WHITE PAPER

THE IMPACT OF AI ON DATA CENTER INFRASTRUCTURE

How AI-driven data centers are transforming the U.S. energy landscape.

Rajan Battish PE, ATD, LEED AP

ABOUT RSP

Founded in Minneapolis in 1978, RSP Architects has emerged as one of the country's most trusted, diverse and agile architecture practices.

The firm's clients are a dynamic cross-section of Fortune 100 global brands and retailers, innovative start-ups, thoughtful non-profits, government agencies, and more.

PREVIOUSLY PUBLISHED

This paper was first published by Johns Hopkins School of Advanced Studies as a capstone to Mr. Battish's Master of Arts in Sustainable Energy.

In this white paper, RSP Principal and Mission Critical Lead Rajan Battish details how AI-driven data centers are transforming the U.S. energy landscape. Once focused on research and enterprise, data centers now represent one of the fastest-growing demands on the electric grid. Goldman Sachs projects AI consumption could reach 500 TWh, with PJM forecasting a 16% increase in demand over the next 15 years. Hyperscale facilities, fueled by investments from major tech firms, refresh hardware every two years and require more power-dense cooling, pushing efficiency metrics higher than traditional centers. While past innovations like virtualization and cloud consolidation kept energy use flat, the scale and speed of AI adoption may overwhelm supply, raising the risk of an "AI bubble" similar to the overbuilding of the Dot Com era.

Adoption, governance, and infrastructure policy will shape the trajectory of AI data centers. With load forecasts difficult to extend beyond five years, utilities face mounting pressure to deliver flexible, sustainable solutions. Emerging technologies such as reversible and quantum computing may soften demand, but bridging solutions like Small Modular Reactors, distributed generation, and smart grid systems will be critical. Collaboration between utilities, regulators, and developers will determine whether AI's energy appetite destabilizes the grid or drives innovation in sustainable power delivery. Mission Critical experts are continuing to drive change and provide solutions to developers, tech companies and communities in this rapidly changing industry that has the power to disrupt nearly every sector.

BACKGROUND AND CONTEXT

This white paper explores how data centers have evolved from being predominantly focused on research and government to incorporating Artificial Intelligence (AI) into everyday life. It provides insight into the impact of data centers on the U.S. electric grid and discusses how sustainable energy can help address future challenges faced by the data center market. With the rise of hyperscale AI data centers, U.S. electricity demand is projected to increase significantly, with Goldman Sachs estimating that AI-driven consumption could reach 500 TWh.

The Pennsylvania-New Jersey-Maryland Interconnect (PJM) anticipates a 16% increase in grid demand over the next 15 years due to data centers. The average across the ten planning utility regions indicates a total net capacity of 128 GW for peak summer demand. This anticipated growth in energy is based on capacity requests from the Federal Energy Regulatory Commission (FERC). The expected grid demand across the ten utility planning regions in 2024 was driven primarily by AI data centers, chip manufacturing, and the battery manufacturing industries.

Global data center demand was expected to remain plateaued; however, McKinsey forecasted that AI data centers are the primary driver of increased electrical demand. According to Goldman Sachs, most of the data centers are byproducts of the Magnificent Seven, which have added about $11 trillion in market cap since 2023. The rapid investment in chips and data centers means that power demand will increase more than the supply.

There are concerns that the speed of AI deployment may create another Dot Com bubble, similar to how the rapid growth of data centers in the late 1990s resulted in an overcapacity. An article in the Wall Street Journal titled “Internet Hype in the 90s Stoked a Power Generation Bubble” noted that the overbuilding led to the bankruptcy of independent power generators and developers of data centers during the Dot Com bubble.

The potential AI bubble will depend on the acceptance and use of AI by businesses and consumers. The applications of AI can range from data-driven marketing and research to development and financial forecasting. Small businesses anticipate AI usage for cost savings, time efficiency, and increased productivity for their staff.

Governance due to the risk of AI abuse may impact the adoption of AI applications, and thus, the growth trajectory may need to be adjusted. In addition to policies for a safe environment regarding AI use, infrastructure risks for utilities can also influence AI growth trajectories. Increased costs for electrical infrastructure will strain the relationship between AI data center developers and utilities that are required to support them. Tariff rate structures and policies for grid stability will need to be developed and implemented. Discussions regarding this matter have already begun with the North American Electric Reliability Corporation (NERC), where recommendation white papers for large loads are underway.

> Traditional data center energy consumption has remained flat since 2010 due to improvements in IT hardware and cooling, as well as a shift from enterprise to hyperscale data centers.

Data centers are buildings designed to host computers, which are frequently updated with newer technologies, impacting the overall electrical load. These hyperscale data centers have an approximate refresh rate of 18 to 24 months, in alignment with Moore's law. Energy consumption will change to align with the refresh cycle; therefore, accurately predicting future load profiles beyond 3 to 5 years is difficult, much less grid planning that is financed for over 10 years out.

An example of this uncertainty can be illustrated by data center growth anticipated in the mid-2010s, when increased technology use prompted corporations to build data centers at their facilities. However, data center growth was later reduced due to virtualization, where multiple server functions were consolidated into a single, higher-compute server. While the total IT compute load increased, it did not grow at the rate previously assumed. The impact of virtualization is evident in the case study outlined in this white paper, where pre-virtualization, the growth rate was approximately 34.3%, reduced to 8.7% after virtualization.

The cloud also disrupted the data center industry by consolidating distributed data centers and increasing compute efficiency. The case shows a negative growth rate of 3.6% as the data center computer became centralized and consolidated. After the consolidation was complete, data center growth stabilized at a nominal rate of 2.3%.

The Power Usage Efficiency (PUE) of data centers has improved from a historical level of 3.0 to the current hyperscale data center level of approximately 1.1. AI data centers exhibit higher power density per cabinet and require mechanical cooling, leading to anticipated higher PUE. Our analysis employs a blended PUE of 1.35 for AI data centers, assuming some loads are cooled by air at 1.1 PUE while others use mechanical systems like chillers, which will push the PUE to 1.5.

Thanks to technological advancements, traditional data center energy consumption has remained flat since 2010 due to improvements in IT hardware and cooling, as well as a shift from enterprise data centers to hyperscale data centers. The report anticipates disruptive impacts on AI data center growth, such as reversible computing, quantum computing, and network speeds. Based on extrapolated trends, the projected 128 GW peak load across ten planning regions could be reduced to 92 GW (28% savings). This load may decrease to 80 GW if efficiency measures are widely adopted.

Bridging power solutions, such as Small Modular Reactors (SMR) and natural gas generation, will be necessary to meet the demands of data centers. Utilities must transition from conventional energy delivery models to dynamic energy management, integrating Distributed Energy Resources (DER), on-site generation (e.g., SMRs), and smart grid technologies. A partnership between data center developers, regulators, and utilities will be essential to meet the infrastructure demands of AI while maintaining grid stability and sustainability.

INTRODUCTION

Our dependence on technology has increased since Moore's law was established in 1965 by Gordon Moore. Moore's Law predicts that the number of transistors in integrated circuits will double every two years. The advancement in compute technology has influenced the development of personal computers, smartphones, and the internet. The technology we depend upon uses a web of infrastructures to connect our devices, communications, and data. The hubs of these interconnects are the data centers.

The Magnificent Seven Companies (Apple, Microsoft, Amazon, Alphabet, Meta, Nvidia, and Tesla) have driven the push for IT compute consolidation using the cloud and now Artificial Intelligence (AI). Prior to 2020, the cloud was the driving force of growth in data centers, where mega campuses of IT compute facilities were created to be the primary hub for information compute and exchange. More recently, AI has become the main driver for growth in the data center market.

