22-Oct-2024 | 14:36 EDT

Data Centers: Surging Demand Will Benefit And Test The U.S. Power Sector

Topic: Digital Infrastructure & Innovation

Region: Americas, EMEA, APAC

View Analyst Contact Information

Key takeaways

- S&P Global Ratings estimates incremental U.S. power demand from data centers could be 150-250 terawatt hours (TWh) between 2024 and 2030.
- That looming demand has taken the power sector by surprise and will require about 50 gigawatts (GW) of new generation capacity through 2030--necessitating about $60 billion of investment in generation and $15 billion in transmission .
- Grid infrastructure will be the biggest hurdle to power supply growth, notably because of long planning and permitting processes.
- Tight power market capacity will support higher prices for longer, and we now view owning generation assets as credit quality positive.

Significant incremental demand for power from data centers will affect power market dynamics and should have positive credit quality implications for companies in the merchant power sector.

Why It Matters: The fast-increasing data collection combined with the rise of cloud services and AI result in a rapid and significant need for new data centers. These are energy intensive and demand significant additional power generation capacity and infrastructure. We view access to power as the most important variable on whether--and how--this growth in data centers will be shaped.

What We Think And Why: In an already tight power market, additional demand will result in tighter supply and higher power prices through the end of the decade. While green energy is favored, we believe this trend will support earnings growth and visibility for all power generators as more long-term contracts are signed. We view this as credit positive for the competitive power markets.

Data Center Demand Is Changing The Power Sector Landscape

Two pictures taken 13 years apart at the Fifth Avenue annual parade in New York City illustrate wisdom for the ages. The first picture taken in 1900 had only one car. By 1913, the parade had only one horse. In 13 more years, the automobile disrupted the horse, which had been used as the main mode of transportation and freight for centuries. While the horse population continued to grow at a diminishing rate through 1920, it started dwindling rapidly between 1920 and 1940.

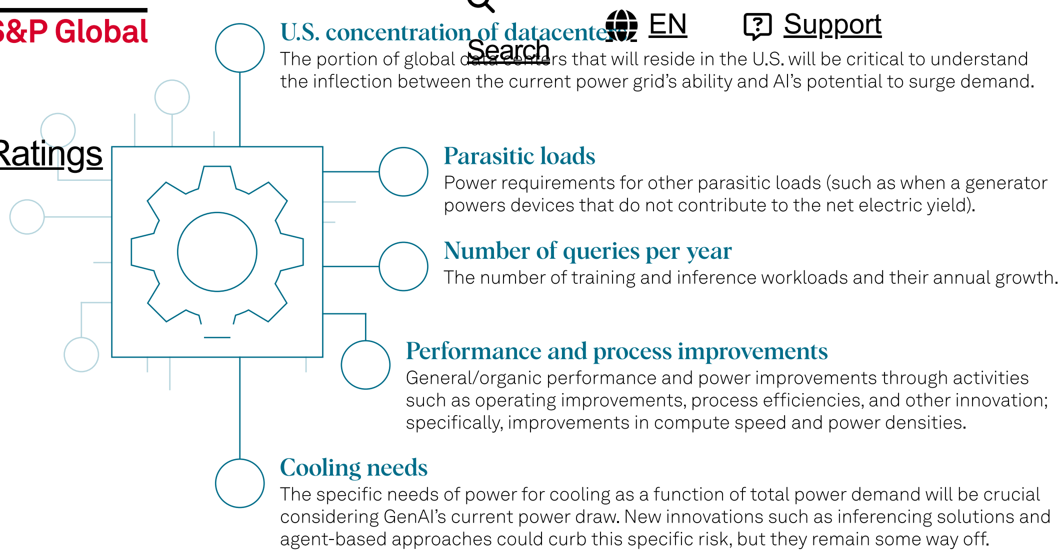
Disruption is a bit like that--you don't see it until it's staring you in the face.

The term "fourth industrial revolution" has been bandied about in the past year. The theory goes that industrial revolutions ensue when hugely disruptive technologies are introduced and swiftly adopted. The previous three included mechanization in the 18th century, electrification in the 19th, and digitization in the 20th. This fourth chapter is characterized by AI and intelligent automation. But now, the voracious need for energy has sucked the power sector into the maelstrom.

With the largest fuel switch in its history, the U.S. power sector was undergoing its own disruption. Now, that need for sustainable power generation is meeting a staggering increase in demand for generation and transmission expansion due to rapid growth of large loads from manufacturing onshoring, heating and transport electrification, and AI-driven data centers.

In this commentary, we discuss this dramatic explosion in power demand. First, we discuss how this incremental demand on the grid could affect power market dynamics. Then we discuss how this environment is affecting the credit quality of companies in the merchant power sector. Importantly, in the appendix, we present our top-down buildout of power demand from data centers through 2030 and our approach to data center demand expectations.

Five key factors influencing U.S. power demand resulting from generative AI



Source: S&P Global Ratings.

The Demand Surge Will Be Uneven

In January 2024, the Pennsylvania-New Jersey-Maryland (PJM) Interconnection, the biggest independent system operator (ISO) in the U.S., increased its demand forecast to a compound annual growth rate (CAGR) of 1.7% through 2030, up from 0.8% the previous January. The news did not grab our attention because we are used to industry revising forecasts and tweaking them later. But then, forecast revisions started accelerating. By June 2024, our affiliate, S&P Global Commodity Insights, revised its power CAGR for the contiguous U.S states to 2.1% for 2024-2030 from a CAGR of 1.2% as recently as January 2024.

S&P Global Commodity Insights' growth forecasts are not uniform. It projects the strongest growth through 2030 in PJM South (7.4% per year), the Electric Reliability Council of Texas (ERCOT) (2.9%), and Southeast Regional Council (3.0%). By contrast, S&P Global Ratings' growth assumptions are lower because we assume meaningful infrastructure constraints.

