Integrating AI-Optimized Data Centers into the U.S. Electric Grid: Impacts and Mitigation Strategies

Omar Ghabayen
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213 USA
Email: oghbayan@andrew.cmu.edu

***Abstract*—This paper examines how the rapid expansion of artificial-intelligence-optimized data centers, with campus loads on the order of hundreds of megawatts and steep ramp profiles driven by cooling and information technology demands, is changing the planning and operation of the United States electric grid. These large facilities are creating unprecedented local and regional load growth and are challenging traditional power delivery and resource planning paradigms. We connect fundamental power systems concepts such as power flow limits, resource adequacy, and ancillary services to recent trends in data center siting, including cluster growth in Northern Virginia and Central Texas, to show why concentrated nontraditional load profiles can stress substations and transmission infrastructure. For a representative AI data center campus that incorporates realistic power usage effectiveness and cooling requirements, we estimate the feeder and transformer capacities required, the impacts on peak demand and annual energy consumption, and the potential benefits of on-site battery storage for ride-through and fast frequency response. We then embed the campus in a simplified dynamic model based on the swing equation to quantify how large power electronic loads affect frequency stability. Using actual locational marginal prices from the PJM Dominion zone, which covers Loudoun County, Virginia, we evaluate the economic value of a 100 MW battery that participates in both peak shaving and energy arbitrage during the 50 highest price hours of the year. The paper concludes by discussing technical and policy mitigation strategies, including grid-interactive uninterruptible power supply systems, demand response for data centers, and fair cost allocation for grid upgrades. The goal is to provide a quantitative and accessible demonstration of how surging AI-driven demand can be integrated into power grids without sacrificing reliability, economic viability, or decarbonization progress.**

***Index Terms*—AI data centers, electric power systems, resource adequacy, ancillary services, battery energy storage, demand response, swing equation**

I. Introduction

Data centers have become one of the fastest-growing electricity loads in the United States, driven largely by the computing needs of artificial intelligence workloads. In 2023, data centers were estimated to consume roughly 4% of all United States electricity, and this share could more than double to around 9% by 2030 under high growth scenarios [1], [2]. Unlike the moderate and diffuse load growth observed in many past decades, this new wave of demand is both rapid and spatially concentrated. About 80% of United States data center load is located in only 15 states, and Virginia alone was estimated to have roughly one quarter of its total electricity consumption serving data centers in 2023 [2].

Loudoun County in Northern Virginia, served by the PJM Dominion transmission zone, is the archetypal example. The region has become known as Data Center Alley, hosting hundreds of facilities that collectively draw several gigawatts of power [1]. Similar clusters are emerging in Central Texas in the ERCOT system. These regions offer favorable land, strong fiber connectivity, and historically low electricity prices. At the same time, the concentration of large power-intensive facilities is straining local grids from distribution substations up through the bulk transmission network and has raised concerns about capacity shortfalls, congestion, and reliability impacts [1], [2], [6].

AI-optimized data centers differ from traditional enterprise data centers in both physical scale and operational profile. They deploy dense clusters of graphics processing units and other AI accelerators, which can push rack power densities into the range of 30–100 kW per rack, well above the 7–10 kW per rack that has been typical in conventional facilities [3]. These centers tend to draw near maximum power on a continuous basis in order to support large training and inference workloads. Rather than following a diurnal or business cycle pattern, their demand profile is characterized by high utilization around the clock. The combination of very high peak power and sustained operation produces a substantial base load on the grid with relatively little variation in time.

In addition, AI data centers rely heavily on power electronics in rectifiers, inverters, and variable-speed drives for cooling equipment. These devices have very limited physical inertia compared to traditional rotating loads and can exhibit different dynamic behavior during disturbances. Large-scale integration of power electronic loads can introduce harmonics, reduce effective system inertia, and complicate voltage and frequency control [3], [7].

From a power systems planning perspective, the surge in AI data center demand intersects several fundamental topics from an upper-division power engineering curriculum. At the distribution level, the ability of existing substation transformers and medium-voltage feeders to deliver tens or hundreds of megawatts to a single campus is often limited. New data centers may require feeder reconfiguration, reconductoring, or construction of entirely new substations. Power flow and

voltage drop constraints must be reevaluated when a large site is added, particularly when multiple sites cluster in the same region.