Top IT companies are looking to invest billions of dollars in the AI data center market, as seen in EIA Graph 12 in Appendix B. The capital spending as a percentage of annual revenue ranges from 10% for Amazon Web Services (AWS) to 28% for Microsoft(MSFT). The AI investments is equivalent if not higher than the average cloud investment by these companies.[6]

The investment in AI will cause tremendous pressure on the US energy industry and specifically the electric grid generation and distribution infrastructure. The AI data center has much higher energy consumption compared to conventional data centers of the past. The current AI computer cabinet can consume 140 kW to 200 KW, which is up to one hundred times more power consumption per cabinet compared to traditional data centers. Future AI server cabinets are anticipated to be two hundred times the power consumption of conventional data centers.

Grid instability is another major concern of AI due to the dynamic nature of AI hardware. An AI data center load profile is anticipated to be similar to a crypto mining facility. The load profile of a crypto mining data center is transitory compared to the fixed demand load of conventional data centers. The comparative step load profiles of crypto and traditional data centers can be seen in Appendix B, Graph 1. The load profiles for smaller AI facilities may not affect the grid, but the hyperscale data centers can impact the grid voltage and frequency due to large step loads and transitory load profiles.[7] This creates an outage risk for the transmission grid that connects vast areas of the country.

The nature of AI compute, with its rapid transitory voltage load signature, poses risks of grid instability to the electrical infrastructure that need to be addressed. Solutions could include using technologies such as supercapacitors and Battery Energy Storage Systems (BESS). To address the large step loads of hyperscale data centers, a BESS integrated with the grid helps minimize voltage deviations, as illustrated in Appendix B, Graph 2.

In addition to voltage and frequency stability concerns with AI data centers, rapid demand load requirements will strain the utility generation capacities. Hyperscale data center project sites can require an initial power of 100MW to a final power of 1GW. These capacities are typical of coal power plants and large nuclear power generation facilities that can take years to construct. The high-power

The need for computing and storage near high-density areas, such as cities or regions with high computing requirements, has created an emerging demand for edge compute data centers.

requirements in short duration will require onsite generation. These bridging power plants are typically natural gas, with a capacity of 50MW to 100MW.

BACKGROUND

This white paper focuses on data center loads that are classified into Traditional, Cloud, Edge, and AI data centers. We will not analyze cryptocurrency mining data centers, though they have similar components, but different uses and functions.

TRADITIONAL DATA CENTER

Data centers were purposely built and designed for specific purposes such as research, financial, email, and application-specific requirements. Traditional data centers typically provide transactional compute and have lower power density compared to modern data centers. The power density is typically less than 10 kW per cabinet, and the total grid power to the facility is less than 10 MW. This is typically the limitation of a single 15kV feeder from the utility and common medium voltage distribution for corporate campuses.

The data centers were usually part of the corporate portfolio of a company and cohabited on corporate campuses. These data centers were purpose-built, and due to high availability requirements, they were not as efficient. The data centers typically operated at Power Usage Efficiency (PUE) from 1.7 PUE to 3.0 PUE.

CLOUD DATA CENTER

Cloud-based data centers have a higher power density per cabinet than traditional ones due to the more powerful compute machines utilized. The typical power density can range from 10 kW per cabinet to 20 kW per cabinet. These cloud-based data centers are purpose-built by hosting companies such as AWS, Microsoft, and Alphabet. These facilities consist of larger buildings with power consumption ranging from 20 MW to 100 MW per site. In areas where data center parks are developed, power consumption can reach up to 500 MW. Due to economies of scale in IT compute and infrastructure efficiencies, these facilities typically operate at a PUE of 1.1 to 1.4.

EDGE DATA CENTER

The need for computing and storage near high-density areas, such as cities or regions with high computing requirements, has created an emerging demand for edge compute data centers. These facilities are smaller than hyperscale data centers and support processing loads closer to the user, minimizing latency and costs associated with transferring large amounts of data from the user to the cloud. Edge provides an interface between regional loads and larger cloud data centers. They are typically purpose-built for up to 60 MW. The demand for these facilities will grow due to smart cities and EV systems.

AI DATA CENTER

The AI compute data centers typically range from 100 kW to 400 kW per cabinet. Due to the high-density compute cabinets, the site power for the AI data center can vary from 500 MW to 2 GW per site. These facilities are purpose-built and usually have a higher PUE (up to 1.35) because of the liquid cooling needs that are assisted by air cooling. The market drivers for the AI data centers include corporate competitive advantages, consumer demand, and investment from governmental agencies such as DeepSeek and Stargate, a $500 billion investment.

Commercial applications of AI support data-driven social media, advertising, and marketing services. Research and development applications run complex calculations and models, such as genome mapping and space research. AI is used for financial applications in forecasting, analysis, and increased productivity.

The increase in investment for data centers is evident as data centers supported the cloud in the mid-2010s to the 2020s, where AI is driving exponential growth. The global AI data center market size is projected to reach $827 billion, compared to $690 billion for the cloud. The investment potential of AI data centers suggests expectations of AI becoming mainstream, similar to the Internet of Things in the 1990s. Statista Market Insights regarding investments in cloud and AI data centers are represented in Appendix B, Graph 3.

HISTORICAL PERSPECTIVE

Data centers were initially designed for government and research universities that required buildings for large mainframes, magnetic disk storage, and network connectivity over copper lines. As compute increased and costs decreased, technology was adopted by commercial users, and data centers became part of the private sector real estate portfolio. Growth began in the 1990s with the Internet of Things and evolved into managing critical business functions such as financial banking and communications. Companies like UPS, which established the Uptime Tier IV availability standard, utilized this technology to help track packages and processing.

The commercial data center market was purpose-built and customized to meet varied resiliency requirements based on business needs. Uptime Institute had developed availability requirements to help industries justify the infrastructure costs associated with data center development. Four-tier ratings (Tier I to IV) were established, with each higher rating offering increased availability at a greater infrastructure cost. These data centers were an asset of a corporate real estate portfolio designed to last over twenty years. The cost of technology was high, and most corporate clients' refresh cycle was four to five years, compared to a two-year refresh for hyperscale data center cloud providers.

As technology costs decreased and compute dependence increased, the expenses associated with data center capital (CapEx) and operational expenditures (OpEx) grew for functions primarily located in the back office of corporations. In the 2000s, data centers imposed a greater economic burden on corporate portfolios, leading to new opportunities for data center developers such as Equinix and Digital Realty, which facilitated the outsourcing of data center services. These facilities provided expansive warehousing for co-location services, where spaces were designated for hosting the owner's IT equipment within dedicated caged

areas. The spaces were leased with Service Level Agreements (SLAs), allowing data centers to be categorized as expenses rather than debts.

Data centers experienced a resurgence in the mid-2000s due to e-commerce, search engines, and the commercial use of IT services. Initially, new technology driven clients from Silicon Valley utilized leased facilities from data center developers. Advances in technology and market expansion increased the demand for larger data centers. These larger facilities provided the necessary computing power and connectivity to meet business needs.

Since electric power consumption represents a significant cost for data centers, the necessity to reduce both CapEx and OpEx has spurred innovations in IT, mechanical, and electrical infrastructure. The industry had begun measuring and improving energy consumption. Solutions have included transformer-less UPS systems, airside economizers, and Data Center Information Management (DCiM) software.

The term Power Usage Effectiveness (PUE) was established in the mid-2000s to help identify and measure energy efficiency. PUE, in its simplest form, is the ratio of the total electricity metered at the grid for a data center facility divided by the electricity consumed by IT equipment:

$$PUE = \frac{\text{Total Facility Power}}{\text{IT Equipment Power}}$$

Data centers are steady-state electric energy consumers, as coined by the term 7x24. They require constant power and cooling for 7 days a week, 24 hours a day. This continuous power consumption is unique to the data centers industry, creating a base load demand for grid operators.

Initially, most data centers utilized dedicated servers for their respective applications. This required a large quantity of IT equipment and operated at low efficiency in terms of compute per watt. When the applications were not in use, the servers remained idle and were not operating at optimal efficiency. Advances in virtualization, which allowed one server to run multiple applications, compared to dedicated servers that ran dedicated applications, helped consolidate IT hardware and optimize server watts per compute. As efficiency gains were achieved at both the chip level and building infrastructure, cloud computing further optimized and expanded the data center market.

