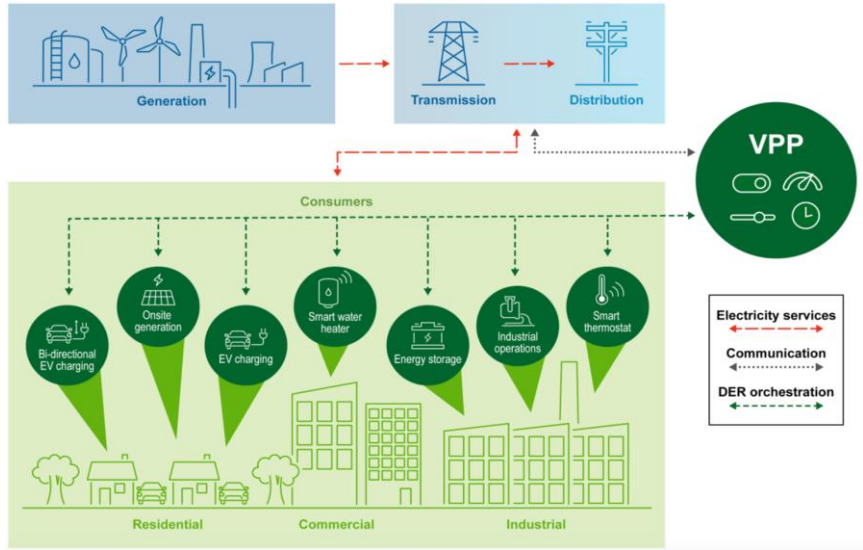
VPPs can adjust energy supply in real time, prioritize renewable sources, and cut reliance on carbon-intensive power generation

With electrical grids facing several challenges…



- Aging infrastructure leads to more power outages.
- **Climate change puts more stress on the grid** due to extreme weather.
- Energy demand growth leads to a traditional grid's inability to handle new loads and integrate renewables.
- **Distributed energy resources (DERs) are becoming more prevalent** given increasing affordability, reliability concerns with traditional grids, and the push for cleaner energy (i.e., solar panels, EVs, battery storage).

…AI solutions like virtual power plants are a potential answer



- VPP tech **aggregates, controls, and optimizes DERs from consumers.**
  - AI-powered solutions can analyze real-time data to optimize energy distribution, predict energy production, and anticipate energy consumption patterns for proactive management of energy resources.
- This in turn **increases grid resilience and reliability,** as excess or surplus energy created from DERs can be redistributed back to the grid.
- The solution **increases electrification and relieves load stress at the grid** at low cost.

Sources: Data and graphics from DOE, Pathways to Commercial Liftoff: Virtual Power Plants (2025).
Credit: Yosafat Partogi, Hyae Ryung Kim, and Gernot Wagner. Share with attribution: Kim et al., "Powering Data" (23 January 2026).

Columbia Business School
Climate Knowledge Initiative

AI Data Center Team

Hyae Ryung (Helen) Kim
PhD in Sustainable Development
Senior Research Fellow, Climate Knowledge Initiative

Ariela Farchi Behar
Master of Sustainability Management
Staff Associate, Climate Knowledge Initiative

Isabel Hoyos
Master of Sustainability Management
Senior Staff Associate, Climate Knowledge Initiative

Una Oljaca
BA in Sustainable Development
Fellow, Climate Knowledge Initiative

Shubhangi Prasad
Master of Public Administration
Fellow, Climate Knowledge Initiative

Yosafat Partogi Simbolon
Master of Business Administration
Fellow, Climate Knowledge Initiative

Clara Zibell
BA in Sustainable Development
Fellow, Climate Knowledge Initiative

Gernot Wagner
Senior Lecturer, Columbia Business School
Faculty Director, Climate Knowledge Initiative
gwagner@columbia.edu

Attribution: Kim *et al*., “Powering Data,” Columbia Business School Climate Knowledge Initiative (23 January 2026).

