ESG RESEARCH PAPER

"One of the main features of GPUs, along with their fast calculation speed, is their rapid rate of energy consumption."

Above all, this growth is attributable to the intense race of hyper scalers, such as Microsoft and Meta, and their investments toward vital components of AI infrastructure, such as GPUs. Graphic Processing Units (GPUs) are best seen as accelerators for Central Processing Units (CPUs) within AI, due to their ability to simultaneously compute billions of certain operations per second. Accordingly, this has caused the data center GPU market size to grow tremendously, with the market currently being valued at $14.3 billion and estimated to grow at a CAGR of about 35% by 2028.

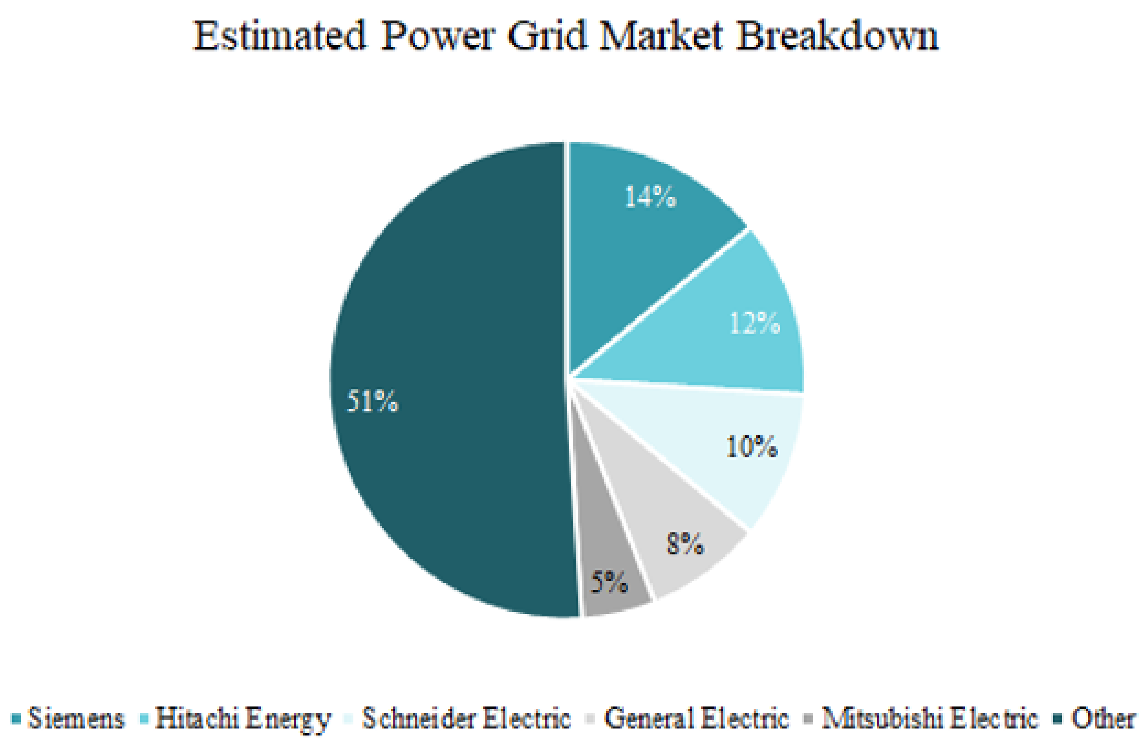
Unfortunately, this high growth comes at a cost. One of the main features of GPUs, along with their fast calculation speed, is their rapid rate of energy consumption. For example, NVIDIA's latest GPU, the H100, consumes up to 700W of power at its peak - more than an average American household. Many energy reports believe that annual data center energy consumption will double by 2026, going from 460 terawatts to over 1000 - about the same as a year's energy consumption in all of Japan. At the moment, the US 2023 forecast of electricity demand is growing at a CAGR of 5% in the next five years - however, this forecast is already double the previous year's forecast (which was 2.6%), and many major utility companies are already expecting this growth rate forecast to double or even triple again in 2024, up to a 10-15% CAGR.

Country	2022 Energy Consumption (in TWh)
China	8,090
United States	4,082
India	1,392
2026 Data Centers (expected)	**1,000+**
Russia	979
Japan	939
Canada	577
2023 Data Centers	**460**

Thus, introducing a barrier to the huge data center ambition of the hyperscalers. Though energy demand is increasing dramatically, the world's energy provisions are not. The global power grid market is set for revenue growth at a meek CAGR of 6% by 2028, compared to the 35% of GPUs and 19% of the overall data center market by 2027.