Cloud computing in the 2010s spurred the consolidation of smaller distributed data centers into larger, centralized dedicated data center facilities. The data center market grew alongside increased streaming services and software applications, which required data center consolidations to leverage economies of scale. The corporate data center footprint shrank and consolidated to the data center service providers. With lower PUE, IT efficiencies, and economies of scale, data centers maintained a flat electrical demand on the grid. The flat load profiles of the data center are represented in Graphs 4 and 5 in Appendix B.

With optimized power and cooling solutions, the PUE remained below 1.15 for cloud data centers. However, data centers grew larger, occupying mega campuses with six to twelve buildings on over 100-acre sites. The power requirements for larger hyperscale data center sites increased to over 300 MW, with each data center building consuming up to 25 to 30 MW of power. Each site is equivalent to

> The demand for AI compute has put tremendous strain on utilities to provide the necessary generation and transmission.

a couple of standard coal power plants and larger than many Independent Power Producers (IPP) can support. The fragmented national grid struggles to support the generation of electric energy needed for these hyperscale data center campuses.

In the 2020s, AI and High-Performance Compute (HPC) changed the dynamics of data centers once again. AI data centers are not a direct replacement for conventional data centers currently, but that may become possible in the future as new software and hardware develop. Typically, AI data centers require lower resiliency since the applications are compute-intensive rather than transactional, as with traditional data centers. HPC is necessary for AI and utilizes different computing components and performance requirements. The growth in AI applications supports emerging markets, machine learning, modernizing processes, learning, and data-driven decisions.

Policies for accountability and ethical issues arising from the manipulation of resources and people must be developed. The advancement of AI might be constrained by protectionist regulations. According to Matt Garman, CEO of Amazon, "What we advocate for is really thinking through not setting regulations or policies that are inadvertently going to lead us to place that you are trying to avoid. It's very easy to construct a scenario where you actually give China the leg up that you are trying to prevent by accidentally holding back the companies there doing all this incredible innovation."[8]

To improve compute efficiency, AI could be integrated into the cloud, though concerns about governance persist. Technology can be outsourced to the cloud rather than relying on private and institutional infrastructures. One challenge of the cloud is that using third-party IT services can be costly due to current technological limitations. The cost of services depends on compute, storage, and networking. Costs are also influenced by geographic locations, data transfer, security, and compliance requirements.

To help mitigate the costs of AI, industries are considering the private cloud or a third party, such as AWS HPC Compute. This potential opportunity has sparked a boom in the data center market, where significant growth is expected for commercial applications. The demand for AI compute has put tremendous strain on utilities to provide the necessary generation and transmission.

DATA CENTER ARCHITECTURE

The buildings that house the IT equipment must be purpose-built to meet computing performance requirements. The key factors that affect the data center infrastructure include IT systems, electrical systems, mechanical systems, and efficiency. IT equipment power supplies are designed for global use to support various voltages. These power supplies convert AC power into multiple DC voltages suitable for IT equipment. Power supplies can be configured for optimal cost efficiency;

therefore, off-the-shelf IT equipment is typically designed with a low tolerance for voltage deviations.

IT SYSTEMS

The servers, storage, and network equipment require specific power and cooling systems to ensure that the equipment operates as intended. There are national design standards that dictate the necessary cooling and power requirements. The American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) standard TC9.9 governs cooling systems, while the International Electrotechnical Commission (IEC) standard 62040 dictates the power conditioning requirements for data centers. The IEC 62368, along with UL, provides standards for server power supplies.

The Information Technology Industry Council (ITIC) established guidelines for recommended power disruption to IT equipment; thus, many data centers require Uninterruptible Power Supplies (UPS) or supplemental power conditioners for the DC bus in the servers. This helps protect IT equipment from sags, swells, overvoltage, and transients. If the power distribution is less than 20 milliseconds, there is no disruption to IT computing. As the duration of disruption increases, the range of proper operation narrows. The power quality requirements for IT equipment are summarized in the ITIC Curve indicated in Appendix B, Graph 6.

ELECTRICAL SYSTEMS

Data centers are typically designed with a power configuration that ensures high availability, incorporating redundant utility power sources, multiple onsite power sources like generators, and UPS systems with batteries. The UPS systems offer power conditioning and ride-through capabilities to support onsite power generation during utility power disruptions. Furthermore, power supply to IT equipment is usually structured with a distributed redundant configuration, allowing it to receive power from multiple paths. This distributed redundant approach is crucial for maintaining continuous 24/7 operations, enabling electrical components to be serviced while still powering IT equipment.

MECHANICAL SYSTEMS

The mechanical systems for data centers typically use various technologies, such as air-cooled chillers, evaporative coolers, and Direct Outside Air Systems (DOAS). Utilizing mechanical cooling increases the PUE of the data center compared to direct air cooling due to the lower electrical energy consumption of evaporative coolers and DOAS.

Most conventional data centers use air to cool IT equipment, as it provides effective cooling at an economical price point for low-density IT cabinets. However, there are limitations to air cooling, making it less effective for high-density computing, such as in AI data centers. AI data centers require air-cooled chillers, although these consume more total electrical energy than alternative direct air or evaporative coolers.

EFFICIENCY

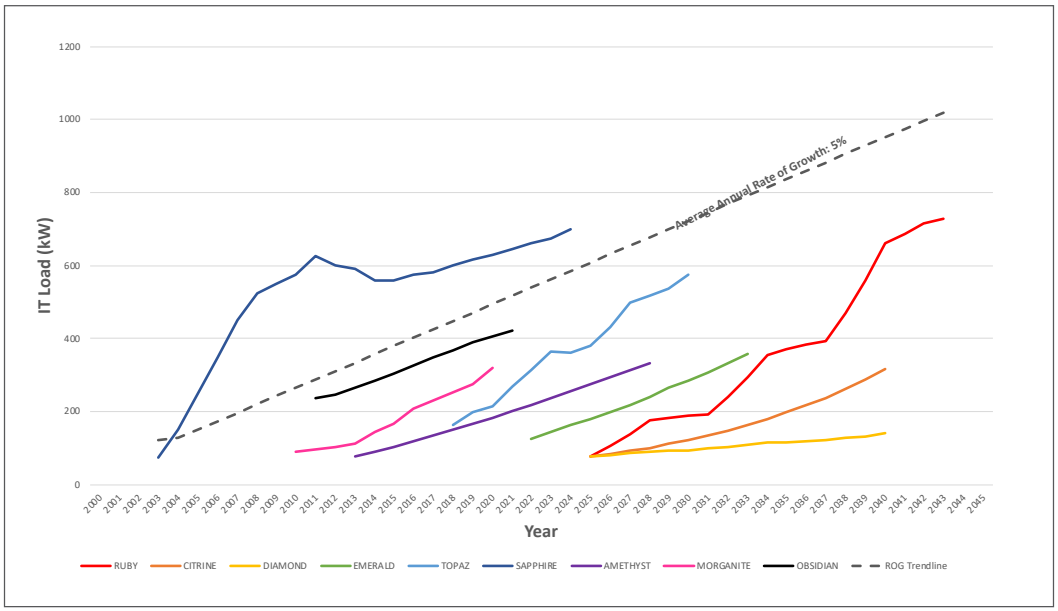
The energy intensity of IT compute decreased exponentially from 2008 to 2022, as seen in Appendix B, Graph 7. A watt reduction in power at the IT compute has a compounding impact on the energy needed for the data center from the grid. More efficient IT equipment will produce less heat and thus require less cooling and less power from the grid. In addition, the parasitic losses on electrical distribution will be lower due to reduced

IT power consumption. Conventional data center cooling has improved over the years to lower total energy consumption.

