From Firm Load to Flexible Resource: Understanding Data Center Flexibility and Its Costs

Elan Trager
Fellow | Evolved Energy Research

EVOLVED ENERGY RESEARCH

Photo by Geoffrey Moffett on Unsplash

Executive Summary

Against a backdrop of immense projected U.S. data center load growth and a grid that has experienced low, predictable load growth for decades [1], many strategies have been proposed to enable the development of new data centers. One solution gaining traction is data center flexibility. By exchanging faster interconnection for the ability to call on data centers for peak shaving, flexibility can allow for greater usage of existing generation and transmission capacity. This can unlock existing grid headroom, defer costly infrastructure upgrades, and accelerate time-to-power for new facilities.

This whitepaper has two goals. First, it reviews the current state of data center flexibility efforts, including public initiatives, tariff designs, and operator-led projects. Second, it quantifies the cost implications of on-site generation and storage for data center operators using a behind-the-meter (BTM) optimization model within the CAISO balancing authority. The modeling assesses generation and storage technology choices under varying flexibility requirements, regulatory environments, and data center penetration levels. The results are not intended as site-specific forecasts or procurement guidance. Instead, they highlight the directional cost impacts of required flexibility, given CAISO operating conditions and the rate schedule applied in the modeling.

This paper is guided by three central questions:

- How do operator costs shift as flexibility requirements increase?
- How does the optimal mix of BTM technologies change as more data centers participate?
- How do these costs change under different simulated regulatory environments?

Key insights from this work:

1. **Data centers are overwhelming the grid**. Interconnection queues are clogged with both new generation and new large loads. Data centers are projected to add tens of gigawatts of demand, but grid expansion is years behind due to long lead times for turbines, transformers, and transmission. The result is a bottleneck: load growth is outpacing the grid's ability to deliver new capacity.

2. **Flexibility is limited but meaningful**. Although the amount of low-cost flexibility available to data centers is finite, it can materially reduce time-to-power for early projects while utility-scale capacity is developed.

3. **Flexibility gets progressively more expensive**. Each additional unit of required flexibility and flexible nameplate data center capacity raises costs for participating data center operators. However, technologies that can engage in arbitrage can help mitigate this impact.

4. **Low penetration favors low fixed-cost options**. When flexible load participation is limited, cost-effective strategies are those built on low capital cost technologies, even those with high variable costs.

5. **Permissive regulation broadens the toolkit**. In environments with fewer regulatory barriers, thermal generation and fuel cells play major roles in providing cost-effective flexibility, while BTM storage emerges as a strong secondary option for high-energy, low-power shifting.

6. **Planning frameworks need to change**. As immense data center load growth approaches, the dynamics of data center demand and flexibility require novel analysis approaches to properly account for the system value they might contribute.

Table of Contents

Executive Summary 1

Glossary 3

Introduction 5

1. Data Center Characteristics and Flexibility 6

1.1 Data Center Characteristics 6

1.2 Hyperscaler Priorities 7

1.3 Types of Flexibility 8

1.4 Delivering Workload Flexibility 9

1.5 Data Center Uptime 9

1.6 Flexibility Dynamics 10

2. Modeling the Operator Costs of Flexibility 12

2.1 Model Setup and Approach 12

2.2 Results 13

Conclusion 17

Appendix: Model Overview, Inputs, and Assumptions 18

A.1 Model Overview 18

A.2 Model Inputs 18

A.3 Assumptions 20

A.4 Limitations 21

References 23

Glossary

Air District: A local California air-quality regulator that issues permits and operating limits for on-site generation.

Balancing Authority (BA): The entity responsible for maintaining real-time balance of supply and demand within a defined grid area.

Behind-the-Meter (BTM): Generation or storage located on the customer side of the utility meter that can offset grid imports or provide flexibility.

Battery Energy Storage System (BESS): Electrochemical storage that charges from the grid or on-site resources and discharges to serve load.

Bridge Power: Temporary on-site generation used to meet load while awaiting grid interconnection or upgrades.

CAISO: California Independent System Operator; the BA operating most of California's transmission grid and wholesale markets. Balancing authority used for this modeling exercise.

Capacity Factor (CF): Ratio of actual energy output over a period to the energy output if operating at nameplate power for the entire period.

Capacity Reservation Charge (CRC): Tariff fee paid to reserve utility capacity for fallback grid service when a customer uses on-site generation.

Capital Recovery Factor (CRF): Factor used to convert a capital cost into an equivalent uniform annual (or period) cost given a discount rate and asset life.

Curtailment: Reduction of load or generation relative to what would otherwise occur; in this paper, load curtailment by data centers during grid stress.

Data Center Flexibility (this paper's usage): Ability of data centers to reduce grid-facing demand at times of system stress relative to a firm-load counterfactual, thereby limiting required supply-side capacity buildout.

Demand Charge: Tariff component based on the customer's maximum measured demand (kW) during a billing window.

Demand Response (DR): Programs that compensate or require customers to reduce or shift load in response to grid conditions or prices.

Emergency Generator: Standby generator permitted primarily for testing/maintenance and true emergencies; non-emergency use is typically restricted by Air District rules.