At the generation capacity level, rapid load growth raises resource adequacy concerns. Utilities and grid operators must ensure that sufficient firm generation and import capability exist to serve these new loads during peak hours without eroding planning reserve margins. National assessments indicate that by 2030 an additional on-peak supply in the range of tens of gigawatts may be required to support data center and AI load growth [1], [6]. The Department of Energy projects that roughly 50 GW of new firm capacity may be required solely to serve data centers by 2030 [6]. This is a significant challenge because data centers are often built in less than two years, while new generation and high-voltage transmission can take many years to permit and construct.

At the operations and reliability level, continuous high-reliability power draw from AI data centers means they participate very little in traditional demand-side fluctuations. Their presence can exacerbate net load peaks, especially in systems with high solar penetration where the most critical period occurs in the evening after solar output declines. The requirement to maintain ancillary services such as frequency regulation, spinning reserves, and voltage control also grows with larger and more concentrated loads. On the other hand, there is increasing interest in viewing data centers as potential flexible resources. With appropriate control and incentives, they may curtail noncritical workloads, shift computation geographically, or use on-site backup systems to ride through disturbances and temporarily relieve grid stress [5], [8], [9].

To respond to these issues, a variety of mitigation and integration strategies is now being explored. Some large data center operators are directly investing in clean on-site generation, such as solar or wind with storage, or considering small modular reactors, in order to supply part of their demand with local resources and reduce net draw from the public grid [1]. Others are coordinating with utilities on demand response programs in which the data center agrees to shed load or temporarily switch to backup power during system peaks or emergencies in exchange for compensation [5], [8]. There are also emerging demonstrations of using the facility infrastructure itself to provide grid support services. Advanced uninterruptible power supply systems that use batteries can be configured to provide fast frequency response and peak shaving when spare capacity is available [5], [7]. These approaches blur the line between load and resource and begin to turn data centers into grid-interactive assets rather than purely passive customers.

This paper investigates the impacts of AI-optimized data center integration along these axes and evaluates both technical and policy mitigation strategies. Section II provides background on key concepts, including data center design and efficiency metrics, uninterruptible power supply behavior, and resource adequacy fundamentals. Section III develops a more rigorous analytical framework that includes frequency dynamics based on the swing equation, a numerical net load model, and an energy arbitrage model using PJM prices. Section IV presents a case study of a 100 MW AI campus in Loudoun County. Section V discusses mitigation strategies, Section VI examines policy and planning implications, and Section VII concludes with key takeaways for power system engineers.

II. Background

A. AI Data Centers and Power Demand Characteristics

The power demand of AI-focused data centers is defined by both magnitude and shape. Modern hyperscale AI facilities are often built as campuses that draw hundreds of megawatts of power, which is comparable to the load of a small city [1], [2]. They concentrate servers and accelerators at much higher densities than conventional data centers, which greatly increases power consumption per unit floor area. A single AI training rack can require between 30 and 100 kW, while legacy enterprise data centers typically operated at about 5–10 kW per rack [3]. As a result, a single data hall can reach tens of megawatts of information technology load in a footprint that once carried only a few megawatts, with corresponding stress on local electrical infrastructure.

AI workloads such as training large language models tend to run continuously and at high utilization. Unlike many commercial or industrial loads with clear daily peaks and valleys, an AI-optimized data center often operates close to peak capacity 24 hours a day. This high load factor means the facility behaves almost like a base load plant from the grid perspective. Although the steadiness simplifies forecasting, it reduces regional load diversity and raises both nighttime and daytime minimum loads.

Another important characteristic is limited interruptibility. Because AI jobs are power intensive and may run for many hours, and because many facilities also serve latency-sensitive cloud services, operators are reluctant to throttle or pause workloads except in extreme situations. In ordinary operation the load is therefore not easily dispatchable downward. AI data centers present large, steady, and largely non-sheddable loads, and their integration requires robust upstream infrastructure and careful planning.

Geographical clustering further amplifies the impact. When multiple massive data centers locate close to one another, as in Northern Virginia, they can collectively impose hundreds or even thousands of megawatts of new load on a single transmission zone. This can strain substation transformers, medium-voltage feeders, and regional transmission import capacity. Utilities may need to accelerate capital investments, including new substations, additional high-voltage lines, and higher capacity transformers, in order to accommodate the growth [1], [2]. The concentration of load also alters power flow patterns and may require network reconfiguration and updated contingency studies to ensure acceptable voltage profiles and thermal margins under both normal and contingency conditions.