If electricity demand is to follow the forecasted growth, from 10-15%, there is a strong chance the power grid market will not be able to keep up. Power grids are essential for data center use, amongst other things, as they are the vessels that deliver electricity from power plants to data centers, homes, businesses, and more across the nation. They both generate and distribute electricity, and areas would not be able to get power without the infrastructure. Will the world be able to keep up with the incredibly fast depletion of energy that data centers bring on? And will AI and data center progress face a rude awakening as the reality sets in that faster data centers and GPUs come with a hefty energy price?

"There is a high chance of AI growth being bottlenecked by this lack of energy supply, causing spending and AI growth to begin decreasing again in the next few years."



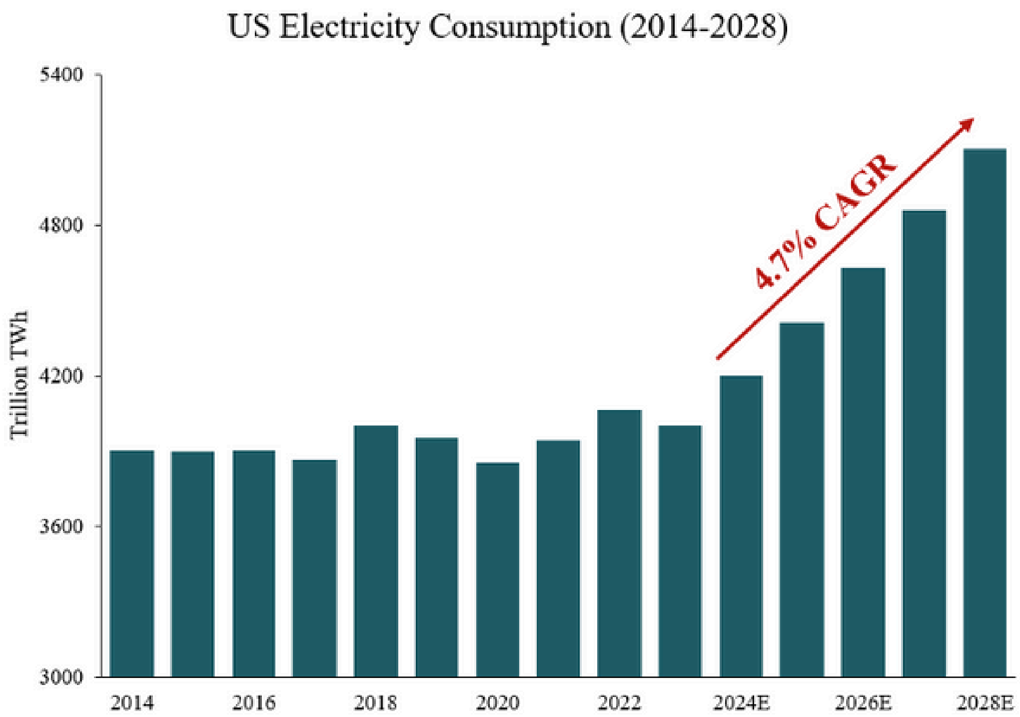
The power grid industry is a complex and overall fragmented market due to ever-changing technology advancements and the energy legislature in different countries, or even different states. The power grid market can be segmented either based on main components (cables, transformers, etc.) or power source (oil, natural gas, coal, hydroelectric, and renewables).

Furthermore, one of the main upcoming subsectors of power grids is the smart grid sector, which is focused on renewable energy power grids. Compared to the overall power grid industry, the smart grid industry is growing at a CAGR of about 10% until 2032, fueled by the government and business desire to transition to more sustainable energy sources. The growth unsurprisingly also aligns with many of the hyperscalers' plans to transition to only using renewable energy to power their data centers in the next few years.

The industry itself has quite a few main players, as certain companies are better able to cater to certain regions and climates so no single manufacturer has been able to monopolize. The main players of the global industry include Hitachi, Siemens, Schneider Electric, and General Electric, however there is still an abundance of smaller players, such as Toshiba and Hyundai Electric & Energy Systems, that focus more on specific regional areas.