Historically, data center cooling loads had a one-to-one ratio of energy consumption; for every watt of energy consumed by IT equipment to produce heat, an equivalent watt of energy was needed to cool the data center. Efficiency gains in cooling were achieved through changes in ASHRAE guidelines regarding recommended and allowable temperature ranges for IT equipment. The recommended cooling requirements range from 64.4 to 80.6 degrees F at the inlet of the server, with an allowable range from 59 to 89.6 degrees F. Cooling technologies such as DOAS, Electrically Computed Direct Current (ECDC) motors for fans, and evaporative coolers have reduced mechanical PUE values to about 10% to 40%, meaning one watt of compute heat requires only 0.1 to 0.4 watts for cooling.

Due to their high power density, AI systems utilize a combination of direct-to-chip liquid cooling and air-cooling systems. The chip-level liquid cooling rejects heat to the heat exchanger and ultimately to the mechanical system in a closed-circuit arrangement. The mechanical cooling for AI compute will increase the PUE, where one watt of compute heat requires 0.35 to 0.70 watts for cooling.

Data Center IT Load Growth Projection



In addition to mechanical inefficiencies, electrical power distribution system losses arise from voltage transformation through equipment such as transformers, UPS, and I2R losses. These losses, along with parasitic loads such as jacket water heaters at generators, can total up to 10% loss for every watt of IT compute, where one watt of compute heat equates to 0.1 watt of electrical losses. Efficiency gains from next-generation technologies, such as Silicon Carbide UPS and

battery at the server, and four-level IGBT inverters have enhanced efficiency and reduced energy intensity to 0.05 watts for every watt of compute.

DATA AND METHODS CASE STUDY

To understand the future needs of data centers driven by AI compute, existing data centers were evaluated to assess their development over the past 20 years. These facilities were planned, designed, and operated with an anticipated Rate of Growth (ROG) for IT compute power, as referenced in Table 3 in Appendix A. The analysis examined data centers across the US from various industries, including Healthcare, Institutional, Corporate, Software IT, and Government. A total of nine data centers were evaluated, focusing on their initial IT KW power and the ROG for each facility. The names of the clients have been replaced with assigned semi-precious names for confidentiality reasons. These facilities are summarized in Table 2 and analyzed in Table 3 in Appendix A.

INDUSTRY DATA CENTER IT GROWTH PROFILE

Each data center's utility power varies based on its resilience and PUE. To normalize data from various data centers, only the IT compute power in kW was used along with its respective ROG. The data center's growth profile, kW per Year, is plotted and illustrated in the plot above. The IT power for each data center was normalized by calculating the slope for each data center's trendline.

After performing a linear regression, a year-over-year ROG was established to develop a consolidated average of 5.4%. The consolidated ROG of 9.1% was calculated over an initial 14 years to align with PJM's projected growth profiles. The kW per Year measured across nine data centers also indicates an average power growth of 3.5x for IT, with a median of 3.1x. The year-over-year growth accounts for various technologies deployed in the data center market with differing market functions.

To develop a granular analysis, the Sapphire data center was chosen to conduct a case study that examines the life cycle of the data center. It was selected due to its ample data and historical context to provide a comprehensive perspective on various technological deployments. Given that AI technology is still emerging, the industry is presently creating infrastructure solutions capable of integrating with evolving AI technology and IT systems, which are not represented in the Sapphire data center. This case study seeks to understand how earlier technologies influenced the growth of data centers.

PROGRAMMING QUESTIONNAIRE

The initial process in analyzing the data center involved developing a programming questionnaire discussed during an interview with the data center manager regarding the requirements of the data center operations over the facility's 20-year history. The results of the data center programming questionnaire are summarized below.

The Sapphire data center case study features an IT area of 11,922 square feet, designed for Uptime Tier III resiliency, with an availability of 99.9%. The data center is equipped with redundant on-site generators, UPS systems, air-cooled chillers, and distribution equipment. Currently, the average load is 600 kW, with a peak of 613 kW and a low of 580 kW. This corresponds to a steady-state base load profile of +/- 5% of the average. This is typical when the IT load remains constant (base load) without significant deviations.

> One of the biggest challenges in AI deployment is the limitations of infrastructure (power and cooling) and the costs associated with IT equipment.

The data center in the case study contained IT equipment such as storage (40%), networking (10%), and compute (50%), including applications like SAP and EPIC. The power density for computers, storage, and networks varies from high to low, respectively. The more compute processes are handled, the higher the power requirements. For example, network cabinets typically require less than 1kW per cabinet, while compute cabinets range from 5 to 6 kW per cabinet. The blade server chassis exhibited increased power density for compute, reaching 16kW per cabinet. With higher-density compute cabinets, the total number is reduced to 30% instead of 50% of the data center. The blade chassis servers enabled higher utilization of the server by allowing multiple applications to operate on a single server, thus increasing efficiency gains. According to Upsite Technologies, blade servers can perform the same work as a standard rack-mounted server with 20% less power consumption.[20]

Sapphire data center aligns with the conventional IT refresh cycle of 4 to 5 years, which is two times slower than Moore's Law and hyperscale data centers. This slower refresh cycle results in a delayed transition to higher efficient computing for this data center.

Initially, the data center predominantly supported enterprise applications and networks, but it is evolving with the increased use of HPC and future AI technologies. Their expected growth in AI is estimated to reach 40%, with a projected threefold increase in power density over the first three years. The data center's power density is anticipated to grow fourfold within 5 years and eightfold in 10 years. This anticipated growth does not align with their historical growth patterns and it does not take into account any efficiency gains due to the limitations of current technology and the costs associated with servers.

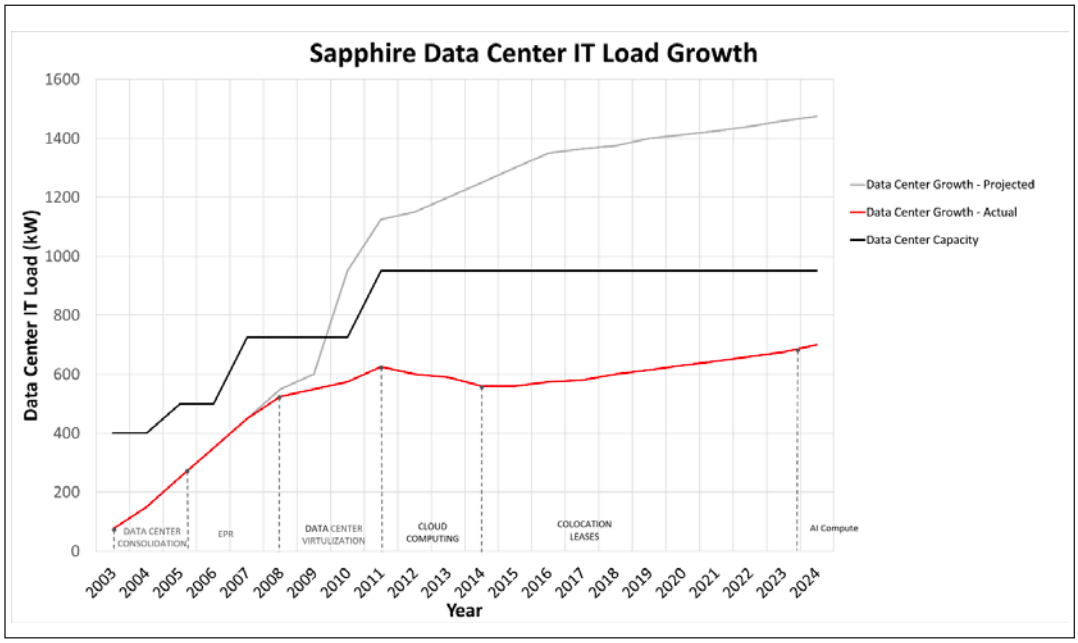
HPC is necessary for AI as they serve different functions in data transaction computing. While AI can support enterprise workloads, there is hesitance to run enterprise applications on AI due to the potential risk of outages. The primary drivers for AI applications are to lead in technology deployment and to attract talent. One of the biggest challenges in AI deployment is the limitations of infrastructure (power and cooling) and the costs associated with IT equipment.