Emergency Generator Runtime (modeled): Annual number of hours during which emergency generators may run to support grid events in sensitivity cases.

Energy Emergency Alert (EEA 1/2/3): NERC defined notifications signaling increasing system stress; higher levels indicate proximity to load shedding.

Firm Load: Demand that must be served under all normal conditions; opposite of flexible/interruptible load.

Fuel Cell: Electrochemical generator producing electricity (and heat) with low NOx emissions.

Headroom (System): Unused generating or transmission capability available above current load; inferred in this paper using load duration curves.

Hyperscaler: Large cloud/AI operator with multi-MW to GW-scale facilities.

Interconnection Queue: The list and study process for connecting new generation or large loads to the grid.

Latency-Sensitive Workloads: IT tasks requiring rapid response times (e.g., real-time services); generally less flexible temporally/spatially.

Load Duration Curve (LDC): Demand at all hours of a given timeframe, sorted in descending order by portion of maximum load to visualize demand variability and potential headroom.

Long-Duration Energy Storage (LDES): Storage capable of delivering for >10 hours; typically lower round-trip efficiency than short-duration BESS.

Nameplate Capacity (Data Center): Maximum continuous electrical demand the facility is designed/permitted to draw.

Net Load: System load minus variable renewable output, in this paper the CAISO load profile subtracted from solar and wind production.

Off-Grid Data Center: Data center supplied entirely by on-site resources without a grid interconnection.

Prime Generator: Non-emergency generator intended for regular or continuous operation at varying loads.

PUE (Power Usage Effectiveness): Data center efficiency metric: total facility power divided by IT equipment power.

Reciprocating Engine (NG/Diesel): Combustion generator technology with fast start and high power density; emissions-constrained in many areas.

Round-Trip Efficiency (RTE): Ratio of energy discharged from storage to energy charged, indicating the losses associated with the storage technology.

Service-Level Agreement (SLA): Contractual agreement between provider and customer that defines level of service and performance requirements along with penalties for failing to meet them.

Short-Notice (Emergency) Flexibility: Rapid, unplanned shedding supplied by existing BTM resources during grid emergencies (e.g., EEA conditions).

Silicon Valley Power (SVP) CB-6: Large-load retail tariff used as a representative schedule in this paper (includes TOU, demand charge, and CRC features).

Spatial Shifting: Moving workloads between sites/regions to relieve local grid stress when latency constraints allow and capacity exists.

Tariff (Retail Rate Schedule): Utility pricing and terms governing customer billing for energy, demand, and capacity reservation.

Throughput-Driven Workloads: Compute tasks that depend on total work completed. Certain throughput-driven workloads may enable data center flexibility without the need for on-site generation.

Time-of-Use (TOU) Rates: Energy prices that vary by time block, enabling valuable arbitrage by storage or load shifting.

Uptime: Percent of time a data center is operational. Often measured by '9s' of uptime and specified in SLAs.

Uptime Tiers (Uptime Institute): Tier I–IV redundancy classifications (N, N+1, 2N, 2N+) defining facility reliability levels.

Uninterruptible Power Supply (UPS): Power-conditioning/ride-through system that maintains IT load during short disturbances.

Workload Flexibility (IT): Ability to defer, queue, or migrate compute tasks without violating SLAs; a complement to BTM resources in providing flexibility.

Introduction

U.S. data center demand is on track to add anywhere from tens to hundreds of gigawatts (GW) by 2030 [2]. In the 2024 U.S. Annual Decarbonization Perspective, Evolved Energy Research (EER) forecasts data centers will account for 10.6% to 13.6% of U.S. energy consumption by 2030 [3]. The rapid expansion of AI inference and training facilities could make even the higher-end forecasts plausible, requiring substantial new transmission and generation capacity. This is particularly true if data centers are treated as firm load.

Yet, this demand boom collides with barriers to grid expansion. Hyperscalers, the high market cap tech companies at the helm of AI data center expansion, face a grid conditioned by decades of slow demand growth [4]. Now, utilities must balance increases in demand from electrification, long supply-side interconnection queues, and a flood of interconnection requests from data centers [5]. As a result, interconnection wait times have ballooned, study resources are strained, and utility load forecasting is a far more difficult and imprecise exercise. Uncertainty in eventual data center construction further complicates forecasting. An EPRI survey found that some utilities derate requested data center capacity, others include all of it, and some remove it entirely from forecasting [6]. This illustrates the difficulty for utilities in adequately sizing capacity investments and for data center operators that prioritize fast time-to-power.

One solution to develop data centers under these constraints is flexibility. Here, flexibility is defined as the ability of data centers to reduce their demand-side load at times of grid stress relative to a firm-load counterfactual. This capability, when aggregated across facilities, can limit the need for new supply-side capacity buildout. While the promise of flexibility is enticing, efforts to determine how flexible data centers can be are ongoing, with numerous public initiatives, innovative tariff designs, and private projects actively exploring it.

To understand the potential for large load flexibility, Norris et al. conducted a study [7] that estimates the U.S. grid could integrate 126 GW of new load with only 1% annual curtailment, demonstrating the significant potential for large load flexibility. While this whitepaper supports a growing discussion indicating the value of flexible large loads, additional study is needed to determine the extent to which data centers can offer flexibility and at what cost.