The Current State of Power Grids and Data Centers:**

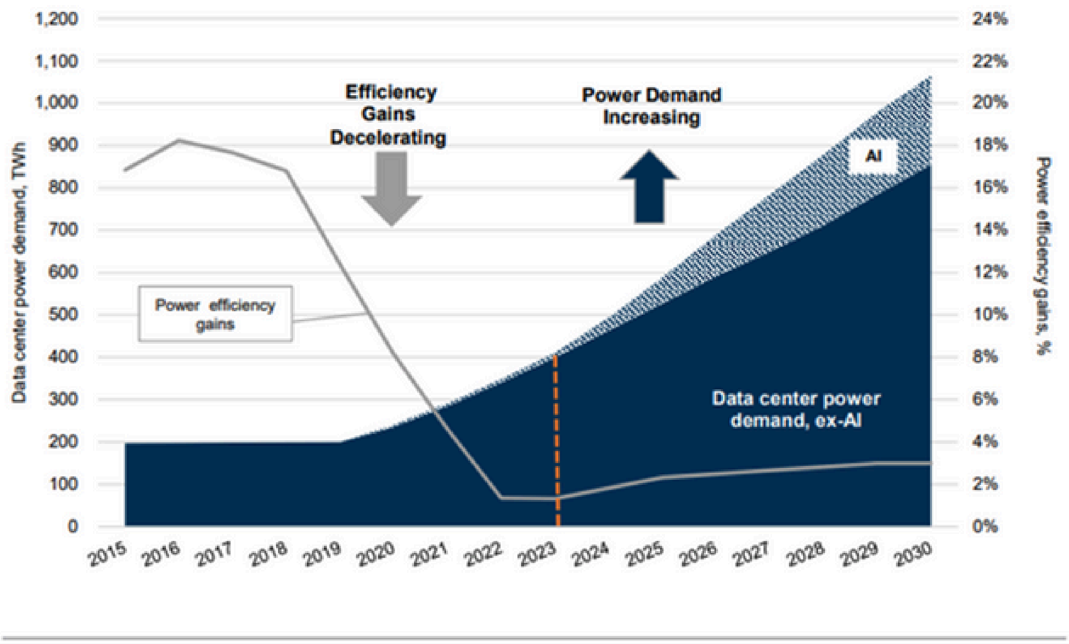
In 2024, across the world, power grid companies have doubled their forecasts for how much power will be needed by the end of the decade, led primarily by data centers. Within the US, for example, energy companies project that data centers will consume about 10% of the entire nations' electricity by 2030. However, it seems this era of severe energy demand is quite an exception to the past few decades. According to many manufacturing plants and power grid companies, growth in electricity demand in the past 20 years had been relatively flat, at around 1% per year, so many companies shrank their departments.



** According to the widely accepted 4.7% CAGR for the next five years (however, this is the most conservative rate, as many utilities companies are expecting even 10-15%)*

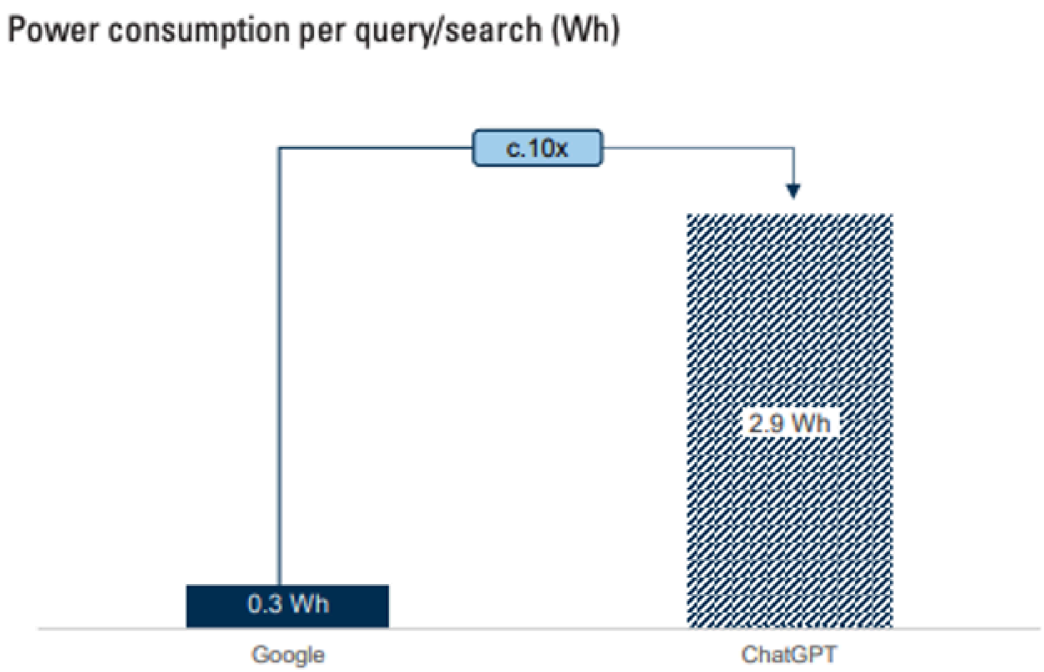
Thus, when the huge influx of demand came in from data centers, most companies were unprepared and slow to expect what was coming. This lack of preparation, by itself, has severely stunted power grid industry growth. Additionally, power grids, along with many other parts of the energy sector, face one large limiting factor - regulatory issues. To build more grids and increase the loads of plants, the industry needs new policies to pass. Unfortunately, the process is extremely time consuming, with companies having to face both state and federal processes to get approved.