AI is an emerging technology for machine learning and performing tasks such as modernizing processes and learning to grow. However, there are challenges with AI, including the ability to make "gut decisions," security breaches, and bots on the market. A commercial risk of AI is the lack of checks and balances or accountability. Policies and regulations for the use of AI will be necessary to address devious intentions with proper supervision and enforcement.

Due to the critical nature of the data center for businesses, both primary and backup data centers are required to ensure business continuity. Initially, data center redundancy for Sapphire was in Data Center #1 and Data Center #2; eventually, this redundancy was outsourced to a developer-based colocation facility in Northern Virginia (NoVA). This became possible with advancements in connectivity speed, allowing transfers to occur with minimal latency and within acceptable limits for data center operations. Although complete migration is feasible, there are limitations to cloud transfer, such as risk conversion, cloud dependency, security concerns, lack of control over the business, and "colocation is one security breach from a knee-jerk reaction." Data centers are an essential function of businesses, making the security of data critical.

The efficiency gains in the Sapphire data center have been driven by improvements in both IT compute and infrastructure. The initial PUE of the data center was 2.7 with campus central water cooling, which decreased to a PUE of 1.39 with the implementation of dedicated waterless chillers and high-efficiency UPS systems. This change resulted in a reduction in annual electrical billings from $ 1.3 million in 2018 to $700K in 2024.18. The efficiency gains in IT compute and infrastructure provide significant financial benefits due to the data center's 24/7 operation, which continuously consumes energy.

The challenge with AI involves physical elements such as hardware and connectivity. Conventional solutions work for most traditional data center applications. Currently, the competition in AI compute is limited to Nvidia, but increased competition may drive down costs over the next three to five years. Presently, the existing AI compute resources are overdesigned for most applications, yet uncertainty about future needs poses a risk to users if the chip underperforms for anticipated applications. The requirements for AI compute in the future could alter the data center market's electrical demand.



GROWTH PROFILE

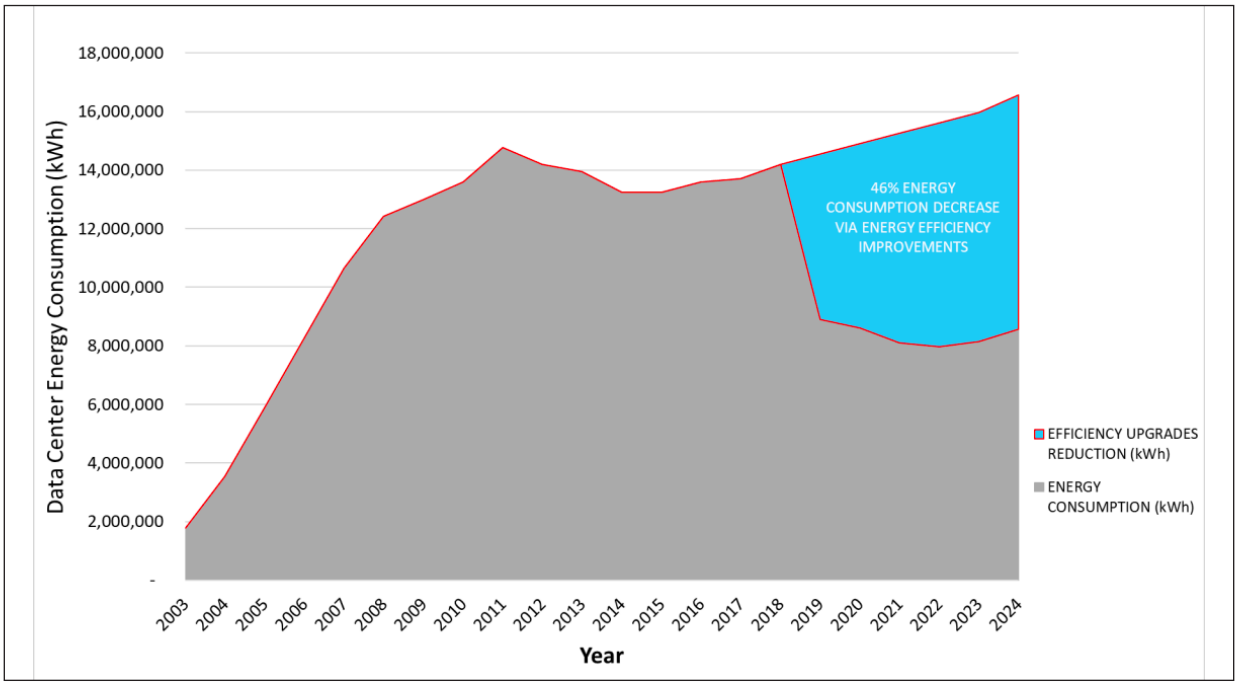
Table 1 in Appendix A indicates the projected growth profile for the Sapphire data center developed in 2003. The growth profile served as a "living" document to help track the data center's growth over the years. The load profile consists of years of operation, planned IT load projected to 2024, and planned ROG. The measured IT load represents the actual IT power consumed by the computers at the data center. The data center capacity column indicates the installed power and cooling infrastructure for IT computing. The total on-site generation depends on IT compute and PUE. The capacity reflects the facility's limitation in delivering the maximum IT that can be installed without exceeding the site infrastructure. The ceiling for the data center IT compute is based on installed on-site generators, chillers, UPS power, and distribution equipment. The plot below shows Table 1 from Appendix A.

SAPPHIRE DATA CENTER IT GROWTH PROFILE

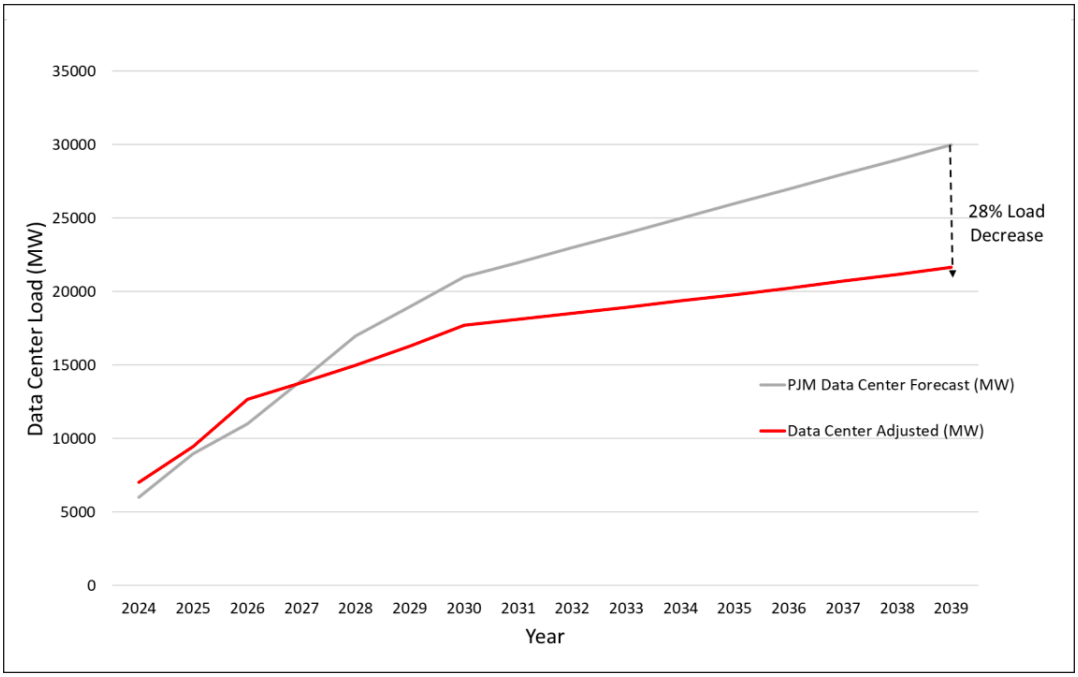
The light gray line represents the planned load growth at the start of the data center's operational date. The black line on the graph indicates the installed capacity of the data center. The sharp rise in the capacity line suggests modifications to the onsite power and cooling systems to accommodate increased IT capacity. The red line shows the metered IT load over time.