B. Energy Efficiency and Power Usage Effectiveness

Improving energy efficiency is an essential strategy for mitigating the impact of data centers on the grid. The industry standard metric for data center efficiency is power usage effectiveness (PUE). PUE is defined as the ratio of total facility power consumption, including information technology equipment, cooling, lighting, and support systems, to the power used by the information technology equipment alone [1], [4]. An ideal PUE is 1, meaning all energy goes directly to computation. Historically, many legacy data centers operated with PUE values between about 1.5 and 2.0, so each kilowatt drawn by servers required an additional 0.5–1 kW for overhead.

Over the past decade, major operators have reduced PUE through improvements in cooling technology, airflow management, and power distribution. State-of-the-art hyperscale facilities now routinely achieve PUE values near 1.1, and some report fleet averages around 1.08 [4]. A lower PUE means that a data center draws less total power for the same information technology workload, which directly reduces its impact on generation and delivery infrastructure.

Strategies to reduce PUE include more efficient cooling systems such as direct-to-chip liquid cooling or advanced air handling, optimization of airflow and temperature setpoints, waste heat recovery where climate allows, and high-efficiency power conversion equipment including modern uninterruptible power supply units and server power supplies [1], [3], [4]. For AI-intensive facilities, cooling is often the dominant overhead load because accelerators operate at high power densities and generate large amounts of heat. Liquid cooling and immersion cooling are particularly promising for these environments and can significantly reduce the energy used by chillers and fans. Workload management, for example shifting nonurgent computation to cooler nighttime hours, can provide additional modest improvements in energy efficiency.

Even with aggressive efficiency measures, the absolute magnitude of AI data center demand remains high. For example, at a PUE of 1.1, a facility that delivers 30 MW of information technology power still requires about 33 MW from the grid. Energy efficiency therefore reduces but does not eliminate the need for substantial grid capacity. Nevertheless, maintaining low PUE is critical for minimizing wasted capacity, because every kilowatt saved in cooling or other overhead is a kilowatt that does not need to be generated, transmitted, and transformed. Some states and utilities have begun to encourage or require new large data centers to meet specified efficiency thresholds as part of interconnection agreements or incentive programs, explicitly linking energy efficiency to system adequacy [4], [6].

C. Uninterruptible Power Supply Systems and Grid Interaction

Large data centers depend on uninterruptible power supply (UPS) systems to maintain continuous service through power disturbances. A typical UPS architecture includes rectifier and charger units, battery strings or other storage, and inverters that provide conditioned power to the information technology load. The UPS can supply power immediately if the main utility feed experiences an outage or voltage sag, bridging the gap until backup generators start or the grid recovers. In normal operation many UPS systems also perform continuous double conversion, which can slightly reduce efficiency, although modern designs often include energy saving modes [5].

From a grid perspective, the UPS equipment in large data centers represents a substantial amount of installed energy storage that is idle most of the time. A facility with a load of 20 MW may have several megawatt-hours of battery capacity available for a few minutes of full-load operation. Traditionally, this capacity has been reserved entirely for emergency backup. Recent research, however, has explored using inverter-driven loads and UPS batteries to provide grid support services such as fast frequency response [7]. With appropriate control and communications, data center UPS systems can be aggregated to deliver fast frequency response and peak shaving. If system frequency suddenly drops, UPS inverters can inject power by discharging batteries for a short interval, helping to arrest the frequency deviation. Likewise, they can charge or discharge strategically to reduce the facility contribution to daily peaks. Pilot projects, including one at a Microsoft data center in Dublin, have demonstrated that grid-interactive UPS operation is technically feasible and can earn revenue in frequency regulation markets [5].

Any such operation must be carefully designed so that the primary backup function is not compromised. Typically, a minimum state of charge is maintained for emergency use, and only the remaining capacity is exposed to grid services. Protection and interconnection standards also apply. The inverters that interface the UPS to the grid can be configured for grid-friendly behavior such as near-unity power factor operation, reactive power support, and voltage ride-through in accordance with IEEE and IEC standards [3], [5]. They also must limit the injection of harmonic currents to comply with standards such as IEEE 519.