Furthermore, AI data centers consume eight times the power of non AI-Powered Data centers and while modern data centers have become more efficient, the rate of power efficiency gain is decelerating and is even projected to backpedal. This sudden increase in consumption will drive costs up unless power generation and power grids can back up this consumption.



Source: Masanet et al. (2020), Cisco, IEA, Goldman Sachs Global Investment Research

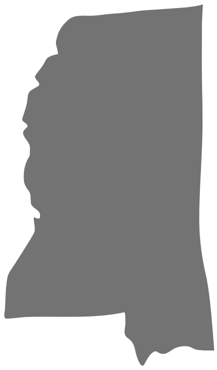
AI had been projected to consume over 20% of the overall data center power demands by 2030. We can already see evidence for this now, as a single ChatGPT search consumes 6 to 10 times the amount of power a traditional google search does.



The US Regulatory Process

Essentially, one of the main issues that delays power grid policy approval is the clashing scopes of the federal and state government. On a federal level, the Federal Energy Regulatory Commission (FERC), is in charge of how much energy is transmitted and generated. However, the actual distribution of this energy, which is what power grids entail, is up to the state government. Often times, the state and the federal government may not be completely aligned which makes the process time-consuming, which means that even if the federal government approves more energy generating processes, the state could stand in the way of the energy being distributed.

For example, in states such as Maryland and Mississippi, they fully support the growth of data centers and have pursued different methods to fuel it. In May 2024, the governor of Maryland passed a law that redefined "generating systems" to also include any backup power sources, such as many power grids. The label of "generating systems" is significant as any system that falls under this term does not require businesses to acquire a Certificate of Public Convenience from state authorities to install more power generation systems. This law should directly benefit hyperscalers looking to build more data centers in Maryland, as it would largely expedite the process for ensuring the center has enough energy. In fact, the passing of this law was a direct response to the Maryland Public Service Commission (PSC) rejecting a data center's request for 160 new power generators.



However, in certain states such as Georgia, the state government recently passed legislation that they were not in favor of more power grids because the return on investment for more data centers within the state was not enough compared to the amount of energy that were consuming. Rather than investing more into the power grid presence in Georgia, the state government decided to halt tax incentives for any new data centers for the next few years, in an attempt to slow energy consumption. These doubts are shared by quite a few other states, such as South Carolina and Virginia.

Since the opinions of state governments vary so widely across the US, the process to get power grids approved can take anywhere from months to years. Unless every state miraculously gains the resources to support the influx of data centers, there is a strong chance that growth in the power grid sector will not be able to accelerate, and therefore not be able to meet the demand of data centers.

Challenges for Power Grid Companies from Hyperscalers' Sustainability Goals

Hyperscalers are making life harder for power grid companies by accelerating the transition toward renewable energy, with most aiming to rely exclusively on zero-emissions power within the next decade. However, their shift to sustainability isn't purely altruistic—it is primarily driven by two factors:

Reputation Management | Cost Savings

Firstly, hyperscalers have faced backlash over their environmental impact, especially on local communities, due to excessive energy and water usage. By committing to greener practices, they can partially appease environmental activists and mitigate resistance from the public. Secondly, while the upfront investment in renewables can be high, research shows that renewable energy often proves more cost-effective in the long run. Tax incentives, subsidies, and declining operational costs provide an additional financial incentive.

Many industry giants are already leading the charge. Meta, for example, runs on 100% renewable energy, setting a benchmark for others. Amazon's data center operations currently use 90% renewable energy, with plans to hit 100% by the end of 2025. While these efforts are beneficial for the environment, they introduce new challenges for traditional power grids.