There are key milestones throughout the life of the data center that shape the IT compute load and power consumption. When the project was initially constructed in 2003, it was intended to consolidate multiple smaller data centers for the entity. Data center growth was high due to the "field of dreams" syndrome: if you build it, they will come. There was interest among the owner's base in consolidating the smaller IT rooms into a centralized data center facility designed for high-availability

Sapphire Data Center Energy Consumption



Adjusted PJM Data Center Load Forecast



infrastructure. The consolidation was inefficient, and the ROG was significant because the servers and applications were not integrated. The "field of dreams" effect occurred from 2003 to 2005. Once the consolidation was complete, the data center expanded to meet business needs, such as the implementation of Electronic Patient Records (EPR). The growth associated with increased application activity occurred from 2005 to 2007.

Around 2008, new technology was introduced to enhance computer efficiency through Virtualization, Storage Area Networks (SANs), and Direct Access Storage Devices (DASD). Virtualization became feasible with the introduction of blade computers, enabling increased computing capacity and allowing software to minimize the total amount of IT equipment. While virtualization significantly curbs growth, the overall net growth of IT compute has increased.

From 2011 to 2015, Sapphire implemented cloud computing that helped reduce the total IT load. This included offloading some data center computing to third-party service providers such as AWS, Google, and MSFT. The number of IT computers decreased by approximately 8% during that time frame. Goldman Sachs Research shows the industry impact of cloud on energy intensity per compute in Graph 8, Appendix B. Although there is no specific study that demonstrates the effect of outsourcing computers from private data centers to cloud-based systems, the energy intensity of traditional data centers has dropped from approximately 2000 kWh per compute instance in 2015 to around 1300 kWh per compute in 2022. Meanwhile, the energy intensity of the cloud decreased from about 700 kWh per compute to approximately 300 kWh per compute over the same period.

The Sapphire data center also shows that after migration to the cloud, some applications needed to operate in the private data center for various reasons, such as application compatibility with the cloud, security concerns, federal and state legislation, and policies. The steady-state growth declined to an organic growth rate of approximately 2.3% in the absence of any business acquisitions, as indicated in Appendix A Table 1.

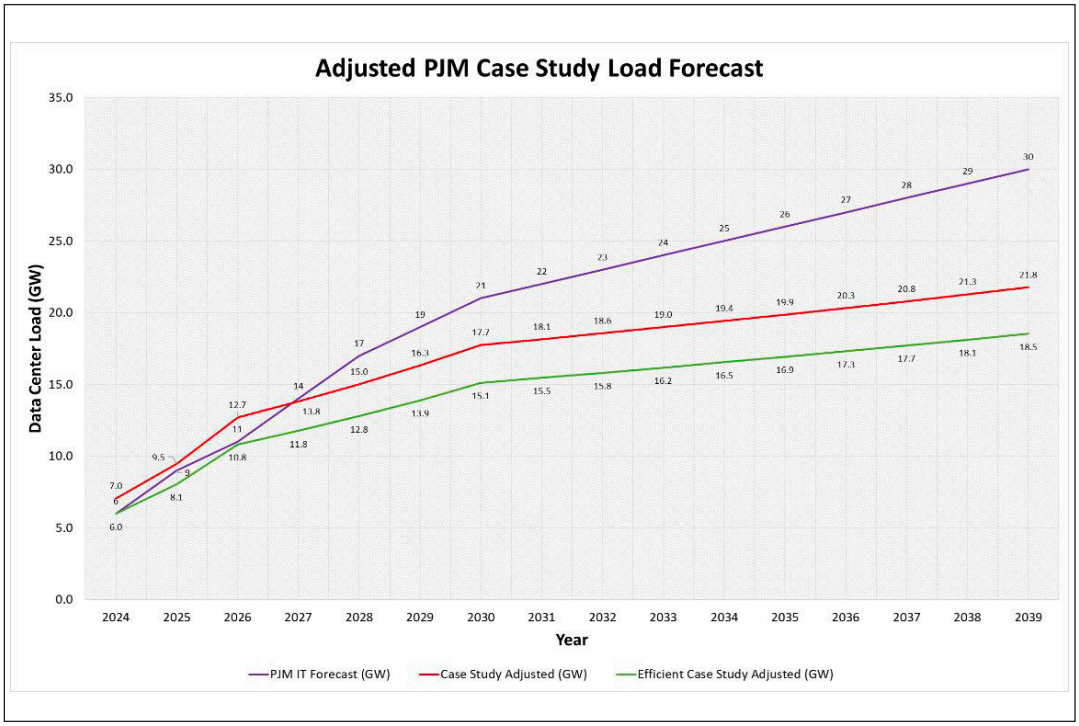
EFFICIENCY GAINS

Data center power efficiency has improved tremendously from a historical level of 3.0 PUE (Power Usage Efficiency) to the current hyperscale data center at approximately 1.1 PUE. Our analysis assumes a blended PUE of 1.35 for AI and conventional data centers, as some loads can be cooled with conventional air-side cooling of 1.1 PUE, while a percentage of data center loads will need to be cooled by mechanical systems such as chillers.

The impact of PUE has a significant effect on the total power observed by the grid and onsite generation. In 2019, Sapphire implemented an infrastructure upgrade, reducing the operational PUE from 2.5 to 1.39. The total annual electrical energy reported by the grid decreased from approximately 16,556 MWh to about 8,575 MWh. This represents a reduction of around 46% in energy consumption, totaling approximately 42,508 MWh over 5 years. The gains in energy savings are graphically represented below, based on the raw data in Table 6 in Appendix A.

SAPPHIRE DATA CENTER ENERGY CONSUMPTION REDUCTION

The energy savings indicated above equate to 6,204 gasoline cars driven for a year, according to the Greenhouse Gas Equivalencies Calculator from the US EPA.



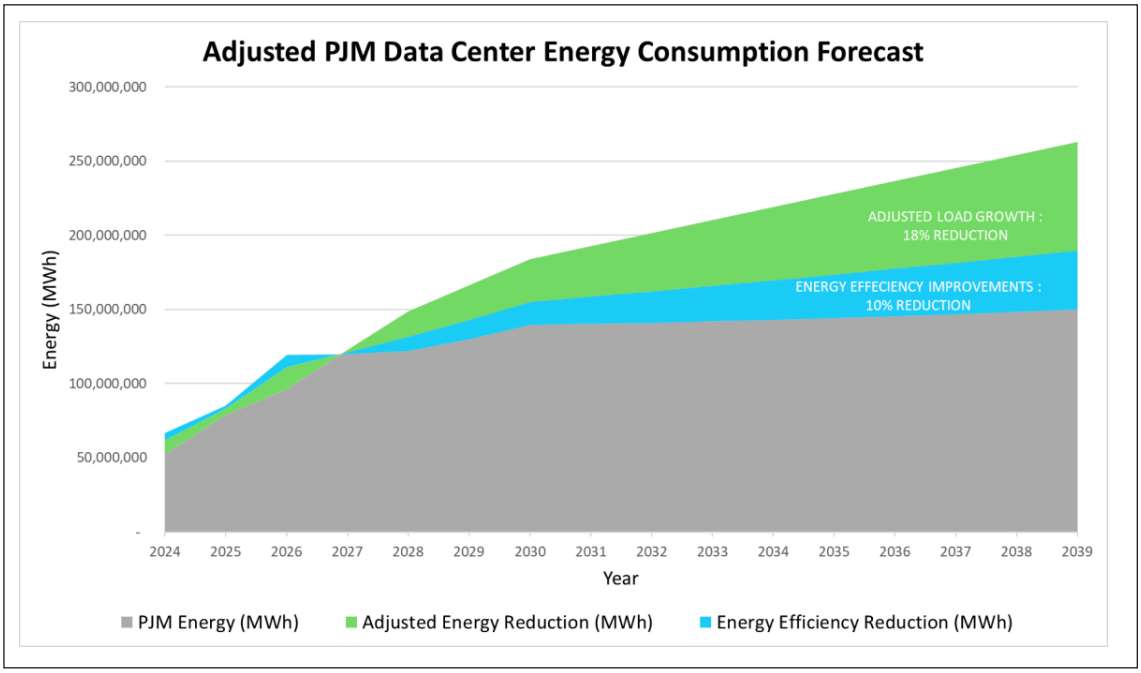
Given the impact of efficiency gains from the Sapphire data center, the efficiency improvements for hyperscale facilities can significantly affect operational costs and carbon footprints due to their size and scale. As shown in Table 5 in Appendix A, there is approximately a 28% total power savings if efficient power and cooling systems are applied to the projected PJM data center load forecast. This is graphically represented below.