UPS systems are therefore demand-side technologies that share many characteristics with distributed energy resources. If their capabilities are leveraged through grid-interactive control, they can help mitigate the impact of data center loads by providing ancillary services and by reducing peak demand, effectively turning a necessary reliability investment into a dual-purpose grid asset [5]–[7].

D. Resource Adequacy and Reliability Implications

The rapid growth of data center load highlights the importance of resource adequacy, which is the ability of a power system to supply aggregate customer demand at peak times with a sufficient reserve margin. In United States regulatory practice, regional transmission organizations, independent system operators, and vertically integrated utilities conduct resource adequacy studies to verify that enough generation, storage, and demand response resources will be available under extreme conditions, such as heat waves or cold snaps,

while meeting a specified reliability criterion often expressed as an expected number of loss-of-load events per decade [6].

Large incremental loads, such as a 100 MW AI campus that operates continuously, can materially shift a region peak demand forecast and reduce available reserves. For example, if a region previously had a planning reserve margin of 15% above its peak demand, several new data centers could erode that margin and force new power plant construction or additional capacity market purchases. The Department of Energy has warned that the existing grid is not fully prepared to meet the projected energy demands of AI and data centers without significant investments in firm capacity and grid upgrades [6]. The same report estimates that roughly 100 GW of new capacity may be needed nationwide by 2030 to preserve reliability, with about half of this requirement attributable to data centers and related computing loads.

The type of capacity also matters. Because AI data centers require highly reliable power around the clock, they depend on firm dispatchable generation such as natural gas, nuclear, or storage with sufficient duration, especially during periods when variable renewable output is low. If new data center loads are met mainly with intermittent resources, additional balancing and storage will be necessary to maintain reliability during calm or cloudy conditions. This reality complicates efforts to retire older thermal plants for decarbonization reasons, since retirements must be balanced against new load growth.

Maintaining reliability also involves operational readiness for high load and contingency scenarios. Grid operators may need to carry more frequency regulation reserves and adjust operating criteria when large data centers come online. Sudden loss of a large data center, for example if many facilities disconnect due to a voltage disturbance, can produce a frequency spike that is the inverse of the more familiar generator trip event [3], [6]. Operators must have procedures and resources in place to manage such events, potentially including fast-acting storage and coordinated load shedding. Some jurisdictions have begun to explore tariffs and contracts under which data centers agree to operate on backup generators during system emergencies, effectively providing demand response by removing themselves from the grid during critical hours [5], [8]. Although this approach raises environmental concerns when diesel generation is used, it underlines the lengths to which planners may need to go in order to protect system reliability.

III. Methodology

This section develops a more rigorous methodology for quantifying the impact of an AI-optimized data center on both steady-state and dynamic behavior of the grid. The focus is on a representative campus in Loudoun County, Virginia, connected to the PJM Dominion zone.

A. System Model and Transformer Parameters

We consider a campus that delivers 100 MW of information technology power at full utilization. The design power usage effectiveness is taken as 1.2, which is representative of modern AI-focused facilities with advanced cooling [4]. The total facility demand at peak is then

$$P_{\text{tot}} = \text{PUE} \cdot P_{\text{IT}} = 1.2 \times 100 \text{ MW} = 120 \text{ MW}. \tag{1}$$

The campus is supplied from the 230 kV transmission system through a dedicated 230 kV to 34.5 kV substation transformer bank. Manufacturer data for transformers in the 120–150 MVA range indicate typical short-circuit impedances of about 10% on the transformer base, with X/R ratios on the order of 40 [12], [13]. We model each transformer with

$$Z_{\text{pu}} = R_{\text{pu}} + jX_{\text{pu}} \approx 0.003 + j0.10, \tag{2}$$

on a 150 MVA base. For a 120 MW campus the load power factor is assumed to be 0.98 lagging, so the apparent power is approximately 122 MVA. Two such transformers can be installed in an $N{+}1$ configuration, where each unit can carry the full load during contingency operation.

On the 34.5 kV side, the three-phase line current under full load is

$$I_{34.5} \approx \frac{P_{\text{tot}}}{\sqrt{3}V} = \frac{120 \times 10^6}{\sqrt{3} \times 34.5 \times 10^3} \approx 2010 \text{ A}. \tag{3}$$

This current is beyond the rating of a single medium-voltage feeder, so the campus demand is divided across four parallel feeders, each carrying roughly 500 A at peak. Voltage drop and ampacity calculations are then performed on a per-feeder basis.