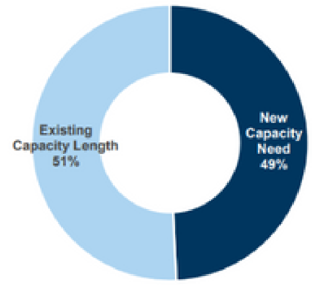
The inconsistent nature of renewable energy—such as limited solar output on cloudy days or insufficient wind for turbines—poses a problem. Legacy grids are designed for steady, reliable energy flows, making them ill-suited to manage fluctuating renewable sources effectively. This mismatch underscores the need for smart grids, which can adapt to these variations.

Smart grids utilize advanced technologies to enable a two-way flow of electricity and data, facilitating real-time communication between energy suppliers and users. Unlike conventional grids, they can dynamically respond to supply and demand shifts, ensuring efficient energy management. By leveraging these innovations, power grids can integrate renewable sources more effectively, paving the way for a more sustainable energy future.

Power Generation Sources and Capital Expenditure

The North American Electric Reliability Corporation (NERC), a nonprofit organization focused on ensuring the reliability of bulk power systems through compliance and enforcement programs, has reported that nearly half of the growing energy demand from data centers will require new generation capacity.

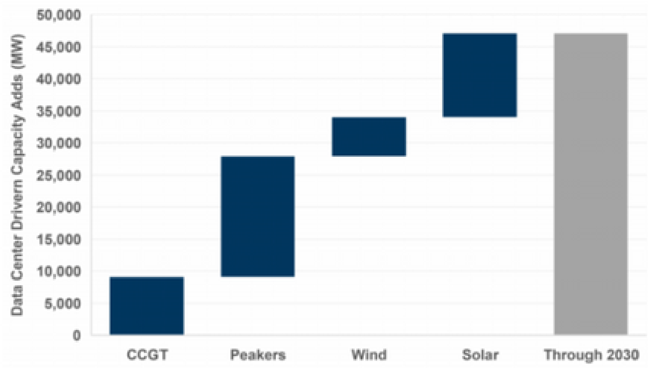
How to meet data center driven power demand, %

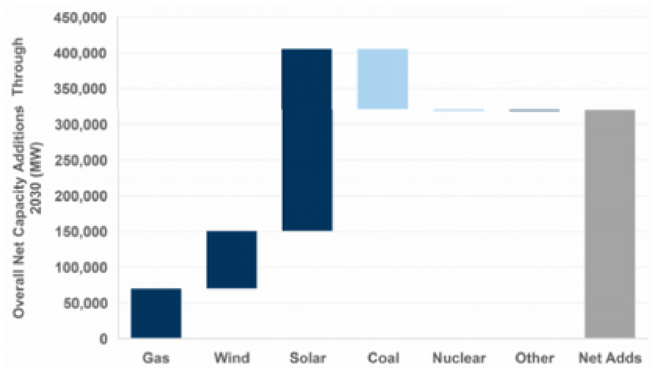


According to Goldman Sachs, 60% of this new demand will be met through natural gas, divided in a 2:1 ratio between Combined Cycle Gas Turbines (CCGT) and Peaking Units. Specifically, 9 GW will come from CCGT, while 19 GW will be supplied by Peaking Units.

The remaining 40% of demand will come from renewable energy sources, split in a 5:3 ratio—with solar energy contributing 13 GW and wind energy providing 6 GW. This expansion will result in an additional 48 GW of capacity between 2024 and 2030.

Notably, the projections do not anticipate growth in America's nuclear power capacity. However, the report indicates a potential slowdown in nuclear plant closures and predicts an increase in the adoption of small modular reactors (SMRs), which could play a more significant role in the future energy mix.





If these projections hold, the U.S. could see a 15% increase in overall net capacity by 2030. However, this growth comes with a significant price tag.

Power Generation Method	Cost per kW
CCGT	$1000/kW
Gas Peaker	$750/kW
Onshore Wind	$1600/kW
Utility Scale Solar	$1300/kW

To support this growth, approximately $50 billion in cumulative investment will be needed—roughly $7 billion annually for new power generation alone.

In addition to generation costs, the distribution infrastructure will require massive upgrades. Estimates suggest that the industry could spend around $720 billion cumulatively on power grid improvements and expansion to support the evolving energy landscape.

Impact on Chip Manufacturing

One unexpected consequence of energy imbalances is the rising energy demand from chip and battery factories, which are critical for producing AI hardware such as GPUs. These factories are highly energy-intensive. For instance, new semiconductor plants in the U.S., built by major players like Intel, TSMC, and Micron, are projected to consume twice as much energy annually as large cities like Seattle—roughly 2.1 GW per facility.

Despite concerns about energy consumption, chip manufacturers show no signs of slowing down. Intel and TSMC are already building four new factories in Arizona, adding to the growing energy burden. This poses a significant risk to the energy supply-demand balance, as these factories are likely to exacerbate the energy shortage.