Columbia Business School
Climate Knowledge Initiative

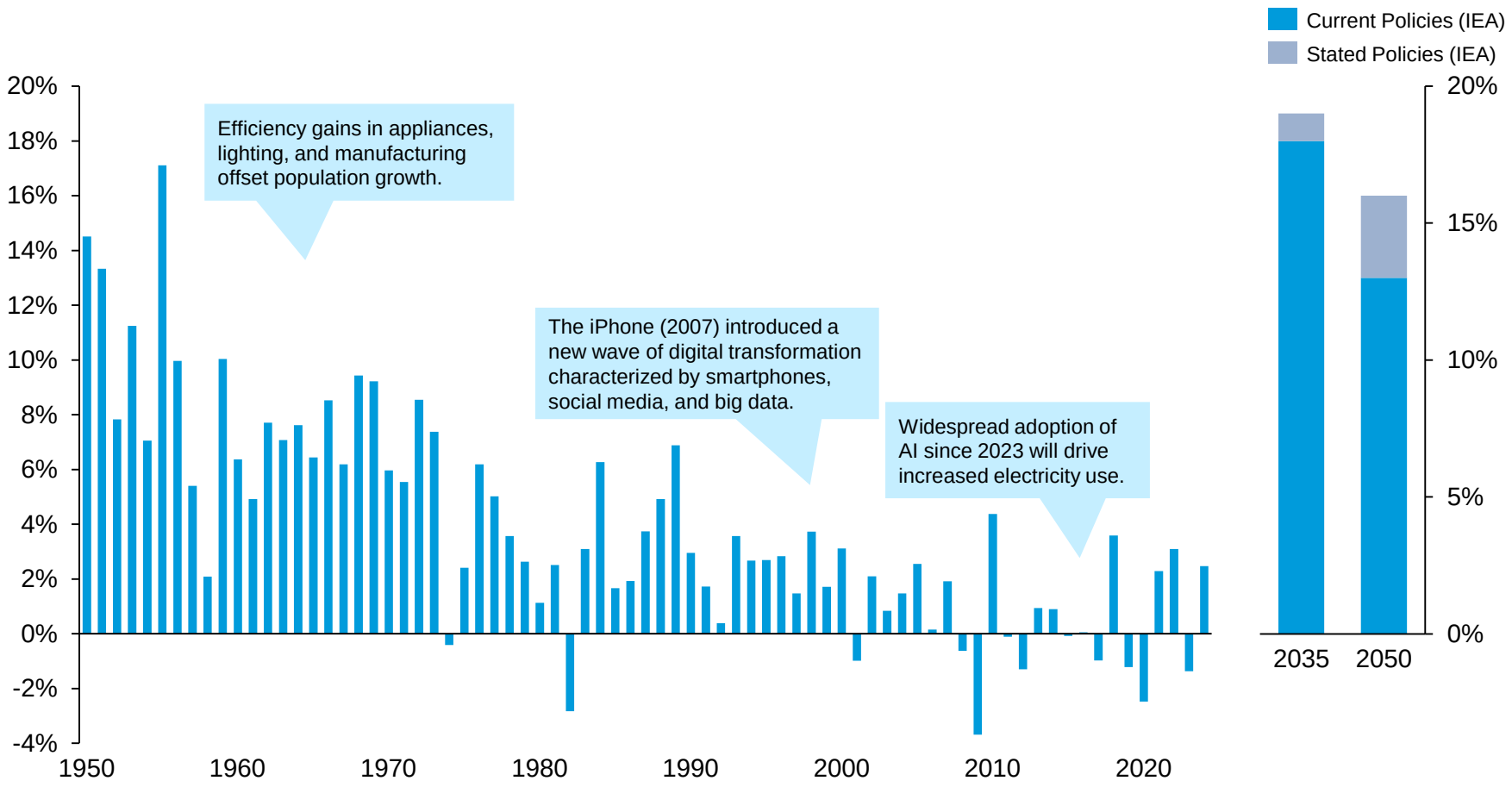
Appendix

Glossary

AI	Artificial intelligence
BTM	Behind the meter
CPU	Central processing unit
DC	Data center
KV	Key-value
FLOPS	Floating point operations per second
PPA	Power purchase agreement
PV	Photovoltaic
MOE	Mixture of Experts
MHLA	Multi-head Latent Attention
MPF	Mixed Precision Framework
MW	Megawatt
LLM	Large language model
EGS	Enhanced geothermal systems
EU	European Union
SOE	State-owned enterprise
GPU	Graphics processing unit
SMR	Small modular reactors (nuclear)
TPU	Tensor processing unit
PUE	Power usage effectiveness

AI boom expected to reverse decades-long downward trend in U.S. electricity growth