B. Frequency Dynamics and Swing Equation

To capture the impact of large power electronic loads on system frequency, we employ the classical swing equation for an equivalent machine that represents the inertia of the interconnection:

$$M\frac{d^2\delta}{dt^2} = P_m - P_e, \tag{4}$$

where δ is the rotor angle, P_m is mechanical input power, P_e is electrical output power, and $M = 2HS_{\text{base}}/\omega_s$ is the inertia coefficient in per-unit seconds. Taking the time derivative of electrical frequency f yields

$$\frac{df}{dt} \approx \frac{\Delta P}{2HS_{\text{base}}} f_0, \tag{5}$$

where $\Delta P = P_m - P_e$ is the power imbalance, H is the inertia constant in seconds on base S_{base}, and f_0 is nominal frequency.

In a conventional system dominated by large synchronous generators, the aggregate inertia constant H_{sys} is on the order of 5–6 s. As rotating machines retire and are replaced by inverter-based resources and loads such as AI servers and their power supplies, the effective H_{sys} decreases. Equation (5) shows that the rate of change of frequency after a disturbance scales inversely with H_{sys}.

To illustrate, consider the sudden disconnection of a 120 MW AI campus from a system with $S_{\text{base}} = 150{,}000$ MW and $H_{\text{sys}} = 5$ s. The immediate power imbalance is $\Delta P =$

+120 MW because generation exceeds load. The initial rate of change of frequency is

$$\frac{df}{dt} \approx \frac{120}{2 \times 5 \times 150{,}000} \times 60 \approx 0.0048 \text{ Hz/s}. \tag{6}$$

If low inertia resources replace part of the synchronous fleet and the effective H_{sys} falls to 3 s, the same event yields

$$\frac{df}{dt} \approx \frac{120}{2 \times 3 \times 150{,}000} \times 60 \approx 0.008 \text{ Hz/s}, \tag{7}$$

a RoCoF increase of roughly 70%. In reality, events in Northern Virginia involving simultaneous tripping of several data centers have produced measurable high-frequency excursions of similar magnitude on the PJM system [3], [6].

Battery energy storage co-located with the campus can be controlled to counteract this effect. If a 100 MW battery injects power for a short interval when the campus trips offline, the net change in load seen by the grid is close to zero, reducing ΔP and hence df/dt. Conversely, during an under-frequency event the battery can temporarily decrease consumption or inject power to provide fast frequency response.

C. Net Load Simulation and Shark Tooth Curve

To study the impact of a flat AI campus load on local net load shape, we construct a simple one-day profile for a PJM Dominion substation. The base case, without the data center, follows a residential and commercial load pattern similar to the well-known duck curve. Normalized net load $P_{\text{base}}(t)$ is low at night, dips around midday due to solar generation, and rises sharply in the late afternoon.

We then superimpose a constant 120 MW block to represent the data center campus:

$$P_{\text{new}}(t) = P_{\text{base}}(t) + P_{\text{campus}}. \tag{8}$$

The result is a profile sometimes called a shark tooth curve: the daytime dip due to solar is partially filled but the evening ramp remains steep, and both minimum and maximum net loads are higher.

Fig. 1 illustrates a representative example of this transformation and also shows the effect of peak shaving from an on-site battery that discharges during the evening peak. A short Python script using NumPy and Matplotlib can be used to generate this visualization and to explore different choices of campus size and battery schedule.

D. Energy Arbitrage Model Using PJM Prices

The final component of the methodology is an economic model for battery arbitrage using historical locational marginal prices from the PJM Dominion zone. Monitoring Analytics, the independent market monitor for PJM, reports that in 2024 the load-weighted average on-peak LMP in PJM was about $37/MWh, while the off-peak average was about $26/MWh [10]. During an extreme heat event on June 29, 2025, real-time LMPs in the Dominion zone reached the administrative price cap of $3700/MWh for several hours, and many of the 50 highest price hours of the year had prices above $500/MWh [11].