This whitepaper includes an introduction to data center flexibility, notable public flexibility studies, and an exploration of the costs of flexibility using a BTM optimization model. The scenarios apply a California Independent System Operator (CAISO) load shape to provide a stylized demonstration of how different regulatory environments and technology mixes influence operator costs. While not a site-specific forecast, the results offer directional insight into the tools data centers can use to provide demand-side flexibility, the tradeoffs of different on-site technology mixes, and the implications of flexible tariff designs for operator costs.

1. Data Center Characteristics and Flexibility

1.1 Data Center Characteristics

Data center operators have a broad range of needs depending on their ownership model, data center workload profiles, sizes, and desired reliability. Each combination of characteristics determines unique spatial and temporal constraints that operators must meet when siting new data centers, in addition to the cost of flexibility [8]. These different ownership models confer varying levels of control over infrastructure, utility arrangements, and operational decision-making—factors that directly influence how, when, and to what extent a facility can adjust its load in response to grid conditions.

EPRI's 2025 whitepaper, Grid Flexibility Needs and Data Center Characteristics [8], provides a clear framework for organizing data center characteristics. They sort data center characteristics by ownership model, workload, size, and reliability, and the combination of these characteristics determine a data center's potential for flexibility. These characteristics are summarized below.

Size

Data centers can be small, medium, or large (hyperscale) entailing <5 MW, 5–20 MW, and >20 MW power requirements. Over time, data centers have grown larger, with a Wood Mackenzie report finding that the average size of proposed data centers grew from 150MW in 2023 to 300MW in 2024 [9]. While many of the documented projects in queues may not get built, this is a striking indication of the rate at which hyperscale data centers are becoming the industry standard.

Ownership Models

Data centers can be owner-operated, co-located, hosted, or cloud. For owner-operated data centers, the data center user owns all physical infrastructure. For co-location, the data center user owns the racks and is responsible for all maintenance for the IT equipment they operate. The data center shell itself is built and managed by the data center owner. For hosted data centers, the service provider owns the physical infrastructure and leases servers to customers, who still carry the responsibility for maintenance and workload orchestration. Finally, cloud ownership models assign all responsibility and ownership of physical infrastructure to the service provider. End users access services through service-level agreements (SLAs) that tie contractual obligations to the services provided.

Workloads

Workload type may most directly determine flexibility. There are numerous ways to categorize workloads, but generally they can be characterized by the latency and throughput they require (note a simplified framing here as compared to [8]). These are not mutually exclusive, and real world workloads are constrained by a combination of the two. To see how these dimensions shape flexibility, it's useful to consider each in turn, starting with latency.

Latency is the delay between when a request is made and a system responds, while throughput is the total processing capacity. Latency is determined by factors like the physical distance between end users and the data center, the network infrastructure used, and the methods data centers use to efficiently manage workloads. Applications that demand extremely low latency must be located close to users or major exchange points. Examples include real-time systems, streaming platforms, and high-frequency trading.

Throughput-driven workloads require very large amounts of computing capacity and continuous data movement but are generally less sensitive to latency. What matters is not the instantaneous speed of response, but the total volume of work completed over a given time, making them more tolerant to being shifted temporally or spatially. Examples include batch processes such as artificial intelligence (AI) training runs, large-scale scientific modeling, video rendering, and certain machine learning (ML) training work-

loads. Because these jobs can often be paused, queued, or scheduled flexibly, throughput-driven workloads represent the highest potential for demand shifting and grid integration compared to latency-sensitive applications.

Reliability

There are a number of standards defining reliability, with the Uptime Institute's reliability tiers being one of the most common [10]. They define 4 tiers of reliability:

- **Tier 1 (N)**: Basic redundancy, with backup generation, an Uninterruptible Power Supply (UPS), space for IT equipment, and dedicated cooling equipment. This provides enough reliability for day to day disruptions, but the facility still has to shut down for maintenance.
- **Tier 2 (N + 1)**: Everything that is included in Tier 1, with additional layers of power and cooling backups to provide additional redundancy. Tier 2 facilities can undergo a degree of equipment failure or removal without undergoing shutdown.
- **Tier 3 (2N)**: Facility can undergo maintenance or removal of any asset in service while maintaining operation. Tier 3 facilities can operate for up to a couple of days without grid power.
- **Tier 4 (2N+)**: Additional redundant layers to all systems. Facilities are "fault-tolerant, fully redundant, and can guarantee a downtime of only 26 minutes annually" [11].

Each combination of data center characteristics informs the amount of flexibility a data center can deliver. As ownership over facilities and workloads increases, so too does the operational autonomy that enables flexibility. As workloads are chosen that have lower latency requirements and less strict required response times, there is more opportunity to decouple temporal and spatial requirements to drive flexibility. A workload and ownership structure that exemplifies this is the hyperscale training of in-house AI models. If these data centers are entirely owned by hyperscalers and carry out primarily training workloads (high throughput), they may have more ability to be sited in areas of greater grid headroom. They may also tolerate more downtime than traditional data centers. This will be discussed further in Section 1.5.