Annual growth in U.S. electricity use (1950-2024) and projected growth by 2035 and 2050, %



Observations

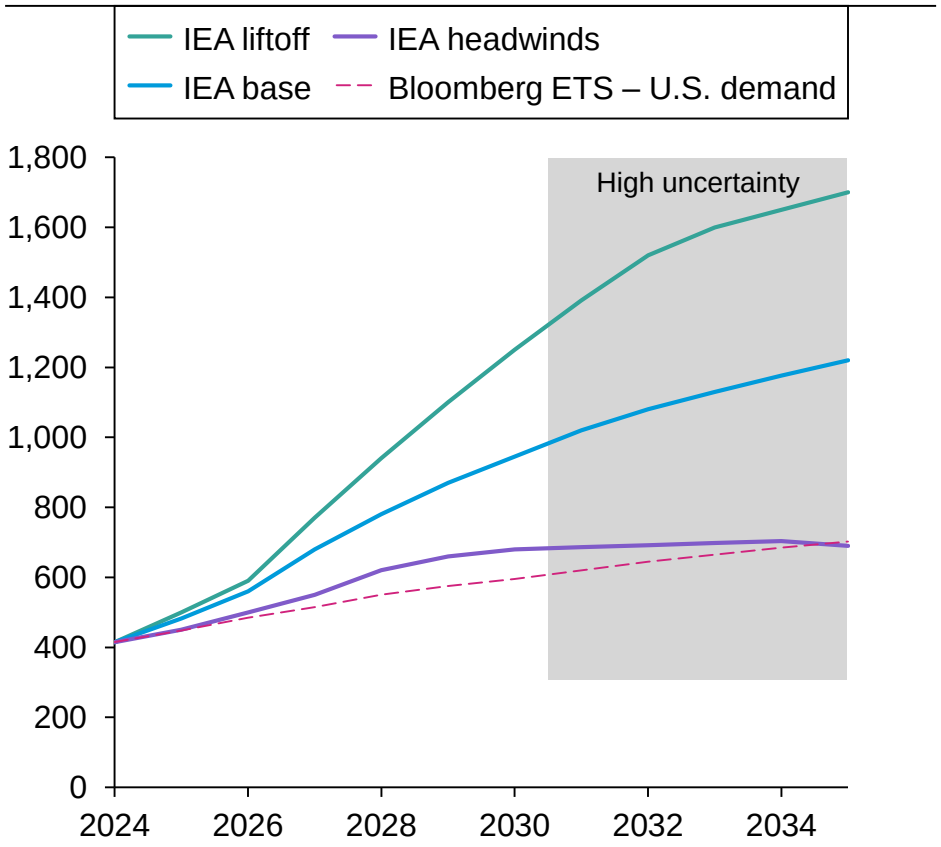
- Total U.S. electricity use was **~4,100 TWh in 2024**, and is expected to increase to **~4,900 and ~5,600 by 2035 and 2050**, respectively, under the IEA Stated Policies scenario.
- Total U.S. population has grown steadily since 1950, from around **150 to 350 million people**.
- Electricity demand from **data centers** in the U.S. is projected to increase by **240 TWh (+130%) by 2030.**

Sources: EIA, Total Energy (2025); IEA, World Energy Outlook (2025); Statista, Resident population of the United States from 1950 to 2025 (2025).
Credit: Isabel Hoyos and Gernot Wagner. Share with attribution: Kim et al., "Powering Data" (23 January 2026).