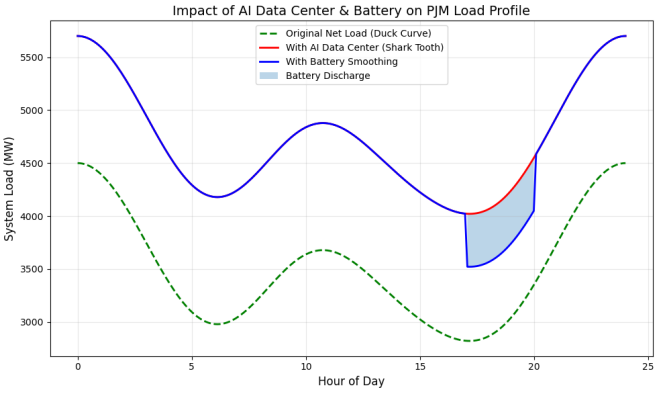

Fig. 1. Stylized PJM Dominion net load profile illustrating the transformation from a "duck curve" (original net load) to a "shark tooth" profile when a large AI data center cluster is added, and the effect of on-site battery peak shaving.

We model a 100 MW battery with a 200 MWh energy capacity connected behind the campus meter. For arbitrage, the operator reserves half of the energy capacity for reliability and uses the remaining 100 MWh for market operations. The battery discharges at 100 MW for one hour during each of the 50 highest LMP hours and charges during low-priced off-peak hours.

Let p_{hi} denote the average price of the top 50 hours and p_{lo} the average off-peak price. For a round-trip efficiency η, the net revenue from one arbitrage cycle is

$$R_{\text{cycle}} = E_{\text{dis}} p_{\text{hi}} - \frac{E_{\text{dis}}}{\eta} p_{\text{lo}}, \tag{9}$$

where $E_{\text{dis}} = 100$ MWh. If we conservatively take $p_{\text{hi}} = \$1500$/MWh (reflecting a mix of price cap hours and lower high-price hours) and $p_{\text{lo}} = \$25$/MWh, with $\eta = 0.9$, then

$$R_{\text{cycle}} \approx 100 \left(1500 - \frac{25}{0.9}\right) \approx \$1.47 \times 10^5. \tag{10}$$

Over 50 cycles this yields approximately $7.4 million in annual gross arbitrage revenue. Additional value can be obtained from frequency regulation payments and local demand charge reductions, which are not explicitly modeled here but are significant in practice [5].

IV. Case Study: Loudoun County AI Campus

We now apply the methodology to a specific case study of an AI-optimized data center campus in Loudoun County, Virginia, connected to the PJM Dominion zone.

A. Campus Load and Grid Interface

The campus delivers 100 MW of information technology power at peak with a PUE of 1.2, resulting in a 120 MW facility demand. The local transmission interface prior to the campus has a peak load of approximately 5000 MW and a minimum nighttime load of 3000 MW, values that are representative of the Dominion zone in recent years [10].

The campus is served by a dedicated 230 kV to 34.5 kV substation with two 150 MVA transformers, each with 10%

per-unit impedance and X/R ratio of 40 [12], [13]. The load is distributed across four 34.5 kV feeders to keep conductor ampacity and voltage drop within ANSI limits. Without this dedicated substation, the existing distribution infrastructure would be unable to carry the additional load, illustrating why utilities often require developers to fund new substations and associated lines.

With the campus online, the regional nighttime minimum rises from 3000 MW to 3120 MW, and the afternoon peak rises from 5000 MW to approximately 5120 MW if the campus operates at full power throughout the day. Over a year, assuming an average facility demand of 115 MW to reflect modest utilization variation, the campus consumes

$$E_{\text{campus}} \approx 115 \text{ MW} \times 8760 \text{ h} \approx 1.0 \text{ TWh}, \tag{11}$$

which is comparable to the electricity consumption of about 90,000 average United States households.

B. Frequency Response Scenario

Consider a contingency where a protection misoperation causes the entire campus to disconnect from the grid and transfer to on-site backup generators. The system experiences a sudden loss of 120 MW of load, resulting in an over-frequency event. Using the swing equation approximation in (5) with $S_{\text{base}} = 150{,}000$ MW and $H_{\text{sys}} = 4.5$ s for the eastern interconnection [6], the initial RoCoF is

$$\frac{df}{dt} \approx \frac{120}{2 \times 4.5 \times 150{,}000} \times 60 \approx 0.0053 \text{ Hz/s}. \tag{12}$$

If multiple campuses around Loudoun County trip simultaneously, the effective ΔP could be well above 1 GW, pushing df/dt beyond 0.05 Hz/s and potentially causing frequency protection relays to operate.