FUTURE AI IMPLEMENTATION

Table 4 in Appendix A displays the IT growth profile based on the PJM forecast in GW for data center power from 2024 to 2039. The values represent the total power requirements observed by the grid associated with data centers. To normalize the data, the IT computer was developed using an average PUE of 1.15. The PJM IT computer's rate of growth was calculated year over year. The ROG related to Sapphire was employed to determine the revised projected IT load.

The projected IT load was multiplied by the projected PUE for the AI data center to estimate the anticipated grid power. The PJM power requirements relative to the Sapphire case study load requirements are illustrated in the plot below. The net delta between the PJM data center growth profile and Sapphire is 72.5%. If an efficient power and cooling solution, such as cloud-based data centers, is implemented, the final anticipated PJM load profile is reduced to 61.8%, as referenced by the green line below.

The energy consumption based on the load profile is presented in Table 7 of Appendix A. The calculation details the energy consumption of the projected PJM load, the Adjusted Energy Reduction based on the anticipated Rate of Growth (ROG) for Sapphire Data Center, and the ROG assuming efficient cooling systems are implemented. The average annual energy savings are projected to be 18.3% for the adjusted ROG, with an additional 9.9% savings possible if efficient cooling



technologies are deployed in AI data centers. The plot below illustrates the energy savings opportunities.

There are significant potential savings in energy based on the current trajectory of power requirements for AI compared to historical trends in data center growth. To validate the growth profile of the Sapphire data center, the average ROG is 8.7%, which aligns with an industry average of 9.1% but is less than the 12% proposed by PJM. Table 4 in Appendix A indicates the total power density growth, another method used to validate the anticipated IT ROG. PJM anticipates total power growth at a factor of 5x from 2024 to 2039. Alternatively, both Sapphire and the industry's average power density growth are approximately 3.3x and 3.5x.

Extrapolating the savings from PJM power projections, the total US industry's projected data center load, based on ten planned regions of 128 GW, should be lowered by 28% to 92 GW. If efficient cooling technology is deployed, the anticipated load capacity should be reduced by 38% to approximately 80 GW. The analysis is summarized in Table 4 of Appendix A.

DATA CENTER IMPACTS

The electrical infrastructure, a product of 18th-century technology implemented for human and industrial consumption, is unable to adjust to the 21st-century technological needs of computers. The rapid deployment of data centers for the cloud and AI creates challenges due to market growth, tariff structures, and grid resiliency.

MARKET GROWTH

NoVA has the highest concentration of data center power requirements in the US. There has been a 500% increase in demand for data centers from 2015 to 2023.11. The high power requirements from AI are driving data center power needs up to 1 GW of power per site. PJM projects a summer forecast of a 23% increase in grid demand over 15 years due to various loads, such as electrification, industrial, hydrogen with data centers being the largest percentage.

The high power requirements for data centers and their speed to market are creating challenges for grid infrastructure. The challenge arises not only from electricity generation but also from limitations in transmission and distribution infrastructure due to restricted spatial diversity. Data centers are spatially concentrated compared to other industries, such as steel plants and warehouses. Data centers are built in specific regions based on factors including available power, fiber connectivity, latency, and municipal support through permitting and tax incentives. The US map, refer to Appendix B Map 1, indicates current and anticipated data centers with the highest density of data centers is found near NoVA and the Pacific Northwest.

TARIFF

Due to high power density requirements for data centers, utilities are struggling with capital outlay for necessary infrastructure upgrades in generation and transmission. Since utilities typically generate revenue through tariff rates, the conventional data center provides steady demand loads, and costs can be passed through energy consumption with initial connection charges.

AI data center loads resemble those of crypto data centers, where the load is variable and can range from zero to 100 percent depending on application requirements. The demand load can be adjusted according to the time of use to optimize tariff costs

for consumers. The traditional tariff based on energy consumption is not effective in capturing the demand capacity that will be required for AI data centers. To address the variability of the AI load, American Electric Power (AEP) has proposed tariffs on data centers as part of the NERC Large Load Task Force (LTTF). The proposal would mandate a 12-year contract with minimum monthly charges. Additionally, there would be an exit fee equivalent to five years of minimum charges.

According to Energy.gov, total data center electricity consumption accounts for approximately 4.4% of total US electrical consumption. Projected growth is anticipated at a parabolic rate, varying from 6.7% to 12.0% of total electricity consumption, as indicated in Appendix B, Graph 9. Total energy consumption is expected to increase from about 200 TWh to nearly 600 TWh over the next five years. Projected energy consumption is likely to double or triple rapidly in an industry that has historically consumed less than 100 TWh. Investments by utilities will create a burden in developing new generation facilities and building transmission grids to accommodate the increased demand and energy consumption. The challenge for utilities lies in the speed of energy delivery from an industry that operates within a highly regulated environment through the Public Utility Commission.

GRID RESILIENCY

The large loads associated with hyperscale data centers have the potential to destabilize the grid. Based on studies by NERC, there are concerns about historical grid instability due to large loads, such as data centers impacting transmission voltages. According to NERC, a case study shows that the large data center step loads could impact a loss of transmission voltage and affect customers served from the connected transmission line within 100 miles of the event. The risk posed by large load variability, whether through transient events or step load response, poses a significant threat to the national grid.

The data center's transient load profile can be viewed as a simulated profile in Appendix B, Graph 11. A significant shift occurs when the data center load is switched to onsite generation due to grid disturbances or load curtailment modes. Qunata Tech's simulated analysis indicates that data centers will experience steep load responses throughout the year because of grid disturbances or maintenance. The 24-hour profile demonstrates a stable load pattern with minimal fluctuations, which corresponds to the prior event documented by AEP, as shown in Appendix A, Graph 10.

The anticipated load profile of the conventional data center, compared to that of the AI data center, demonstrates a simulated load variability as indicated in Appendix B, Graph 11. Over a three-hour period, the expected load of the AI data center will fluctuate by 25%. The load fluctuations in the graph are shown in 15-minute increments. The electrical noise sent back to the grid from the IT equipment can vary up to 1.5 times the inrush current in increments of milliseconds. This increased current can reduce the life expectancy of the power electronics and battery systems commonly found in data center electrical systems and can destabilize the grid and/or onsite generation. Consideration of supplemental power conditioning, such as supercapacitors and/or BESS, should be included to stabilize the grid. This additional infrastructure requirement will lead to increased capital costs for supporting the data center loads.

DISCUSSION

To address the high power demand requirements for the data center, on-site generation will be necessary to mitigate demand loads and provide time for the grid to expand. On-site generation, including DER (BESS, fuel cells, gas generation), will be essential for supplying bridging power. The on-site generation system can also help stabilize the grid as AI load fluctuates due to application demands or grid disturbances from other high-load users, such as data centers and industrial plants.

The practicality of onsite renewable energy is not viable for data centers due to the high-density load profile of data centers and the low-density power output from renewable infrastructure. Future solutions that may be viable include SMR, hydrogen fuel cells, gas generation, molten salt batteries, and/or pumped hydro (site dependent). SMR is the most pragmatic solution for hyperscale data centers because of its higher power density. These SMRs typically reach up to 300 MW, are cost-efficient compared to larger reactors, and reduce grid demand. Onsite generation is also more energy efficient than the conventional generate, distribute, and consume model due to savings in line losses from the generation source to consumption locations.