If energy supply becomes constrained, chip manufacturers may struggle to meet hyperscalers' growing demand for GPUs and other critical components. This bottleneck in chip production could stall progress in AI development and limit future investments.

Impact on Hyperscaler Capex & Operations

Hyperscalers are already facing rising energy costs to power their massive data centers. For example:

- Microsoft reported an additional $800 million in energy expenses for the year.
- Amazon cited a 200 basis point margin drop due to higher electricity and natural gas prices.

These rising operational costs suggest that hyperscalers may need to increase capital expenditures (capex) in the short term to adopt energy-efficient infrastructure and offset long-term energy expenses. One potential shift could involve accelerating the adoption of energy-efficient hardware, such as ARM-based server chips, which consume less power than traditional x86 chips.

However, smaller data centers will likely struggle to keep pace with rising energy costs and the need for continuous upgrades to stay efficient. This could drive further consolidation of the data center market, with hyperscalers increasingly monopolizing the space as smaller operators find it difficult to remain competitive.

Over time, hyperscalers may need to reduce their capex to balance energy costs, potentially slowing down new data center construction and equipment purchases. Combined with the chip supply bottlenecks mentioned earlier, hyperscalers could be forced to scale back their growth until the power grid and energy infrastructure can catch up.

In the next five years, operational and R&D costs are likely to rise further as hyperscalers search for solutions to manage energy consumption and sustain growth. The energy crisis could also push them to rethink their strategy, shifting focus from expansion to optimization and sustainability.

What are Hyperscalers Doing to Handle the Energy Increase?

Though the sheer increase in data centers usually more than offsets any energy efficiency progress that a hyperscaler makes, hyperscalers have been vying for ways to slow the energy drain through certain improvements. Data center cooling systems usually make up about 40% of data center energy use, meaning that changes within these systems would make a substantial impact on energy consumption. Unsurprisingly, many hyperscalers have taken advantage of this and begun adopting much more energy efficient cooling systems than just traditional air-conditioning systems, like Microsoft's immersion cooling method, which can dramatically decrease energy consumption.

For example, Microsoft's new immersion cooling method uses a much simpler hot liquid process rather than fans to cool systems. This therefore takes away lots of power-consuming equipment, such as server fans and cold water pumps. In fact, they estimated that each process would take only 1.014 MW, rather than 1.55 MW, reflecting a 35% decrease in energy consumption. Other hyperscalers have tried other methods to reduce energy consumption, with Equinix implementing weather-adapting technology in their cooling systems which increase their energy efficiency by about 9%.

While these new systems have already made significant improvements, it is likely that hyperscalers will continue to look for new cooling systems to keep their energy efficiency up, such as Microsoft. Microsoft has recently been exploring the usage of adiabatic cooling - using outside air and humidity instead of water for cooling. This technique would use little to no water, also helping Microsoft's efforts to become water positive, and be significantly more efficient because it would use the outside climate rather than pure energy, and would be able to automatically adapt to any environment. As of right now, these systems are being tested in Arizona, with exact results yet to come out, however experiments like this are a clear indication that we can expect research expenses to continue to increase, and likely cooling system infrastructure to eventually boost capex spending.

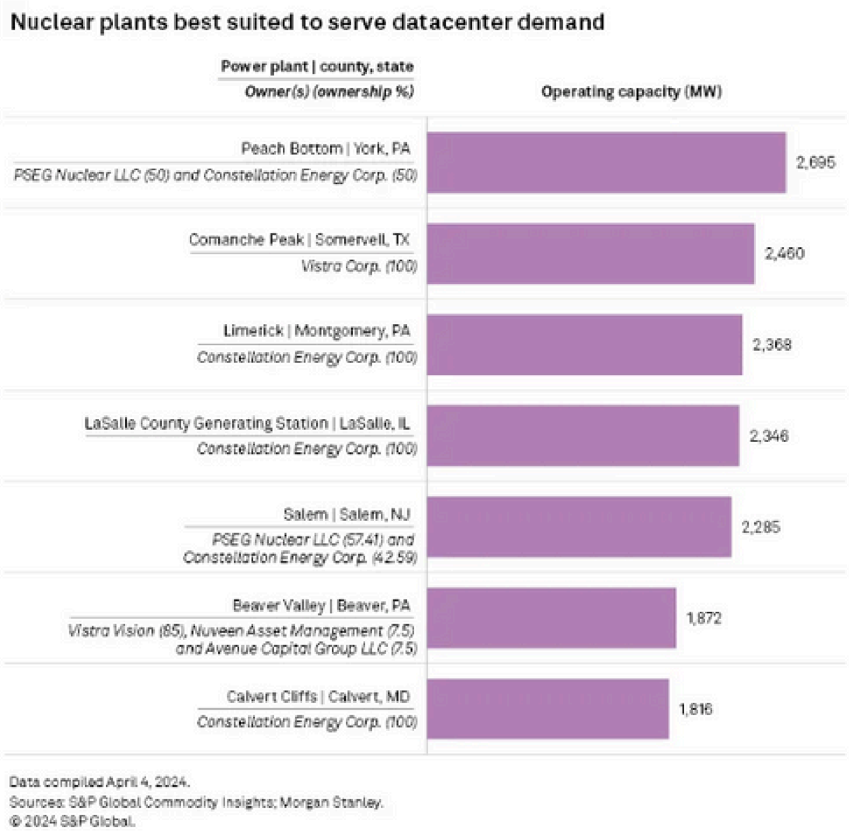
Long-Term Future Outlook: Nuclear Energy