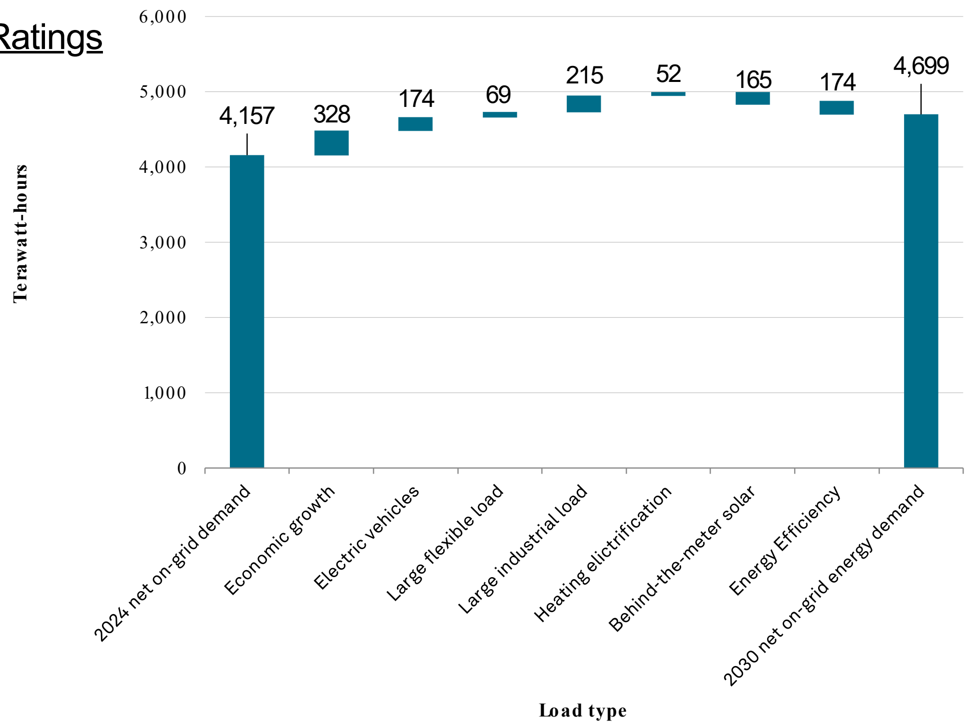
Why are power markets focusing on data center growth?

Compounding is deceptive; it hides growth in a manner humans do not intuitively comprehend. This incremental differential in CAGR over a six-month period is a significant revision, especially when compared with demand that had stagnated. Over the past decade, the increase in demand from customer growth and electrification needs was offset by energy efficiency (e.g., building codes and efficient appliances) and behind the meter solar installations.

This growth surge is spurred by large loads (see chart 1). Electric vehicles and large flexible loads (e.g., electrolytic hydrogen and industrial processes) will drive significant demand between 2030 and 2050. However, over the next few years through 2030, much of the incremental demand will come from data centers. Data center-related load may make up 20%-30% of all net new demand added between any two consecutive years until 2030. That is why we are focusing on this demand--data centers will influence near-term power demand growth most.

Chart 1

U.S. power demand through 2030



Sources: S&P Commodities Insights and S&P Global Ratings.

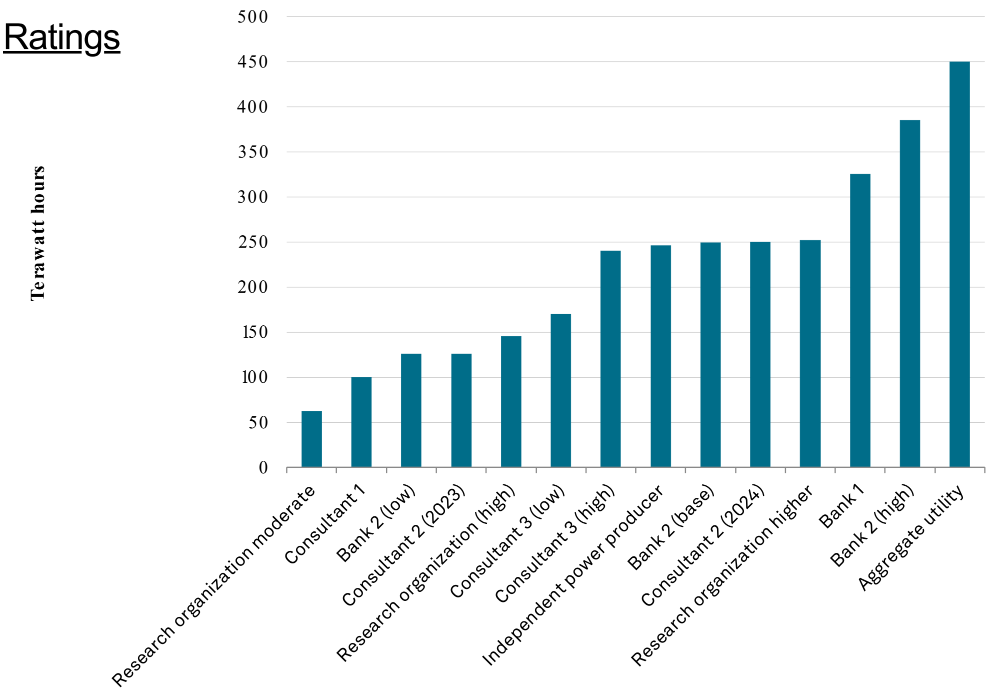
The demand estimate is staggering but could end up lower due to infrastructure constraints

There are many diverging expert opinions regarding the power industry's expected growth rate (see chart 2). From data centers specifically, aggregate U.S. estimates for incremental power span 100 TWh-300 TWh through 2030, with some outlier estimates that go even higher. We present our incremental demand estimate from data centers in table 1 in the Appendix.

With demand at about 170 TWh through 2024 and 170 TWh-250 TWh incremental through 2030, data centers will comprise 7.5%-8.75% of U.S. power demand by 2030 (S&P Global Commodity Insights estimates aggregate U.S. power demand in 2030 of about 4,700 TWh). For perspective, the annual consumption of New York State and California is 150 TWh and 250 TWh, respectively.

Chart 2

Estimates for new U.S. data center demand from 2023-2030



Source: S&P Commodities Insights.

This range is dispersed widely for two reasons:

- There are many variables involved, and these variables could take on a range of values.
- Some forecasts are unconstrained, while others are tempered by physical and logistical constraints, the biggest one being power availability.

What Does This Dramatic Growth Mean For The Power Sector?

Meeting generation demand cannot be done with renewables alone

The continually increasing need for reliable, uninterruptible power for computing speeds required for AI workloads is not helpful for the sustainability story. Even as a meaningful amount of renewable power is being placed on the grid, we expect there will be significant need for natural gas-fired generation if this demand is to be addressed. Not only is gas generation needed to meet part of the demand, but dispatchable gas resources will also serve a critical balancing role as more renewable capacity enters the market. If the world wants all workloads to be powered only by sustainable power in 2030 from currently available technologies, it will have to temper its AI initiative.

We believe all power generators with legacy assets benefit from increased AI power use. However, coal-fired generation will continue to decline even as some assets may have longer lives. Generators that produce renewable power will benefit if they can place generation on the grid quickly. Here, their ability to get sited and secure equipment, as well as placement on the interconnection queue is key. They have an advantage over new fossil capacity because recently issued U.S. Environmental Protection Agency (EPA) rules place significant costs on any new fossil-fired generation capacity (see box 1).

Box 1: Recent emission regulations for new combined-cycle gas turbine (CCGT) builds

Investors are reluctant to finance new builds because of long construction duration and difficulty in raising capital for new thermal assets. In particular, the EPA's ruling earlier this year is a significant overhang. On April 25, 2024, the EPA finalized new source performance standards (NSPS) for new gas-fired plants (i.e., those with capacity factors greater than 40%). These plants now must meet detailed NSPS requirements based on "high-efficiency" combined-cycle combustion technology (CCGTs) immediately upon startup and then install carbon capture and sequestration (CCS) by Jan. 1, 2032. The CCS equipment must capture 90% of all carbon-dioxide emissions. This will inevitably reduce the chance that new base-load gas plants can be built. Opponents of the EPA power plant rule are asking the Supreme Court to freeze the regulation as courts decide whether the agency overstepped its authority.