Grid management is another option to optimize interconnection. Google recently indicated that it is working with PJM to deliver Google AI with Tapestry tools and insights to process new project applications, which will reduce the time required and help new capacity come online faster.20 The primary goals are to bring more energy capacity onto the grid promptly, drive efficiency and affordability, and integrate diverse energy resources. As more DER solutions become mainstream, utilities will need to transition from conventional power generation and distribution to energy management. Data centers with on-site generation can provide load flow from multiple sources, which can function as power producers or power consumers based on real-time grid demands.

CONCLUSION

AI data centers are constantly evolving due to advancements in software and hardware for IT computing. This ongoing change makes it difficult to predict how data centers will affect the grid. As demonstrated through the Sapphire data center case study and broader industry trends, enhancements in IT hardware and cooling technologies have historically reduced energy consumption growth. Forecasts, such as those from PJM, may overestimate future grid demand if they fail to account for expected efficiency gains and disruptive innovations. This will burden the IPPs and utilities with stranded capacity and long-term investments without returns. Unfortunately, consumers will bear the cost burden through tariffs.

Efficiency gains can already be observed with technological solutions such as reversible computing. Reversible computing has the potential to reduce energy consumption by four thousand times as it becomes commercially viable. Other transformative changes include quantum computing and energy-efficient compute alternatives such as Deep Seek. The highly efficient cooling systems currently deployed in cloud data centers should be leveraged to further reduce energy consumption in AI data centers.

Alternative software solutions and network connectivity that allow for lower latency may enable data centers to be spatially distributed

rather than concentrated in specific regions. This spatial diversification can help leverage underutilized areas of the grid and reduce competing demands, such as edge computing, building electrification, smart cities, and electric vehicles.

Regardless of efficiency gains, the deployment of AI will increase and create grid challenges due to high power density and a volatile load profile. On-site generation, including SMRs and distributed energy resources, will be critical in bridging the gap between grid capacity and AI demand. Collaboration among data center developers, utilities, and regulators must replace the siloed approaches of the past. The future of data center energy use will not be determined solely by AI's growth but by our collective ability to deploy smarter, more sustainable infrastructure.

A pragmatic approach to AI growth must be considered, involving strategic investments, flexible energy policies, and the adoption of emerging technologies. Solutions to support AI energy needs are not solely the responsibility of data center companies; rather, they require a partnership among utilities, regulators, and industry to ensure that AI becomes an innovation that propels us toward a cleaner, more efficient energy future instead of a burden on the grid.

ABOUT THE AUTHOR

Rajan Battish PE, ATD, LEED AP has more than 30 years of experience in innovative design and project management on data centers and mission critical projects. A recognized industry expert in power infrastructure and supply distribution, Rajan is a frequent speaker and author on energy efficiency, resiliency and intelligent automation.

Rajan leads RSP's Baltimore office and oversees a multi-disciplinary team of talented, experienced professionals who have designed, engineered and delivered mission critical facilities for public utilities and financial institutions, healthcare systems, academic and university-related research facilities as well secure and defense-related agencies.

SOURCES AND CITATIONS

1 "Is Nuclear Energy the Answer to AI Data Centers' Power Consumption? | Goldman Sachs." Home | Goldman Sachs, 23 Jan. 2025, https://www.goldmansachs.com/insights/articles/is-nuclear-energy-the-answer-to-ai-data-centers-power-consumption.

2 "Generational Growth AI, Data Centers, and the Coming US Power Demand Surge." Goldmansachs.Com, Goldman Sachs, 28 Apr. 2024, https://www.goldmansachs.com/pdfs/insights/pages/generational-growth-ai-data-centers-and-the-coming-us-power-surge/report.pdf.

3 Mooney, Molly. "Load Forecast Development & Use n PJM." In.Gov, PJM Resource Adequacy Planning Department, 6 June 2024, https://www.in.gov/iurc/files/IN_meeting_PJM-Load-Forecast_06062024.pdf.

4 Wilson, John D., et al. Strategic Industries Surging: Driving US Power Demand. GridStrategies. Accessed 25 Apr. 2025.

5 IEA analysis based on Masanet et al. (2020), Malmodin (2020), Hintemann & Hinterholzer (2022) and reported energy use data from large data centre operators.

6 Gallagher, Dan. "Meta and Microsoft: AI Spending Champs Won't Be Tapping the Brakes." Wall Street Journal, 1 Nov. 2024.

7 "Incident Report: Considering Simultaneous Voltage-Sensitive Load Reductions." Nerc.Com, North American Reliability Corporation, 8 Jan. 2025, https://www.nerc.com/pa/rrm/ea/Documents/Incident_Review_Large_Load_Loss.pdf.

8 Shehabi, A., Smith, S.J., Horner, N., Azevedo, I., Brown, R., Koomey, J., Masanet, E., Sartor, D., Herrlin, M., Lintner, W. 2016. United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory, Berkeley, California. LBNL-1005775

8 "Inside Amazon's AI cloud Strategy." Wall Street Journal. Accessed 24 Apr. 2025.

9 IEA (2024), Efficiency improvement of AI related computer chips, 2008-2023, IEA, Paris https://www.iea.org/data-and-statistics/charts/efficiency-improvement-of-ai-related-computer-chips-2008-2023, Licence: CC BY 4.0

10 "PJM Long-Term Load Forecast Report." www.Pjm.Com, PJM Resources Adequacy Planning Department, 25 Jan. 2025, https://www.pjm.com/-/media/DotCom/library/reports-notices/load-forecast/2025-load-report.pdf.

11 Stone, Adam. "Inside the Divisive Debate Surrounding Northern Virginia's Data Centers." Northernvirginiamag.Com, 9 Apr. 2025, https://northernvirginiamag.com/news/2025/04/09/inside-the-divisive-debate-surrounding-northern-virginias-data-centers/#:~:text=%E2%80%9CWe%20have%20been%20letting%20data,-Virginia%20Technology%20Council%20(NVTC).

12 IEA (2024), Spatial concentration index for selected infrastructure categories, 2010, IEA, Paris https://www.iea.org/data-and-statistics/charts/spatial-concentration-index-for-selected-infrastructure-categories-2010, Licence: CC BY 4.0

13 Bell, David., Bauer, Rich. Planning and Operating the Grid of the Future. American Electric Power, 2024, https://www.nerc.com/comm/RSTC/LLTF/LLTF_Kickoff_Presentations.pdf.

14 Shehabi, A., Smith, S.J., Hubbard, A., Newkirk, A., Lei, N., Siddik, M.A.B., Holecek, B., Koomey, J., Masanet, E., Sartor, D. 2024. 2024 United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory, Berkeley, California. LBNL-2001637

15 "Evolving Load Characteristics and Reliability Consideration ." Quanta-Technology.Com, NERC, 14 Nov. 2024.

16 Porat, Ruth. "Our Investment in AI-Powered Solutions for the Electric Grid." Google, Google, 10 Apr. 2025, https://blog.google/inside-google/infrastructure/electric-grid-ai/.

17 Frank, Michael P. "The Future of Computing Depends on Making It Reversible - IEEE Spectrum." IEEE Spectrum, IEEE Spectrum, 25 Aug. 2017, https://spectrum.ieee.org/the-future-of-computing-depends-on-making-it-reversible.

18 Data Center Programming Questionnaire, Company Sapphire, Data Center Manager, October 30, 2024

19 IEA (2024), Investment in data centres in the United States, January 2014 to August 2024, IEA, Paris https://www.iea.org/data-and-statistics/charts/investment-in-data-centres-in-the-united-states-january-2014-to-august-2024, Licence: CC BY 4.0

20 Partida, Devin. "3 Energy Efficient Strategies to Consider for Data Centers." Upsite Technologies - Data Center Cooling Optimization, https://www.facebook.com/UpsiteTech/, 19 Jan. 2022, https://www.upsite.com/blog/3-energy-efficient-strategies-to-consider-for-data-centers/.