1.2 Hyperscaler Priorities

This paper is not meant to thoroughly explore the current barriers to U.S. grid expansion, but it is worth noting some of the most significant challenges, as they inform the interplay between data center operators and current grid conditions.

- **Transformer lead times**: Wood Mackenzie reported in 2024 that "transformer lead times have been increasing for the last 2 years – from around 50 weeks in 2021, to 120 weeks on average in 2024. Large transformers, both substation power and generator step-up transformers, have lead times ranging from 80 to 210 weeks" [12].
- **Turbine supply shortages**: As the Rocky Mountain Institute summarizes, "Mitsubishi states that turbines ordered today will not be delivered until 2028–2030. Siemens reports a record backlog of €131 billion (U.S. $148 billion). And GE Vernova has announced new turbines will not be available until late 2028 at the earliest" [13].
- **One Big Beautiful Bill impacts**: In our analysis for the REPEAT project, Evolved Energy Research projects that the OBBB will significantly increase future costs of energy and reduce deployment of renewable energy resources [14].
- **Slow interconnection timelines**: Total active capacity in U.S. interconnection queues is nearly 6TW as of 2023 [5]. Additionally, "the median duration from interconnection request (IR) to commercial operations date (COD) continues to rise, approaching 5 years for projects completed in 2022-2023".
- **Transmission permitting length**: A review of 30 transmission projects found an average completion time of 10 years [15].

These obstacles for integrating new loads into existing infrastructure make project expenses and timelines less predictable and diminish the likelihood of introducing larger loads without increasing rates for other customers.

As grid expansion becomes more difficult, time-to-power has become a defining priority for hyperscalers and other developers. A 2025 survey by Bloom Energy found that "[data center] leaders are now ready to invest 50% more than seven months ago if that means they can access power faster for upcoming data center projects" [16]. As demand for AI infrastructure grows, the ability to secure electricity quickly often outweighs other considerations such as site optimization or long-term operating costs, given the high opportunity value of immediate compute capacity.

Time-to-power is not the only priority influencing hyperscaler decision-making. Other energy-related considerations, such as total cost of energy over the facility's life, access to low-carbon or renewable power, regulatory certainty, site expansion potential, and alignment with corporate sustainability commitments, also play important roles. Beyond energy, factors like proximity to talent pools, tax incentives, availability of fiber and network latency requirements, and exposure to extreme weather events can significantly shape siting and design choices. While these priorities remain important in the broader decision process, the growing focus on flexibility stems most directly from time-to-power requirements, making time-to-power the main lens through which this paper examines data center flexibility.

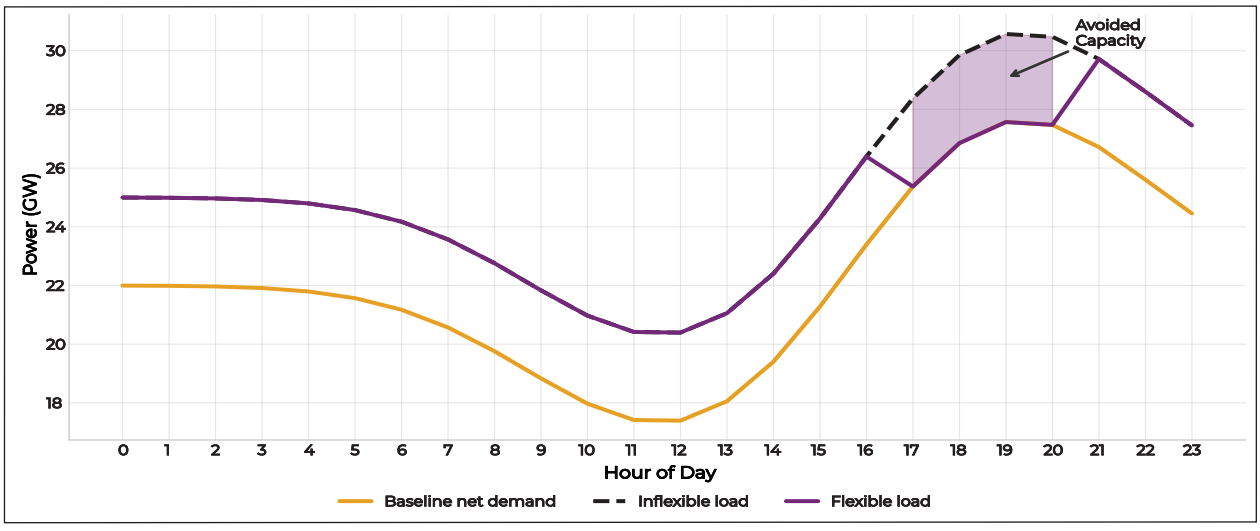
1.3 Types of Flexibility

Flexibility in data centers can take different forms depending on how much advance notice operators receive to adjust load. While both reactive (or short-notice) and planned flexibility can help balance the grid, only the latter provides the ability to plan for cost-minimizing solutions. Broadly, these two categories can be described as follows:

1. **Short-notice flexibility**: when a grid event requires a data center to operate without grid power without enough notice to adequately plan additional capacity past backup generators. Emergency flexibility is triggered by grid emergencies such as CAISO Energy Emergency Alerts (EEA) 1 or 2 [17]. In these situations, the grid operator signals that reserves are low and voluntary or mandatory curtailments are needed to avoid outages. Data centers with on-site backup generation or storage can shed load from the grid and operate independently, if permitted to do so by the local Air Quality District and properly incentivized.