We expect these new requirements to present significant headwinds for new gas-fired plants. However, we see the potential for extensive litigation, especially after the Supreme Court recently struck down the Chevron doctrine. Regardless, this poses the potential for further delays or incremental capital costs for baseload gas-fired builds. Estimates for new combined-cycle, gas-fired generation now top $1,500/kW from $1,000/kW as recently as 2019.

Any higher cost for new entrants means existing generation will earn higher net revenues for the same capacity, all else equal. That is why companies with a large generation footprint are seeing significant tailwinds to their stock prices this year.

Based on simplified assumptions that include 50% of this incremental capacity coming from natural-gas-fired units (40% CCGT at 70% capacity factor [CF], 10% peakers at 9% CF), 20% from wind (40% CF), and 30% co-located solar/battery solutions (35% CF), the grid would need about 50 GW of incremental generation supply at about $60 billion of investments just to address the data center demand through 2030.

We expect the largest deployment of gas-fired generation in regions like ERCOT, but the PJM Interconnection also cannot support growth without gas-fired generation. ERCOT has already set up its Texas Energy Fund that is providing lower-cost loans for new gas-fired generation. The PJM may contemplate a similar program.

U.S. solar and wind projects had average capacity factors of about 25% and 38.0%, respectively, in 2022 (the latest yearly data reported by the Energy Information Administration).

Transmission is the biggest constraint

Unlike Europe, the U.S. does not have an energy problem, but it does have an energy infrastructure problem.

Northern Virginia (NoVA), in utility Dominion Energy Inc.'s territory, currently has more data centers than the next five biggest markets in the U.S. combined. To address this area's long-term needs, the PJM Interconnection opened a competitive auction for transmission projects. It garnered about $5 billion of transmission projects to support about 7.5 GW of upgrades, split between several utilities.

Not all regional transmission organizations need the same level of investment as PJM, but if we extrapolate the results of the PJM Interconnection's transmission auction as the potential investment needed for the transmission grid, the estimate implies that the grid would need about $15 billion of transmission capital expenditure through 2030 to support growth.

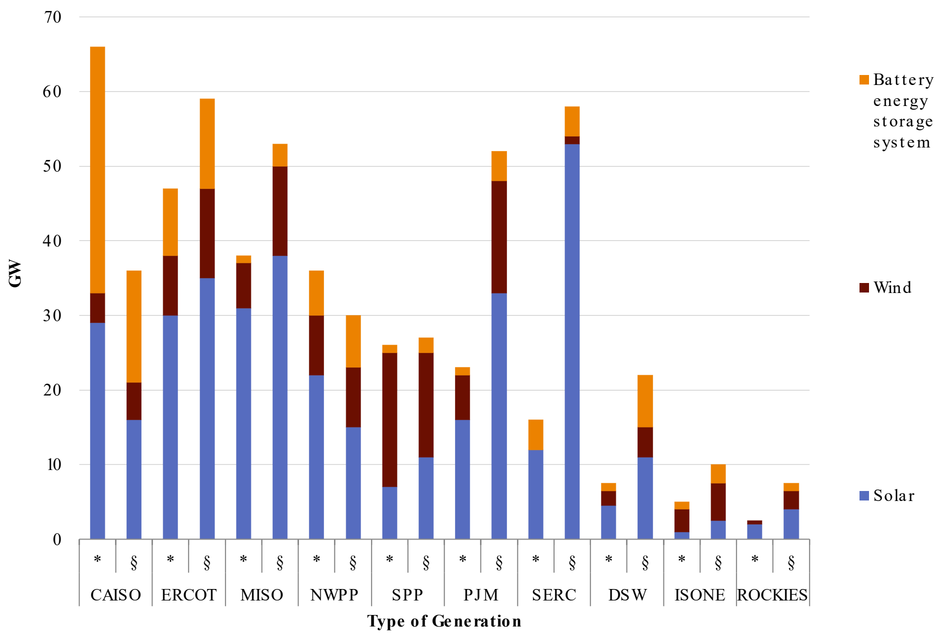
Given poor results for recent new builds and weak economics, there have been no active interconnection requests for natural gas plants since 2021, a first since 1997. Besides, the expansion of transmission infrastructure assets is a long-term planning process that requires permitting and siting and is typically done at a measured pace. It requires regulatory approvals that often take many filings and considerable time (see charts 3 and 4). Moreover, reliance on foreign suppliers, lingering COVID-19 pandemic-related shortages, and insufficient domestic manufacturing capacity all contribute to delays in transmission project development.

We expect the slower pace of transmission infrastructure upgrades cannot fully meet all the immediate growth needs of the high-tech industry, thereby limiting its growth rate. Market estimates in Chart 2 vary widely not because of the unconstrained (i.e., aspirational) power demand, but because of the interconnection constraint.

Recognizing this dire need, in March 2023, the GRID Power Act was introduced in the House of Representatives. If approved, U.S. grid operators could fast-track their review of power plants in their interconnection queues if the plants will improve grid reliability. Under the Act, the Federal Energy Regulatory Commission would have 60 days to review proposals from regional transmission organizations (RTOs) and ISOs for specific projects that would be pushed to the head of interconnection queues. The RTOs and ISOs would have to show the proposed projects bolster grid reliability, address power supply shortfalls from retiring power plants, or meet growing power demand.