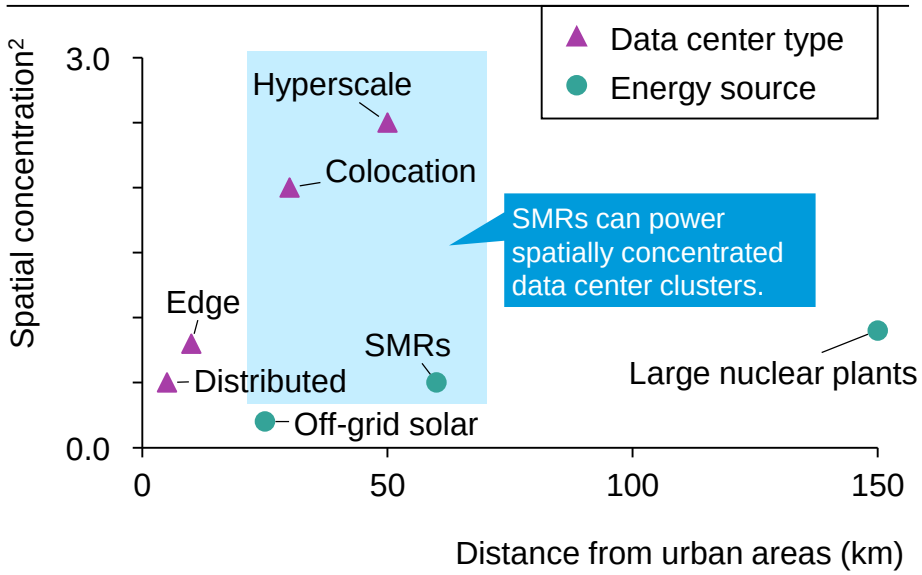
Columbia Business School
Climate Knowledge Initiative

Off-grid SMRs are a mid- to long-term solution for increased energy demand from hyperscale and colocation data centers

Projected global data center energy demand[1], TWh



Urban proximity and spatial concentration[2]



Data center types

	Training AI	Inference AI
Hyperscale	Primary	Common
Colocation	Limited	Common
Edge	Not suitable	Primary
Distributed	Not suitable	Primary

Observations

- Long timelines for transmission connection, gas turbines (more than seven years), and power supply stress the need for **off-grid solutions.**
- **Off-grid solar is a strong option for training-only data centers,** when uptime requirements are slightly relaxed.
- The **small size and modularity of SMRs** allow these to be sited closer to data centers. This is ideal for **inference AI** workloads, which have **low-latency and high-redundancy** requirements.
- The data center construction timeline is up to seven years in the United States. SMR technology is expected to be deployed after 2030.

[1] IEA Scenarios are based on projected AI uptake, energy sector bottlenecks, and efficiency improvements. [2] Spatial concentration = electrical output/land area used.
Sources: IEA, Energy and AI (2025); IEA, Electricity (2025); Baranko et al., Fast, scalable, clean, and cheap enough (2024); Ember, Global Electricity Review (2025); BNEF, How AI Influences US Data Center Power Demand (2025); Gartner, Power Shortages Will Restrict 40% of AI Data Centers by 2027 (2024); Last Energy, On-Site Nuclear Power (2023).

BTM geothermal is a potential low-carbon baseload energy source, can play key role in meeting growing U.S. power demand

Geothermal satisfies key baseload energy needs, which is critical as demand grows

1 High-capacity factor

~70-90% (can exceed 90% for next-generation geothermal power), higher than all power sources except nuclear

2 System inertia

Rotating turbines spin continuously due to high inertia, building resistance to frequency changes and helping maintain grid stability