2. **Planned flexibility**: when a data center has an adequate planning horizon to site the needed capacity for planned periods without grid energy use. Planned flexibility can help avoid utility-scale capacity buildout and bring on flexible loads as the capacity is built. Examples include reducing peak-hour demand to avoid triggering new capacity upgrades (**Figure 1**), shifting workloads to off-peak periods, or operating partially from on-site resources to lower interconnection requirements. The modeling presented later reflects this type of flexibility, assuming foresight of peak periods and demand response calls. This foresight, created through planned flexibility, enables least cost deployment of flexibility tools.

Figure 1: Illustrative curtailment of flexible load



1.4 Delivering Workload Flexibility

Flexibility can be provided in three main ways:

1. **Spatial shifting**: shifting workloads across locations. This involves moving IT workloads from one data center to another when the local grid is under stress. This approach is most viable when latency requirements are not stringent, when sufficient unused capacity exists at other facilities under the operator's control or in contractual partnership, and when the destination site has the necessary hardware and software to execute the workload shift. Implementing spatial shifting at scale requires sophisticated workload orchestration software capable of managing complex service-level agreement (SLA) requirements across sites.
    i. Example: Google claims to have shifted workloads in Europe in 2023 to other locations during high grid stress [18]. This was followed by a progress update in August of 2025, with Google noting an improved ability to shift machine learning (ML) workloads for demand response [19].

2. **Temporal shifting**: adjusting consumption within a given site over time. This refers to altering a site's energy consumption profile within a day—shifting workloads away from peak price or peak demand periods. Temporal shifting is generally not cost effective for most IT loads unless SLAs explicitly allow it and data center operators are well compensated, but it may be feasible for in-house data centers or cloud providers that operate under flexible contracts. Non-IT loads, such as cooling systems, present limited but measurable opportunities for temporal shifting through strategies like pre-cooling or short-term load modulation, though these opportunities are modest compared to shifting IT workloads.
    i. Examples: Google's recent workload shifting likely involved temporal adjustments , "shifting non-urgent compute tasks — like processing a YouTube video — during specific periods when the grid is strained" [19] . EPRI's DCFlex initiative with Emerald Conductor demonstrated a "25% reduction in cluster power usage for three hours during peak grid events while maintaining AI quality of service (QoS) guarantees" [18], [20].

3. **BTM generation and storage**: deploying on-site generation or storage to supply part or all of the facility's demand during periods of high grid stress, or while waiting for required upgrades. Examples include battery energy storage systems, emergency or prime generators, and, in some cases, on-site renewable generation. BTM resources can be dispatched independently of grid conditions, making them a reliable source of flexibility where permitted by regulation and air-quality standards.
    i. Examples: NREL research in 2025 explored battery energy storage systems and accompanying orchestration software for demand response in virtual simulations [20]. xAI ran an estimated 35 gas turbines as one of its data centers waited for interconnection in Memphis [21], [22]. This is known as "bridge power", used as xAI waited for necessary substation upgrades to receive full contracted service [21]. A Bloom Energy survey of data center operators found that approximately 30% of data centers will rely on on-site generation as a primary supplemental source by 2030, reflecting anticipated constraints on grid availability [16].

1.5 Data Center Uptime

The modeling in this paper assumes that data center operators require high uptime. "Uptime" is the percentage of time a system is fully operational, often expressed as a "number of nines" (e.g., "five nines" for 99.999% uptime, meaning the data center is expected to be able to run for 99.999% of the time in a year or all except about 5 minutes). Many SLAs today promise 99.9%–99.999% uptime, with four to five nines considered the industry standard.

Despite industry standards, not all workloads need this level of reliability. Certain ML and AI workloads, as demonstrated by Google's shifting of machine learning workloads, can tolerate greater downtime. However, this particular flexibility is likely difficult to scale beyond hyperscaler-owned facilities running in-house workloads. One promising application is specialized AI training data centers at the hyperscale level. If these facilities can operate without four or five nines and have temporally flexible downtime, they could unlock new possibilities such as off-grid data centers.

One analysis estimates over 1.2TW of data center potential for off-grid data centers with solar micro-

grids in the U.S. Southwest alone [2]. The authors calculated prices of $109/MWh for an off-grid data center with 90% of lifetime energy demand met by solar, only a $23 premium to using large-scale off-grid gas turbines. While the prospect of off-grid data centers is promising, the most viable candidates are likely AI training facilities with downtime tolerance. For these facilities, an off-grid solution could be possible, especially if it's the fastest path to power and the lost revenue from additional downtime is minimal.

For loads requiring high uptime, a grid connection is incredibly valuable. While the benefits of large, distributed energy systems are well known, it is worth noting why data centers in particular benefit from securing grid capacity, as it is the core reason for grid-connected flexibility.