Chart 3

2030 outlook versus capacity with signed IAs

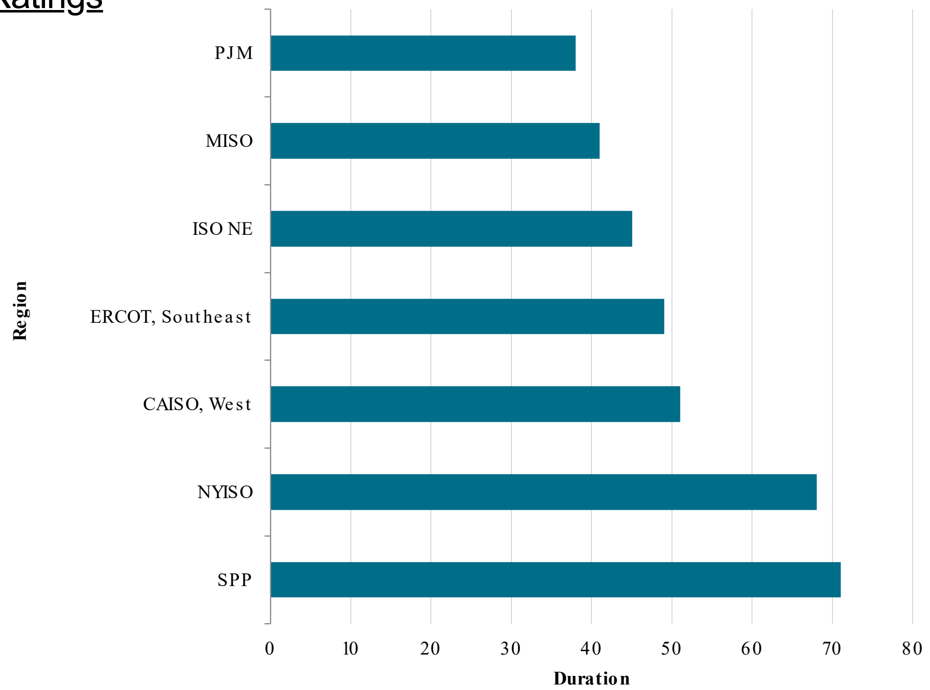


*Signed IA. §Adds by 2030. Source: S&P Commodities Insights; IA-Interconnection Agreement

Chart 4

Interconnection queue by region

Interconnection queues could delay data center buildouts



As of April 2024. Duration--Average time from queue data to proposed online date. Source: S&P Market Intelligence.

Location matters, and power delivery will follow

There are a few key factors that drive data center siting--fiber access, water (and climate) for cooling, availability of workforce, and cheap energy. However, grid connectivity and transmission access are the chief drivers.

As we mentioned above, Tier 1 markets such as NoVA are facing power transmission and distribution constraints due to the age of the grid, long interconnection queues, and shortages in land and material supply to build out transmission infrastructure. As a result, we expect most growth will predominantly be in the Southeast, Midwest, and Western U.S.

The highest concentration will still be around the Washington D.C. metro area because of historical reasons. Initial activities commenced near the metropolitan area exchange (MAE)-East area. MAE-East was the first internet exchange point, beginning with four locations in Washington, D.C. in 1992, then expanding to three Virginia locations: Vienna, Reston, and Ashburn.

It is not a coincidence that the largest concentration of data centers is in Ashburn, NoVA. NoVA has 40% of data center capacity in the U.S., but faces constraints because there are significant delays in grid connectivity at Ashburn. The delays are due to bottlenecks in Dominion Energy's transmission infrastructure, which will not be meaningfully upgraded until 2026. A similar MAE-West was set up in the South San Francisco Bay Area. MAE-West has limited ability for incremental baseload power and is therefore not a good site for data centers.

Because of these interconnection issues, new data center capacity is coming up in atypical locations due to the latency agnostic nature of training workloads and the availability of renewable power. Several co-location power providers are following them to these atypical locations and building generation facilities to help hyperscalers--the largest cloud computing providers--accelerate time to market. We see the prospects of renewable power-backed data centers in such computing locations.

However, many inferencing data centers that are low latency are strategically being placed near large population centers (edge computing) to support the use of multiple applications, such as AI for autonomous vehicles, robotic surgery, or fraud detection. These require reliable power.

Hyperscalers are now building out enormous amounts of data center capacity in Iowa, Idaho, North Dakota, and West Texas, partially to provide capacity for cloud services and training. To serve California, developers of location-agnostic data centers are building in Nevada and Arizona, where land and electricity are cheaper.

How Are The Power Markets Responding?

The nature of contracting is evolving

Data centers that are supplied power by renewable counterparties are all front-of-the-meter (FTM); that is, connected to the grid installations, as opposed to ones that take the generation behind-the-meter (BTM); that is, inside the fence, or off-grid).

Currently, technologies lag aspirations: renewables co-located with battery solutions do not and will not provide 24/7 power for several years. A renewable energy center that co-locates with a data center will provide up to about 70% of the data centers' need directly. The balance will come from the grid. A renewable complex that is a combination of solar, wind, and storage portfolio could provide a high level of firmness, but no data center would risk it unless it is a latency-agnostic training data center.

Between cheap, firm, and clean power, current technologies allow any two. The implications are below:

- With time to market as the primary driver, the most desired power source is existing nuclear generation--it provides firm and clean power at scale.
- Next comes renewable power because it delivers decarbonized energy. With co-located renewable generation, a data center will need access to the grid and therefore must be located near liquid hubs where the developer/aggregator is confident it can procure clean power from the grid.
- Then comes gas-fired generation. It is dispatchable (uninterrupted) but not clean. Many data centers' sponsors will simply ignore the sustainable part.

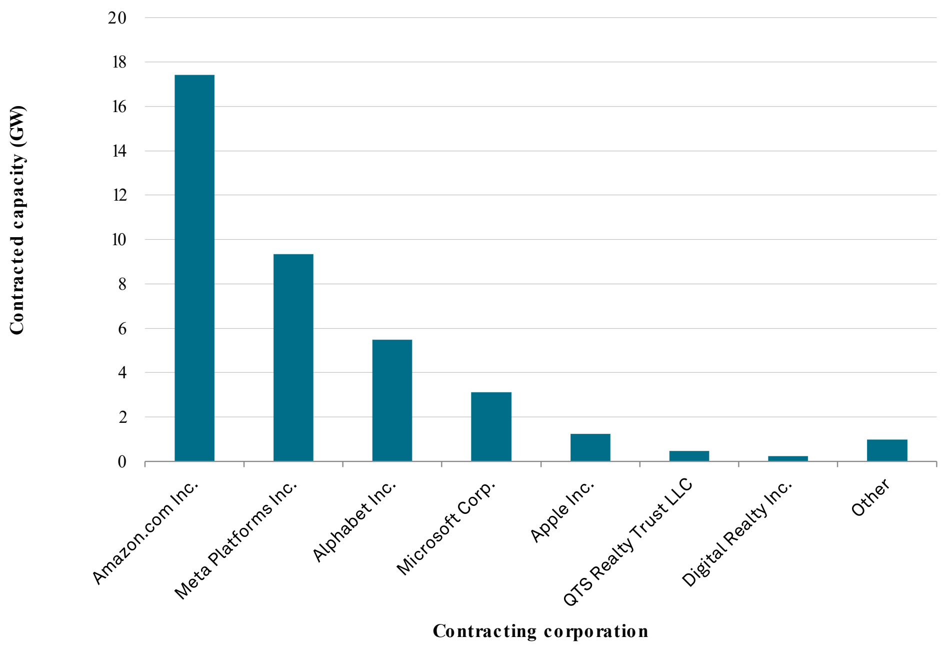
Hyperscalers have aggressive sustainability targets (e.g., Google has a target of carbon-free 24/7 energy by 2030) and primarily rely on the grid to meet their demand. However, the grid still cannot reduce its carbon intensity at the rate the hyperscalers desire. As a result, most hyperscalers and data center companies are using power purchase agreements (PPAs) or virtual PPA (VPPA) contracts to meet their renewable targets.