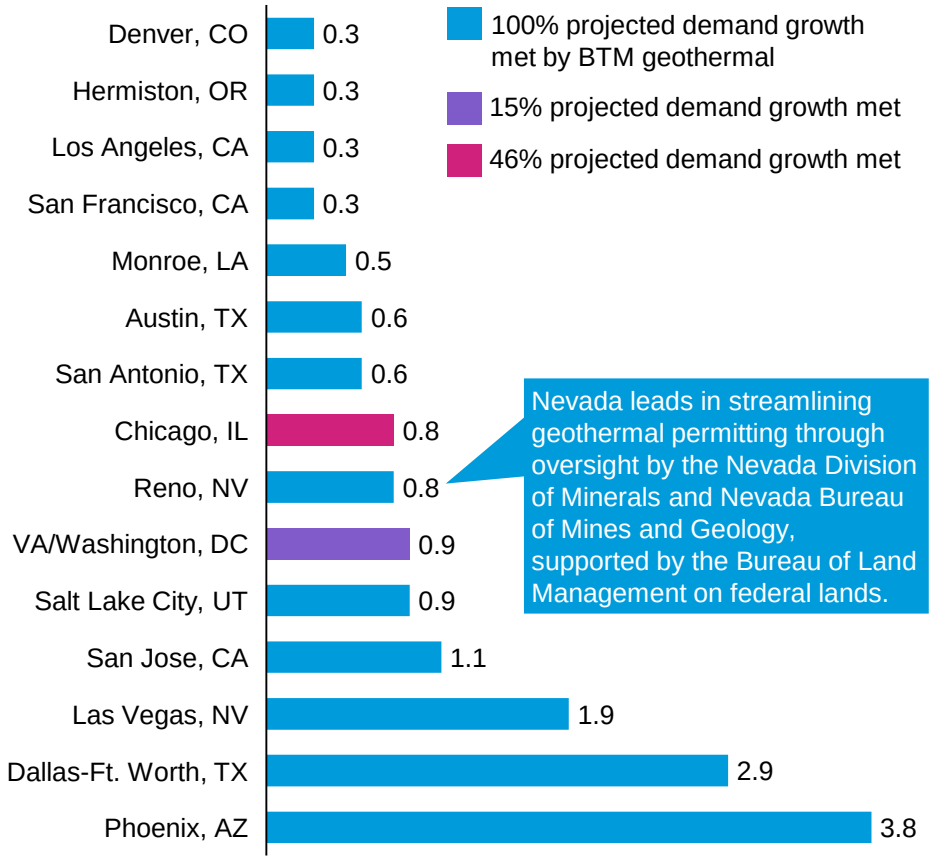
3 Extended service lifetime

Lifespan of 30-50+ years compared to 30-35 years for solar PV, 30 for wind turbines,10-40 for coal plants, and 20-40 for other fossil fuel plants

4 Sustainable operation

Lowest land use (**7.5 km2/TWh/year**) among renewables; low lifecycle GHG emissions – **37 gCO_2e/kWh** vs. 400-1,000 **gCO_2e/kWh** for non-renewables

Behind-the-meter EGS potential for hyperscale data centers in 15 largest U.S. markets[1], GW



Observations

- EGS potential is based on subsurface conditions and geothermal costs.
- **55-64% of projected hyperscale demand growth** could be met with geothermal, representing **15-17 GW of new capacity.**
- While much of the potential is in the West, only **Atlanta** and **New York City** show promise for BTM EGS.
- This scenario estimates a **national weighted average LCOE** for new geothermal capacity of **$78-$85/MWh.**
- If developers site data centers near high geothermal potential areas, EGS could meet all projected load growth by 2035 with a **31-45% lower LCOE.**
- **Direct cooling via GHPs** could further reduce data center electric loads using little water consumption.
- Streamlined permitting, better data and supply chains, and federal incentives are needed to make this shift possible.

[1] Assumes ongoing clustering of data centers, willingness to pay a 20% green premium above the regional electricity rate, and use of dry-cooled EGS and optional solar and BESS.
Sources: Rhodium Group, The Potential for Geothermal Energy to Meet Growing Data Center Electricity Demand (2025); SLB, Beyond levelized cost (2025).