Grid Redundancy

Grid-connected data centers benefit from more than just electricity supply. They draw reliability from modern power-system features such as reserve margins, grid balancing to smooth disturbances, emergency demand response, and other services that minimize outage frequency and duration. Replicating this reliability off-grid demands far greater redundancy, especially when generators must provide continuous power rather than serve as backup.

In the off-grid analysis referenced above, prime generators were sized at 125% of the data center's capacity, supplying any load not met by Battery Energy Storage Systems (BESS) and solar. This may well work for the specialized AI training data centers that are the subject of the analysis, as they may be able to tolerate limited downtime, but for most of the remaining data centers that demand high reliability, achieving five nines would likely be prohibitively expensive. With the grid connection, significant layers of redundancy are already provided front-of-meter.

Siting

Data center siting is constrained by features such as proximity to population centers, access to infrastructure, and potential for expansion, particularly in terms of generation capacity and transmission. Often, data center infrastructure is located in or adjacent to land-constrained urban centers without sufficient or cost-effective space for localized generation like solar. In places like Santa Clara, California, which hosts a substantial concentration of California's data centers, large-scale on-site renewable energy development is impractical beyond, perhaps, rooftop solar. These geographic limitations make a grid connection essential for maintaining high uptime and ensuring that future demand growth at a given site can be met.

For data centers requiring high uptime and low time-to-power, grid-connected flexibility offers a means to accommodate a limited but meaningful share of new load. Once time-to-power is considered, the next key question is cost: What are the cost implications for data centers under different flexibility requirements and penetration levels within a balancing authority? This is the subject of the modeling component of this paper.

1.6 Flexibility Dynamics

The Cost of Flexibility

Modeling the options data centers have to provide flexibility can help to quantify the economic impact of flexible operations (e.g., load shifting or BTM generation) for both data center operators and utilities. For data center operators, it measures the costs, savings, and time-to-power trade-offs of different flexibility tools and tariff structures. For utilities and planners, it shows how flexibility can defer costly infrastructure upgrades, support lower energy prices for ratepayers, and enable the development of new data centers.

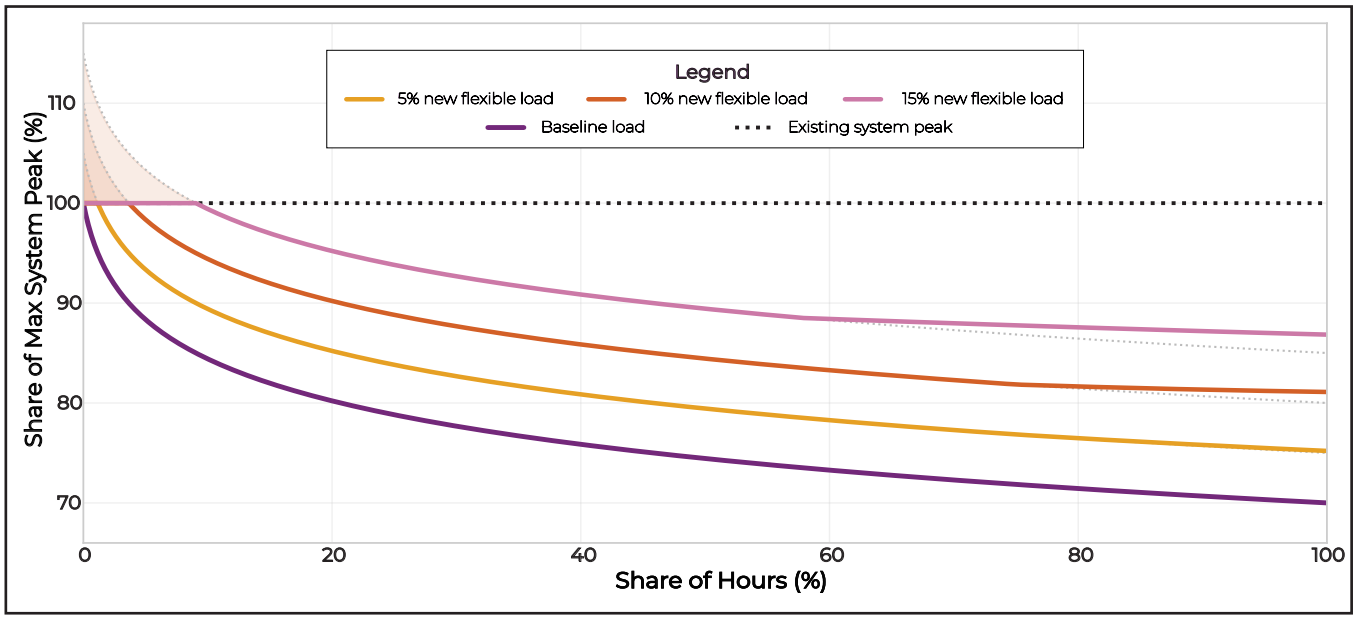
Understanding the cost impacts of flexibility is essential for designing interruptible tariffs and demand response programs that both incentivize new loads and preserve grid reliability. In doing so, it reframes data centers from passive consumers into active grid resources.