Depending on where their data centers are being set up, some hyperscalers like Microsoft focus primarily on VPPAs, while others like Google procure substantial electricity needs through physical PPAs.

Chart 5

U.S.-based renewable capacity contracted by technology and web services companies

Hyperscalers have been active in contracting renewables capacity



As of February 2024. GW--Gigawatts. Source: S&P Commodities Insights.

In a VPPA (see box 2) the green wrap could be in the form of an unbundled renewable energy certificate (REC) or could include "regulated green tariffs" or "competitive green power products" (i.e., bundled electricity and RECs from utilities in regulated markets or bundled electricity and RECs from retail providers in deregulated markets). The focus has shifted recently to VPPAs due to a larger, diverse pool of available power in deregulated markets and growing limitations on the number of PPAs in regulated power markets.

Box 2: The mechanics of a PPA contract

When a data center company decides to pursue a contract, it either enters a physical PPA or a VPPA. With a physical PPA, the hyperscaler (or data center company) takes title to the physical energy at a specified delivery point on the electric grid. The physical energy can then be transmitted from that specified delivery point to the data center's meter.

In these contracts, the data center guarantees that the power provider will receive a fixed price for their energy. In exchange, the buyer receives RECs for every MW hour (MWh) of clean energy that is generated and sold.

PPAs are long-term contracts, typically spanning 10-20 years, enabling the project developer to secure long-term financing and build the project. Once produced and placed on the grid network, power is fungible.

In a VPPA, the data center company may be serviced by a co-located renewable energy center for, say, 35%-50% of its needs. For the balance, the data center purchases "green wraps" with the power provider. So, with a VPPA, the energy does not physically flow from the power unit to the buyer. It is merely a financial contract, which is why it is also called a financial PPA. The energy is sold on the wholesale electricity market at a defined settlement location (node, trading hub, or load zone). The data center continues to get their electricity from the local utility company at the utility's rate.

A VPPA is settled financially as a fixed-for-floating swap or contract-for-differences. At the end of the predetermined settlement period, the floating market price vs. fixed VPPA price will be calculated and the power company or the data center will pay the difference, depending on whether it was higher or lower than the VPPA price. The power producer receives one REC for every MWh sold. Under most VPPAs, the RECs are immediately transferred to the buyer.

This means that while the market tends to think of all grid MW equally, from the perspective of a data center, clean and firm MW are valued much more. We will get into that next.

There is potential for more nuclear PPAs

While co-location involves the pairing of any type of power plant with any large load, nuclear plants are large (scale) and generate carbon-free power (clean). They are also often configured with at least another unit (redundancy) and capable of running for up to two years (uninterrupted). Put simply, nuclear units have the highest reliability and availability compared with other resources (see appendix 1).

On March 4, 2024, Talen Energy Corp. announced a data center transaction with Amazon Web Services (AWS) that co-locates its Susquehanna nuclear unit next to its data center and takes the nuclear plant inside the fence, or BTM, to serve the data center. This is the proverbial epiphany moment for the sector. The market recognized it as a gamechanger for the independent power producer (IPP) sector because it highlights the value of clean baseload generation in a growing demand environment.

The read-through for the sector was in the 10-year power supply contract with AWS' data center, and the incremental revenues paid by AWS for the emission-free energy certificates (EFECs). We note the MW involved are significant--AWS has minimum offtake commitments (120 MW) that could rise to 960 MW in stages. The purchase of the EFECs resoundingly underscores that hyperscalers value firm and clean generation. The fact the Susquehanna nuclear station is a dual-unit site and provides power supply redundancy that a data center would need makes the transaction extremely attractive to the hyperscaler.

Late last month, Constellation Energy Generation LLC announced a second nuclear power data center transaction. Unlike the Susquehanna transaction, the three-mile island plant (835 MW) transaction is an FTM, 20-year PPA with Microsoft, commencing in 2028, and the first run rate year is 2030. The contract is a unit contingent PPA with PJM West as the benchmark hub. Interestingly, this deal is unique because of additionality, as a retired resource is being recommissioned. Notably, the PPA excludes the transmission charges that Microsoft would pay separately.

The Credit Quality Of The Merchant Sector Benefits

A rising tide lifts all boats. We view the dramatic demand increase as a credit favorable for all merchant generation providers. Stock prices and credit quality of players in the broader sector are already benefitting. Even after factoring in upgrades in 2024, 20% of the credits still have positive outlooks. Conversely, only 5% of IPP ratings had negative outlooks as of June 2024 compared with 15% as of year-end 2021 and 24% in December 2020.

We expect capacity markets to remain tight, benefiting prices

In most parts of the U.S., capacity for generation supply is increasingly appreciated. We see that in the sharply elevated resource adequacy payments in California and in the last PJM Interconnection capacity market auction price, which increased 9x over prices set in the previous auction. The explanations are similar. Over the past years, increasing supply and demand imbalance has worked itself up given no material recent investments. However, a longer line of reasoning also includes long interconnection queues. There is also insufficient future transmission planning and now the problems have simply cascaded, resulting in several RTOs short on MW capacity for power delivery.

There just are not enough renewable MW that can be set up to accommodate the demand increase, especially after incorporating their effective load carrying capability (ELCC) accreditation impact. That also means a substantial amount of MWs of gas-fired generation is needed. There is both a willingness and ability problem. A developer must be willing to take a 35-year directional bet on gas-fired generation in an environment pushing toward decarbonization goals. Even if investors are somehow convinced to finance projects on the back of merchant energy revenues, new gas-fired entrants will also be delayed by the lengthy interconnection queue. In markets like the PJM, the generation queue is dominated by resources with low accredited capacity value.

In our view, the capacity shortfall cannot be met solely with renewables given limited contribution post-ELCC weighting. While demand response and energy efficiency have historically carried a lot of weight, we believe saturation points are fast approaching and may be challenged to keep up with rapid AI and data center-driven load growth. Consequently, we expect capacity prices to remain elevated compared with years past, benefiting all generators.

The December 2024 PJM auction has been postponed. Based on the parameters presented, we believe the prices set in the auction would have been higher--potentially topping $300/MW-day--and higher than our current assumptions for capacity year 2026-2027. While future capacity auction outcomes are yet to be set, bilateral capacity is contracting at high price levels in New York and California.

Owning baseload generation assets is now a credit positive

Unregulated, long generation companies are well positioned to respond to the demand. In the sector, we believe Constellation Energy and Vistra Corp. are best positioned to benefit. We expect Talen Energy Corp. and PSEG Power LLC to benefit as well, but to a lesser extent because of their smaller generation footprints. We also expect developers such as NextEra

Energy Inc., Brookfield Renewable Partners L.P., Clearway Energy Inc., and Pattern Energy Group L.P. to allocate significant capital to firming power. Pattern Energy's Sunzia project is well positioned in our view, because there can be no energy transition without transmission.