Now suppose the campus includes a 100 MW battery that is normally operating at zero net power but is configured to inject power for a few seconds in response to a rapid frequency rise. If the battery ramps to 100 MW in the opposite direction of the lost load, the net change in power seen by the grid is near zero and the RoCoF is reduced accordingly. This simple calculation shows why grid operators and standards bodies are increasingly interested in using inverter-based loads and storage to provide fast frequency response [7].

C. Duck Curve to Shark Tooth and Battery Smoothing

We construct a stylized daily net load profile for a Dominion zone substation that serves residential and commercial customers and a mix of solar generation. Without the data center, the net load exhibits a midday dip and a steep evening ramp, the classical duck curve. Adding the 120 MW campus shifts the entire curve upward by that amount and makes the evening ramp even steeper, creating a shark tooth profile.

Using the Python script described earlier, we generate two curves: $P_{\text{base}}(t)$ for the original net load and $P_{\text{new}}(t)$ for the case with the campus added. We then simulate battery operation where the on-site storage charges during late night hours when LMPs are low and discharges during the top 50 price hours of the year. When the battery discharges at 100 MW for one hour during the evening peak, the shark tooth is partially flattened and the ramp rate is reduced. This visual evidence connects directly to the concept of using demand-side resources to mitigate renewable-driven ramp challenges.

D. Battery Arbitrage and Financial Benefits

Using the PJM Dominion LMP data summarized in the methodology, we estimate the annual energy arbitrage value of the battery. The average off-peak price is taken as \$25/MWh and the average price of the 50 highest hours is approximated as \$1500/MWh, based on the mix of price-capped hours near \$3700/MWh and other high-price hours above \$500/MWh [10], [11].

For each of the 50 cycles, the battery discharges 100 MWh during a high-price hour and charges during a low-price hour. With an efficiency of 90%, the annual arbitrage revenue is

$$R_{\text{year}} = 50 \times 100 \times \left(1500 - \frac{25}{0.9}\right) \approx 7.4 \times 10^6 \text{ USD}. \tag{13}$$

From the data center operator perspective, this revenue offsets part of the cost of the storage system, while from the system perspective the battery is providing peak shaving and potentially fast frequency response. If the battery also participates in PJM frequency regulation markets, additional annual revenue on the order of hundreds of thousands of dollars per 100 MW is plausible based on historical regulation clearing prices [5].

V. Mitigation Strategies

The case study suggests several mitigation strategies that can be generalized to broader data center integration.

A. On-Site Storage and Grid-Interactive UPS

On-site battery systems and advanced UPS architectures can provide ride-through capability, peak shaving, and ancillary services. By maintaining the campus load during short disturbances and by modulating net demand in response to system conditions, storage reduces both capacity and stability impacts. The same hardware that is needed for reliability can be used to support the grid, provided that control and market frameworks are in place [5]–[7].

B. Flexible Workloads and Demand Response

Although many AI workloads are intensive, some fraction is flexible in time or location. Operators can design scheduling frameworks that slow or defer noncritical jobs when the grid is stressed and accelerate them when conditions are favorable. Paired with real-time pricing or explicit demand response programs, this flexibility can significantly reduce peak contributions. Recent work shows that coordinated demand response across multiple data centers and retailers can flatten system load and reduce capacity requirements [8], [9].

C. Energy Efficiency and Advanced Cooling

Continued improvement in energy efficiency, especially in cooling, remains one of the simplest ways to reduce impact. Every reduction in PUE translates directly into lower facility demand for a given level of computation. Advanced liquid cooling, improved airflow management, and waste heat utilization reduce overhead consumption and lower both base load and peak load [1], [4].

D. Colocation with Clean Generation and Microgrids

Colocating data centers with new renewable generation, or developing campus-level microgrids that include gas turbines or future small modular reactors, can reduce the net burden on the public grid. In some designs, on-site generators run only during grid peak or contingency events, while in others they provide continuous power with the grid acting as backup. Such configurations raise policy questions about emissions and cost allocation, but they are likely to play a role in high growth regions [1], [6].

VI. Policy and Planning Implications

Technical solutions alone are not sufficient. Policy and planning frameworks must align incentives so that data centers contribute to, rather than undermine, reliability and decarbonization goals.