As hyperscalers like Google, Microsoft, Meta, and Amazon commit to powering their data centers with 100% carbon-free energy by 2030, many are turning to nuclear energy as a viable long-term solution.

Microsoft has recently signed an agreement to partially power one of its data centers using nuclear energy, supplemented by nearly 100% carbon-free electricity at other times. Similarly, AWS has made significant investments in nuclear energy, purchasing a data center campus adjacent to a nuclear power plant in Pennsylvania for $650 million. This strategic move allows AWS direct access to carbon-free energy, enhancing its existing wind and solar projects.

Green Energy Partners is also capitalizing on this trend by planning a combined data center and energy campus near the Surry Nuclear Power Plant in Virginia. This innovative project will feature on-site power generated from small modular reactors (SMRs) and hydrogen production, creating a hub that leverages the existing nuclear infrastructure.

Additionally, Constellation Energy, the largest nuclear power operator in the U.S., is exploring the construction of new nuclear capacity at its current reactor sites to meet the increasing demand from data center customers. As the energy landscape evolves, nuclear energy is poised to play a critical role in helping hyperscalers achieve their sustainability goals while ensuring a reliable and carbon-free power supply.



The Role of Public Spending

As anticipated, the government is working rapidly to meet the growing demand for energy generation and major power grid infrastructure. According to Bloomberg, the U.S. is expected to increase its total generating capacity by 80% by 2035, with a remarkable 96% of this growth anticipated to come from renewable sources, particularly solar energy. This dramatic expansion will inevitably drive progress in other industries, especially in storage solutions to accommodate the influx of power into the grid.

The Biden-Harris administration has already initiated construction on ten major transmission projects aimed at enhancing the storage capacity, resilience, and speed of the U.S. power grid. Unlike previous upgrades that primarily involved installing new transmission wires, most of these initiatives focus on integrating advanced technologies, such as high-performance conductors capable of carrying over twice the power of traditional wires. This modernization will also better equip the grid to handle renewable energy, facilitating a smoother transition as generation from these sources increases.

However, these improvements come with a substantial price tag. The Grid Deployment Office of the Department of Energy (DOE) is administering $10.5 billion in funding for these projects, marking the largest investment in U.S. power grids to date. This announcement, made in late May 2024, followed the Biden-Harris Administration’s commitment of $3.5 billion for power grid enhancement in October 2023. Clearly, public investment in power grids is ramping up quickly. Given the U.S.'s previous unpreparedness for this energy surge, the fact that the government tripled its investments in just eight months indicates a commitment to further investments in the power grid throughout 2025 and 2026, especially as data centers evolve and energy consumption predictions continue to rise.

Overlap with Private Spending

This surge in public spending, particularly focused on renewable energy, aligns closely with hyperscalers' objectives to transition fully to clean energy. The top five hyperscalers—Amazon, Google, Meta, Microsoft, and Apple—collectively account for over 57% of global corporate wind and solar capacity, and their push for renewable energy has significantly influenced the sector. Given their substantial share of energy consumption in the U.S., their increasing demand for renewables will likely compel the DOE to allocate additional funds specifically for renewable energy initiatives.