For example, we believe for every $1/MWh price move, Constellation's EBITDA would rise by $200 million, all else equal. This is about 4.5% of its expected EBITDA. Our calculations make a number of simplifying assumptions, like ignoring any dynamic hedging of its economic generation. We also ignore the fact that about 55 TWh of its generation is serving load at $32MWh-$34/MWh under Illinois' carbon mitigation credits plan (a price below the $40/MWh-$45/MWh production tax credit [PTC] floor before inflation adjustments).

Conversely, the New York units are already earning a higher price compared with the PTC through the state's zero-emission credits program. Therefore, prices in New York have to rise above $60/MWh before Constellation can benefit. However, these calculations illustrate that its nearly 180 TWh of clean, baseload nuclear generation is now seen differently because it is valued more for its 24/7 clean generation.

We have not assumed any hyperscaler contracts in our forecasts for Constellation but provide these details for context. If we assume a floor price of about $45/MWh (a generator would argue they are assured of that in the wholesale market through the PTC, and the hyperscale must match it) and convert $270/MW-day capacity price into $/MWh at 95% capacity factor ($12.5/MWh), then add ancillaries and a premium for clean attributes and long-term contracting, we believe these PPAs would be $75/MWh-$85/MWh.

The Three Mile Island transaction announcement is higher because it brings the added risk of bringing a decommissioned unit back into operations at a significant front-ended capital spend. Also, the longer the term, the higher the price would be.

A similar calculation for Vistra, which assumes about 75 TWh of nuclear and renewable generation (including Energy Harbor Corp.'s 30 TWh) and 105 TWh of gas-fired generation, also results in a $200 million-$210 million uplift for every $1/MWh increase in power prices. At current prices, we expect generation would also potentially increase, so the range could be a little wider. Our calculations assume about 25 TWh of coal-fired generation that the company has classified as "sunset" generation intended for retirement. Most of this retirement is mandated, but we believe some could eventually convert into natural gas units.

Baseload nuclear generation has premium value

While a four-hour battery and nuclear MW are considered similar in a capacity auction (adjusted for accreditation values), the market seems to have a preference. The two recent nuclear data center transactions and the sentiment that more are coming has laid that question to rest.

Not only does a nuclear unit expect higher capacity prices in future PJM auctions, it also now has options. Unregulated nuclear units could withdraw from PJM's capacity auction by participating in a fixed resource requirement (FRR) approach. This is assuming a load-serving entity initiates the withdrawal of load from the capacity auction and contracts with that nuclear generator bilaterally to include in its FRR plan (i.e., to fulfill the capacity obligation for its load via FRR). We believe it is reasonable to expect this. Because the data center will sit adjacent to the nuclear units, transmission and distribution grid charges are saved. As a result, the long-term PPA pricing the generator will negotiate will be higher than the wholesale power price it would otherwise attract. However, this PPA price will still be lower than the retail power price the hyperscaler would normally pay.

The more nuclear units commit to PPA transactions, the tighter the rest of the market becomes. Capacity prices for generation go up either way, whether it is lower supply due to co-location or more demand for grid.

Even if a state decides it does not want data centers, that state's customers could pay higher capacity prices and transmission costs if other PJM states make decisions to add data centers (see also box 2).

The ability to secure sites with grid interconnections and expertise in equipment procurement are key

Given how pervasive transmission issues are, we believe, at least for now, it is all about securing generation sites that come with grid interconnections infrastructure. The time it takes for new generation capacity to go from planning to commercial operations has increased because grid interconnections are harder to find.

We believe companies such as NextEra Energy are skilled in this respect because they have an early mover advantage in securing sites and advanced placement in interconnections queues. NextEra also has experience and expertise in constructing and commissioning renewables. We view AES Corp. as another company with significant advantages and believe it is well positioned for renewable development. NRG Energy and Vistra could also convert several older coal-fired sites that already have transmission access and brownfield infrastructure.

A growing need opens the door for new solutions

The prospects for higher demand are spurring innovation. Within the next five years, we expect data centers to become grid assets that can regulate energy consumption. We also see a place for small modular nuclear reactors (SMRs), hydrogen fuel cells, and self-sufficient microgrids, although they appear to be at scale in 2030 and beyond.

While it looked unlikely as recently as a year ago, with this data center push for 24/7 power, we now expect a new generation of SMRs and microreactors to drive modest nuclear growth in the U.S.

The sector continues to face obstacles, however, such as supply chain bottlenecks, financial challenges, and regulatory hurdles. To resolve this challenge, nuclear developers may pivot to collaboration with technology and data center operators/owners. In April, microreactor manufacturer Oklo Inc. signed a preliminary deal with data center operator Equinix, under which Oklo could provide up to 500 MW of nuclear power. There are about 30 current SMR projects in the U.S., ranging from early to advanced development.

Google recently announced that it is signing the world's first corporate agreement to purchase nuclear energy from multiple SMRs to be developed by Kairos Power. The initial phase of work is intended to bring Kairos Power's first SMR online by 2030, followed by additional reactor deployments through 2035. Overall, this deal will enable up to 500 MW of new 24/7, carbon-free power.

Because of growing investor queries, we will explore the topic of SMRs in a separate article soon.

Appendix 1

Box 3: Compute speeds and power density

Compute speed and power density are two distinct metrics, but they are also related. Compute speed refers to the amount of computational power a central/graphic processing unit (CPU/GPU) can deliver per second. Specifically, this refers to the floating-point operations per second undertaken by the server. It is a measure of raw processing capability.

Power density, however, refers to the amount of electrical power consumed per unit of volume (e.g., kilowatts per rack (kW/rack). It measures how much power is required to operate the server in a given space. A higher compute density usually requires more power, which can affect power density. However, discussing them together gives a more complete picture of a GPU's efficiency and performance characteristics.

Table 1

An estimate of potential U.S. incremental power needs from data centers in 2030		
Variable/input	**Values**	**Remarks**
Data center queries per day in 2030 (A)	1.56E+10	We assume current aggregate queries (inferences) across all search engines at 11 billion per day (we conservatively assume Baidu at 1 billion/day but it appears to have more than that). We also assume the inference rate grows 6% annually through 2030. A higher initial assumption at, say, 14 billion queries per day increases incremental power need to about 225 terawatt hours (TWh) from 175 TWh.
Average number of tokens in an inference reponse (B)	400	Tokens are pieces of words, which could be whole words or parts of words, depending on the complexity of the word. Text generation for writing or summarizing generate 200 to 600 tokens per inference. We took the average.
Floating point operations (FlOPs) per token (C)	6.5E+14	Depends on model size and sequence length. This estimate is for one GPT-3 inference with a standard sequence length. The estimate applies to the largest GPT-3 model (175B). This assumption varied across experts we spoke to and can change results meaningfully.
FLOPs per second server performance (D)	3.20E+16	An 8GPU H100 system provides 32 petaFLOPS of deep learning compute performance.
AI rack power consumption in kilowatts (kW) (E)	70	kW per rack: The average of power requirement for AI racks dedicated to inference (lower) and training (higher). Compared to our base estimate of 172 TWh, U.S. incremental power demand would be 250 TWh in 2030 at 100/kW, all else equal.