The Relationship Between Flexibility and System Headroom

Load duration curves (LDCs) illustrate the available headroom within a system by ranking all hours by portion of system peak and stacking them into a curve. The resulting curve shows the share of hours where load was greater than or equal to the intersecting share of peak capacity. As noted in Norris et al., "a steep LDC suggests high demand variability, with peaks significantly exceeding typical loads, while a flatter LDC indicates more consistent usage" [7]. U.S. load duration curves suggest significant headroom exists, since utilization factors are low enough to accommodate new flexible loads without driving proportional new capacity additions. At the same time, not all apparent headroom is practically usable. A portion of unused capacity reflects generators that are uneconomic to operate continuously, so achieving 100% utilization of system capacity is neither realistic nor desirable. For this reason, the following analysis treats the LDC not as a forecast of perfectly usable headroom, but as a framework to explore how costs evolve as flexible loads progressively saturate available capacity.

Within each balancing authority, there are different degrees of headroom for flexible loads, and a central concept to this paper is understanding that initial flexible loads on grids with steep load duration curves have to shift their load a small amount in order to significantly reduce utility-scale upgrades to transmission and generation capacity. At the same time, flexibility is not necessarily a long term strategy. As penetration of flexible loads increases, the marginal cost of required flexibility increases correspondingly, as data centers have to build generation to cover larger energy generation requirements (the shaded portion of **figure 2**). Increased amounts of flexibility provided come at greater cost to the flexible loads participating. This is also true for battery energy storage systems (BESS). As seen in **figure 2**, with each additional flexible load, the system utilization increases, but so too does the amount of energy that flexible loads must either shift or generate BTM. With each 5% increase in flexible loads, the shaded portion that represents load that they bear responsibility curtail (by shifting, reducing, or generating BTM) increases non-linearly.

Figure 2: Illustrative load duration curve



2. Modeling the Operator Costs of Flexibility

2.1 Model Setup and Approach

This modeling is designed to explore cost implications of BTM storage and generation for data center operators under different flexibility requirements, data center development, and air pollution regulatory constraints within the California ISO balancing authority area. While other dimensions of flexibility are valuable to consider, rigorously quantifying the costs of load shifting, data center load profiles, and the characteristics of shiftable load is difficult given a lack of available data. This, combined with a goal of near-term applicability for the modeling discussed, determined the focus on on-site generation for this work. In doing so, this section explores:

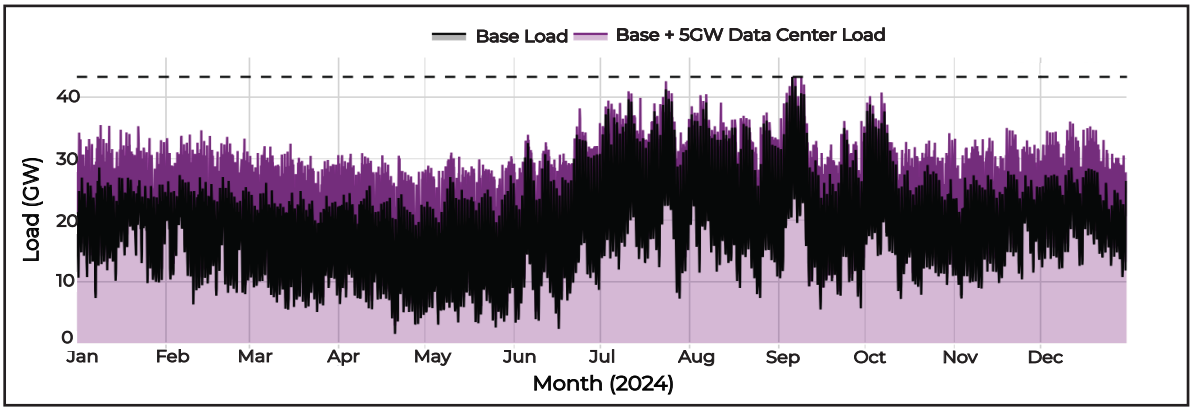
Cost sensitivity: How operator costs scale with increasing flexibility requirements and data center penetration.
Technology choice: Which BTM technologies are most cost effective under strict vs. permissive air quality and permitting regulatory regimes.
Role of backup generators: The effect of limited emergency backup generator use on technology selection and operator costs.

Modeling Approach Overview

- Pulled CAISO net load for 2024 from the OASIS API.
- Selected a rate schedule from a location representing a large portion of California data center load. The rate schedule used for this analysis is the SVP CB-6 Large Load schedule, since Santa Clara, CA contains a high density of data centers. This provides a simplistic representation of the flexibility dynamics we hope to explore, with a set of underlying assumptions detailed in the Appendix.
- Assembled a technology set with near-term deployment potential (<2 years of wait time).
- Developed a BTM technology optimization to produce cost minimizing buildouts of technologies needed to achieve a particular flexibility target within the CAISO load profile. Model incorporated all dispatch across the one year timeframe for this analysis (**figure 3**).
- Conducted sensitivity analysis across emergency generator runtime, flexibility, and nameplate capacity of new data centers.
- Repeated analysis for each scenario.