With aggressive investments in renewable energy and energy efficiency expected to continue accelerating through 2026 and potentially until 2030, analysts anticipate that the U.S. will need to invest at least $50 billion in new generation capacity by 2030 to adequately support data centers. This funding is currently being facilitated by the Bipartisan Infrastructure Deal, approved in November 2021, which allocated over $65 billion for clean energy transmission and power grid improvements over the next decade. However, this investment aims to enhance energy capacity not just for data centers but also for general public use and other energy consumers. If data centers alone require at least $50 billion in investments within the next five years, there is a strong possibility that the original $65 billion investment will fall short, necessitating further public spending on power grid upgrades.

This new generation capacity will also demand enhanced storage solutions, requiring an acceleration in public spending within the power grid sector. Furthermore, this will likely increase both the volume and frequency of investments. To meet the needs of hyperscalers, the U.S. government has expedited the approval processes for transmission line and power grid permits, potentially reducing project development timelines by years. With this streamlined approach, hyperscalers may seize the opportunity to ramp up their energy consumption, assuming that the government will continue to prioritize capacity expansions. Consequently, we could see a cyclical relationship emerge, where hyperscalers push for increased capacity, prompting faster government action to meet demand until a sustainable equilibrium is achieved.

IMPACT ON ITAD

Transition to Smaller Data Centers (Edge Computing)

In response to the surging demand for data and the escalating energy consumption associated with large hyperscaler data centers, many analysts predict that hyperscalers will increasingly shift towards smaller, more distributed data centers known as edge data centers. The fundamental principle of edge computing is to process data closer to its source, thereby improving both speed and energy efficiency.

Smaller data centers inherently manage less data, which helps free up bandwidth and enables faster, more efficient operations. This proximity to data generation significantly reduces energy consumption, particularly in terms of data transport. Additionally, edge data centers can lower their cooling requirements due to their smaller size and reduced infrastructure. Some innovative edge facilities are able to eliminate high-energy cooling systems entirely, utilizing local water supplies or the natural climate for temperature regulation. This is especially impactful given that cooling systems typically account for about 40% of energy consumption in traditional data centers.

Moreover, many edge data centers can adapt to fluctuations in resource utilization. When demand decreases, they can enter a dormant state, conserving energy rather than operating continuously at full capacity, which is common in traditional data centers. The absence of on-site personnel further minimizes the risk of human error, which is responsible for approximately 60% of issues in conventional facilities. With traditional data centers often running 24/7, they tend to waste resources during periods of low demand. Given these advantages—both in terms of operational efficiency and enhanced reputation—it's unsurprising that the edge data center sector is projected to experience substantial growth, with a CAGR of around 23% over the next five years.

How Edge Data Centers Affect IT Asset Disposition (ITAD)

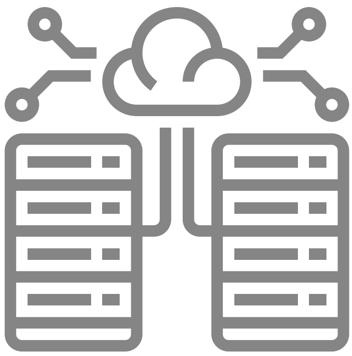
The more efficient nature of edge data centers leads to longer lifecycles for IT assets, resulting in extended refresh cycles compared to traditional hyperscaler data centers. However, this shift also presents new opportunities for local, on-site IT asset destruction, as the costs associated with transporting IT assets from multiple locations are likely to exceed those of handling assets at a single, centralized site. Analysts expect companies to increasingly seek local solutions for IT asset disposition.

Additionally, the proliferation of edge data centers creates significant opportunities for Third-Party Maintenance (TPM) services. As the number of edge data centers rises, managing maintenance across diverse locations will become a considerable challenge for larger organizations. Recently, there has been a trend toward vendor consolidation, with companies operating edge data centers increasingly searching for TPM providers capable of delivering comprehensive support across multiple sites.

Conclusion

As data center energy consumption is projected to more than double by 2026, the power grid industry is expected to grow at a much slower rate. This discrepancy suggests that 2026 could serve as a pivotal reality check for many hyperscalers, revealing that their energy efficiency improvements may not be sufficient to keep pace with the rapid expansion of data centers.

Despite significant investments made in 2024, hyperscalers are likely to experience a slowdown in progress by 2025. The increasing energy demands of their data centers, coupled with constraints faced by chip factories, will hinder their ability to secure the necessary power. Consequently, by 2026, capital expenditure (capex) spending may begin to decline, shifting the focus toward research and development rather than new infrastructure investments.



The complexity of scaling the power grid infrastructure means that we do not anticipate the system will be able to support robust growth until approximately 2030. At that point, as the power industry stabilizes, we can expect to see a renewed acceleration in capex spending, allowing hyperscalers to effectively align their energy needs with their operational growth.