FLOP utilization (F)	50%	The average FLOPs per second performed by the model relative to peak throughput of the underelying training infrastructure. Large models with more complex layers have lower FLOP utilization. Improving FLOP utilization increases energy efficiency.
Days in a year (G)	365	
Hours in a day (H)	24	
Seconds in an hour (I)	3,600	
Tokens generated per second in millions (J)	72.2	(A*B)/(H*I)/10E6
FLOPs per second (K)	9.39E+22	C*J
Number of servers needed (L)	2.93E+06	K*F/D
Number of AI racks needed (M)	5.87E+05	L/5; there are typically five servers per rack (5xH100 configuration per rack).
Power required to operate AI racks in gigawatts (GW) (N)	41.1	M*E/10E6; includes servers, storage, networking switches, central processing unit (CPU) nodes, etc.
Noncomputing power load (O)	8.2	This is a crowdsourced estimate from experts at 20% of AI rack consumption and is for the remaining infrastructure in a data center, including communication equipment, lighting, and security systems.
Power for cooling (P)	32.9	Cooling is 40% of power needs (consenses estimate) of a data center.
Total power in GW (Q)	82.2	N+O+P
GW to TWh conversion (R)	698.1	Q*G*H/10E3; incremental power demand from data center workloads. Assumes 3% downtime for latency agnostic workloads.

Performance to power ratio (S)	1.125	This is representative of the efficiency improvement factor every year; compared to our 172 TWh estimate, demand would be 200 TWh if improvement is 10% per year and 155 TWh if demand is 15% per year, all else equal. The NVIDIA DGX A100 system is listed at 5 petaFLOPS and consumes 6.5 kW max, or 1.30 kW per petaFLOPS.The NVIDIA DGX H100 system does 32 petaFLOPS and consumes 10.2 kW max, or 0.32 kW per petaFLOPS. The new generation NVIDIA DGX B200 system using the new Blackwell chips is listed at 72 petaFLOPS (training) and consumes 14.3 kW max, or 0.20 kW per petaFLOPS.
Global incremental demand of power for AI data centers in 2030 in TWh (T)	344.3	R/(S^6); improvement in six years. Perfomance/power is typically a two-year average. We used annual improvement.
U.S.incremental power demand in TWh (U)	172.2	Estimated at 50% of global incremental demand
U.S. data center power consumption in 2024 in TWh (V)	170	We assumed most of the erstwhile use is classic cloud and not AI.
U.S Power demand growth from AI data center workloads (2024-2030)	12.4%	(((U+V)/V)^(1/6)-1)
Sources: NVIDIA website and presentations, Advanced Micro Devices presentations (AMD), Societe Generale Group, Mckinsey & Co., expert conversations, S&P Market Intelligence, and S&P Global Ratings.		

We believe expectations are healthier than forecasts because they provide a vision of the future stripped of all false precision. Our goal is to wade into the weeds to understand what demand potential exists and the parameters that influence this demand most.

In developing this estimate (see table 1), we did what any reasonable analyst would have done. We ran queries on ChatGPT, which likely sent a server at a remote data center into floating-point burns. As the responses came in, we refined the prompts for increasingly nuanced responses, and finally crosschecked data against reliable sources.

We arrived at the following conclusions:

- S&P Global Ratings assumes a lower growth rate than presented in Chart 1 due to infrastructure constraints. There could be upside to our base case if the purchase and utilization of servers is unconstrained. Conversely, there could be downside to our base case if power efficiency is higher than expected or power/compute speed efficiencies lead to fewer servers purchased than we expect.
- Data center hardware is constantly evolving because general purpose workloads employ CPUs for 75% of their use, whereas AI training workloads employ GPUs for 75% of their use.
- Power density for AI workloads can vary widely from 35 kW-300 kW/rack. Models like ChatGPT can consume 80 kW/rack or more compared with 15 kW for classic cloud.
- Hyperscalers, which are increasingly adopting AI, have a greater need for high-density infrastructure. Currently, their large facilities have an estimated average density of 36 kW/rack, which market consultant International Data Corp. (IDC) estimates will approach 50 kW by 2027. IDC expects many AI cluster requirements to increase to 80 kW-100 kW/rack by 2030.
- New AI innovations have increased computing speed and max power consumption per server. However, as demand for AI training grows over the next several years and for inference longer term, demand growth is exceeding the efficiency improvements, increasing high-power AI server power intensity.
- Continual efficiency improvement cannot be emphasized enough. For example, NVIDIA Corp. has started focusing on energy efficiency. It emphasized how Blackwell's greater processing speed meant that power consumption during training was less than with the Hopper H100 series. Training the latest ultra-large AI models using 2,000 Blackwell GPUs would use 4 MW of power over 90 days of training, compared with 8,000 older GPUs for the same period, which would consume 15 MW of power.

A matter of efficiency but also equity

Some critics claim that taking a generation unit BTM is no different from the "take private" of a vital infrastructure asset, and there has been a lot of debate about FTM and BTM nuclear arrangements (the two announced nuclear transactions are examples of one each). We don't believe this is just a grid connected versus inside-the-fence discussion, but a larger philosophical issue. We can separate this discussion into questions of efficiency (i.e., the lowest cost way of achieving an outcome) and questions of equity (i.e., the fairest way to achieve the outcome).

Efficiency demands that the price of power reflects the lowest cost way of generating, transmitting, and distributing it. There is evidence that new demand, whether grid-connected or inside the fence, pressures existing infrastructure and creates the need for new supply. But the potential trend for nuclear units to seek co-location suggests it is both quicker and easier to build a data center with inside-the-fence interconnection facilities to existing generation than it is to build the new generation and transmission needed to support the data campus if it were to interconnect on the grid.

Under a co-location configuration, the data center pays for all transmission connections (such as high-voltage substations) that are inside the fence and get power without lengthy delays. With the nuclear unit now supplying the data center load, deliverability on the transmission grid is freed up for added resources, typically wind and solar projects, connecting to the grid. So, co-locating data centers at nuclear plants could potentially reduce curtailment of wind and solar output because of inadequate capacity on the transmission system from remote locations to load centers.

In contrast, connecting any large load on the grid would need extra-high-voltage lines, resulting in potentially higher transmission investments that are now allocated to all grid customers as per existing transmission tariffs and cost allocation rules. Similarly, the geographic proximity inherent in co-location also likely reduces line losses due to transmitting longer distances from the power plant to the data center.

Based solely on the issue of efficiency, co-located solutions seem logical. Equity, however, is an altogether separate matter.