First, cost allocation mechanisms should ensure that entities driving load growth pay an appropriate share of the required infrastructure upgrades. This may include direct contributions for new substations and lines, higher demand charges, or special tariffs for very large customers.

Second, interconnection rules and reliability standards may need to be updated to include ride-through requirements and expectations for large loads. Just as grid codes for generators now require low-voltage ride-through, analogous requirements for data centers could reduce the risk of mass disconnection events.

Third, market designs should reward flexible demand and grid-supportive behavior. Examples include capacity credits for loads that can curtail, ancillary service markets that treat storage and controllable loads on a comparable basis with generators, and tariffs that encourage data centers to shift load to times of surplus renewable generation.

Finally, planning processes should integrate data center development timelines and locations into long-range resource and transmission planning. Transparent communication between developers, utilities, and regulators is critical to avoid both overbuilding and underbuilding of infrastructure.

VII. Conclusion

AI-optimized data centers represent a new class of large, steady, and rapidly growing loads on the United States electric grid. Their concentration in regions such as Loudoun County in Northern Virginia challenges traditional assumptions about load growth, power flow patterns, resource adequacy, and ancillary service needs. Through a more detailed analytical framework and a case study of a 100 MW campus, this paper has quantified how such facilities affect substation design, transmission loading, peak demand, and frequency stability, and has evaluated the technical and economic value of co-located battery storage using actual PJM price data.

The analysis shows that without mitigation, large clusters of data centers can quickly erode planning reserve margins and stress local infrastructure. At the same time, the same facilities possess assets and flexibility that can help maintain reliability, including on-site storage, uninterruptible power supplies, and schedulable workloads. By deploying grid-interactive storage, improving energy efficiency, participating in demand response, and in some cases colocating with new generation, data centers can become active partners in grid stability rather than passive stressors.

Policy and planning frameworks that align costs and incentives are essential. When combined with sound engineering design, they can allow continued growth of AI and digital infrastructure without sacrificing reliability or decarbonization progress. For students of power systems, AI data centers provide a timely and concrete application of core concepts such as power flow, the swing equation, resource adequacy, and ancillary services, and they highlight the importance of interdisciplinary thinking at the interface of computation and energy.

References

[1] Electric Power Research Institute, "Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption," EPRI, Palo Alto, CA, USA, 2024.

[2] J. Egan, "EPRI Report Sees Dramatic U.S. Electric Demand Growth from Data Centers," *Industrial Info Resources*, Jun. 2024.

[3] Z. Yang *et al.*, "Electricity Demand and Grid Impacts of AI Data Centers: Challenges and Prospects," *arXiv* preprint arXiv:2509.07218, 2025.

[4] Meta Platforms Inc., "2024 Sustainability Report," 2024.

[5] EnerSys Corp., "How Data Centers Support Grid Stability and Peak Shaving," technical blog, Nov. 2025.

[6] U.S. Department of Energy, "Resource Adequacy Report," Washington, DC, USA, Jul. 2025.

[7] S. Subedi, M. Blonsky, Y. Son, and B. Mather, "Cost Benefit Analysis of Grid Supportive Loads for Fast Frequency Response," in *Proc. IEEE PES Grid Edge Technologies Conf.*, 2023, pp. 1–5.

[8] P. Zhang, P. Yang, Z. Zhao, C. S. Lai, L. L. Lai, and Y. Cao, "A Framework for Several Electricity Retailers Cooperatively Implement Demand Response to Distributed Data Center," *IEEE Trans. Smart Grid*, vol. 14, no. 1, pp. 277–289, 2023.

[9] T. Yang, Y. Hou, S. Cai, J. Yu, and H. Peng, "Multi Data Center Tie Line Power Smoothing Method Based on Demand Response," *IEEE Trans. Cloud Comput.*, vol. 12, no. 4, pp. 983–995, 2024.

[10] Monitoring Analytics LLC, "2024 State of the Market Report for PJM," Independent Market Monitor for PJM, 2025.

[11] Monitoring Analytics LLC, "High Real Time Prices in PJM on June 29, 2025," Independent Market Monitor for PJM, press release, 2025.

[12] Xcel Energy, "High Voltage Substation Transformer Specifications," technical data sheet, accessed 2025.

[13] IEEE Standards Association, "IEEE Standard Requirements for Liquid-Immersed Power Transformers," IEEE Std C57.12.10-2017, 2017.