A 1 GW data center costs $30 billion-$35 billion, with chips replaced every five years. Even if a PPA contract is about $100/MWh, it is less than 10% of the lifetime costs of a data center. With time to market as a key factor, hyperscalers would be willing to pay an above-market price and grab any generation that provides firm power at scale. The importance of speed to operate the data center makes the economics compelling.

Under such an arrangement, the data center is served by an existing generator, taking its capacity out of the supply stack that was otherwise serving all customers in a regional grid, with BTM being a cost effective way to meet data center demand growth. We expect this to be the rational outcome in competitive markets.

However, nuclear units that are part of a regulatory framework where the generation asset is dedicated to franchised customers who pay cost of service rates are unlikely candidates for such strategies because the regulator would not allow it. The right to serve a territory as the incumbent utility also comes with a regulatory contract to serve customers equally (albeit, all classes of customers do not pay the same rates).

Even in competitive markets, public utilities commissions are now stressing supply concerns in bilateral contracting (i.e., data centers bringing their own generation--sustainable or fossil--to the grid along with their load)

Still, as recently as 2020, prospects for nuclear generation appeared to be on the ropes. On the capacity front, nuclear MW were considered the same as any other MW without any credit for their firm(er) product. On the energy front, prices remained low due to the proliferation of zero-marginal cost resources. As a result, we expected several nuclear plants to shutter. Then came the approval of zero-emission credits/carbon mitigation credits in many states and the federal nuclear production tax credits. Now comes the possibility of long-term PPAs. Not only do PPAs offer the nuclear generator a higher power price than what the wholesale market would deliver, they offer a price that is not subject to the vicissitudes of merchant markets. A nuclear generator--and its shareholders--is certainly entitled to what a competitive market agrees to pay.

Yet, this raises other equity issues. A utility customer pays a retail "wires charge" that includes several non-bypassable charges. These charges often support state programs, which are convenient funding mechanisms for initiatives that have no relationship to distribution and transmission itself. So, while one could argue that co-locating should not justify avoiding these non-bypassable charges, the reality that BTM customers do not receive transmission and distribution (T&D) services would make it harder to charge them for these other resources that are billed through the T&D line.

Finally, through the federal PTC, retail ratepayers have been paying for and receiving the clean attributes of nuclear generation. If nuclear assets contract long term with data centers, seen through the prism of equity, some may argue that the clean attributes are now subsumed in the PPA pricing that the offtaker pays and should no longer be imposed on ratepayers. That said, these future PPA transactions would either begin after the zero-emission credit programs come to an end or at prices likely higher than the PTC floor price, meaning there will be no ongoing support for the plants.

While we remain disinterested observers of public policy, we remain interested to the extent it affects the credit quality of companies we rate.

Related Research:

- **Data Centers: Welcome Electricity Growth Will Fall Short Of U.S. Data Center Demand**, Oct. 22, 2024
- **Data Centers: More Gas Will Be Needed To Feed U.S. Growth**, Oct. 22, 2024
- **Data Centers: Computing Risks And Opportunities**, Oct. 22, 2024
- **Gridlock: Interconnection Queue Backlog Adds Risks For U.S. Not-For-Profit Power Sector**, Oct. 8, 2024
- **Merchant Nuclear Power Update: American Beauty Or American Reality?**, Sept. 13, 2017

This article does not constitute a rating action.

CONFIGURATION. In no event shall S&P Parties be liable to any party for any direct, indirect, incidental, exemplary, compensatory, punitive, special or consequential damages, costs, expenses, legal fees, or losses (including, without limitation, lost income or lost profits and opportunity costs or losses caused by negligence) in connection with any use of the Content even if advised of the possibility of such damages.

Some of the Content may have been created with the assistance of an artificial intelligence (AI) tool. Published Content created or processed using AI is composed, reviewed, edited, and approved by S&P personnel.

Credit-related and other analyses, including ratings, and statements in the Content are statements of opinion as of the date they are expressed and not statements of fact. S&P's opinions, analyses, and rating acknowledgment decisions (described below) are not recommendations to purchase, hold, or sell any securities or to make any investment decisions, and do not address the suitability of any security. S&P assumes no obligation to update the Content following publication in any form or format. The Content should not be relied on and is not a substitute for the skill, judgment, and experience of the user, its management, employees, advisors, and/or clients when making investment and other business decisions. S&P does not act as a fiduciary or an investment advisor except where registered as such. While S&P has obtained information from sources it believes to be reliable, S&P does not perform an audit and undertakes no duty of due diligence or independent verification of any information it receives. Rating-related publications may be published for a variety of reasons that are not necessarily dependent on action by rating committees, including, but not limited to, the publication of a periodic update on a credit rating and related analyses.

To the extent that regulatory authorities allow a rating agency to acknowledge in one jurisdiction a rating issued in another jurisdiction for certain regulatory purposes, S&P reserves the right to assign, withdraw, or suspend such acknowledgement at any time and in its sole discretion.

S&P Parties disclaim any duty whatsoever arising out of the assignment, withdrawal, or suspension of an acknowledgment as well as any liability for any damage alleged to have been suffered on account thereof.

Ratings

S&P keeps certain activities of its business units separate from each other in order to preserve the independence and objectivity of their respective activities. As a result, certain business units of S&P may have information that is not available to other S&P business units. S&P has established policies and procedures to maintain the confidentiality of certain nonpublic information received in connection with each analytical process.

S&P may receive compensation for its ratings and certain analyses, normally from issuers or underwriters of securities or from obligors. S&P reserves the right to disseminate its opinions and analyses. S&P's public ratings and analyses are made available on its Web sites, **www.spglobal.com/ratings** (free of charge), and **www.ratingsdirect.com** (subscription), and may be distributed through other means, including via S&P publications and third-party redistributors. Additional information about our ratings fees is available at **www.spglobal.com/usratingsfees**.

Any Passwords/user IDs issued by S&P to users are single user-dedicated and may ONLY be used by the individual to whom they have been assigned. No sharing of passwords/user IDs and no simultaneous access via the same password/user ID is permitted. To reprint, translate, or use the data or information other than as provided herein, contact S&P Global Ratings, Client Services, 55 Water Street, New York, NY 10041; (1) 212-438-7280 or by e-mail to: **research_request@spglobal.com**.

Contact the analysts:

Aneesh Prabhu
Primary Credit Analysts, CFA, FRM, New York

P. + 1 (212) 438 1285

E. aneesh.prabhu@spglobal.com

Sudeep K Kesh
Primary Credit Analysts, New York

P. + 1 (212) 438 7982

E. sudeep.kesh@spglobal.com

Vishal H Merani
Secondary Contacts, CFA, New York

P. + 1 (212) 438 2679

E. vishal.merani@spglobal.com

Pierre Georges
Secondary Contacts, Paris

P. + 33 14 420 6735

E. pierre.georges@spglobal.com

Cassidy Deaver
Research Assistant, New York

P.

E.