By modeling various scenarios, from the usage of emergency generator capacity to limiting carbon and NOx intensive technologies, it becomes possible to identify on-site generation mixes that balance grid needs and regulatory requirements with the operational requirements of data centers. Our goal through this work is to examine the cost implications of flexibility and discuss its relation to the present value of compute for data center developers. Note that the scenarios presented use only on-site generation and storage to meet the curtailment requirements demanded by flexibility.

Figure 3: CAISO net load with flexible data center load



Scenarios and Sensitivity

Each scenario is developed to represent a potential regulatory environment and explore flexibility within a static grid that has limited short-term capacity expansion ability. Scenarios capture flexibility within the California grid, however, they don't contain the spatially resolved components that could adequately capture California on-site generation regulatory requirements, making these scenarios directionally valuable but not directly actionable. Cost outputs are informed by a particular industrial rate schedule (SVP CB-6), rather than a system analysis. All scenarios are run across a flexibility sweep that holds nameplate capacity fixed at 10 GW and a nameplate sweep that holds flexibility fixed at 100%. For a complete breakdown of scenario parameters and methods, refer to the Appendix.

Scenarios:

- **Permissive regulatory environment**: All generation technologies enabled, no additional costs past U.S. average permitting costs already included in technology specs. Far from the California regulatory reality.
- **Strict air pollution limits**: All NOx emitting technologies removed; remaining include fuel cells, BESS, and solar. More closely representative of California reality, however fuel cell costs are likely understated due to the difficulty and resources required for permitting.
- **Low emission (solar + storage)**: BESS and solar are the only technologies remaining. Solar is constrained to rooftop feasibility due to land constraints arising from data center proximities to population centers.

Sweep Parameters:

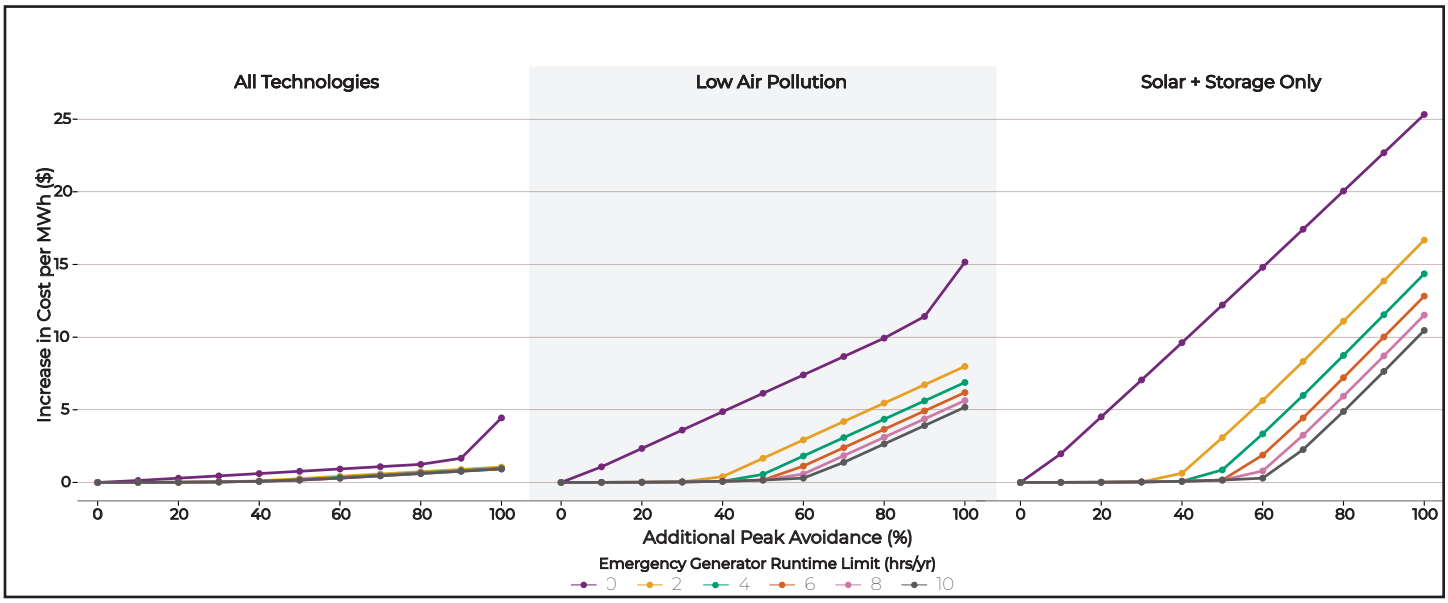
- **Flexibility**: flexibility measures the avoided contribution to peak load. 20% flexibility indicates that 80% of data center nameplate capacity must have properly provisioned transmission and generation capacity ready to serve the data centers in times of peak. 100% flexibility corresponds to no change in peak load, just higher utilization of existing capacity.
- **Nameplate capacity (data center)**: the maximum continuous electrical demand all data center facilities are designed and permitted to draw under normal operation, expressed in kW. It reflects the grid-facing rating (service entrance/transformer/UPS distribution limits), which is included in the model as a flat load.
- **Emergency Generator Runtime**: The maximum number of hours per year that backup (emergency) generators are assumed to be allowed to run for load shedding in addition to their baseline testing and maintenance. In California, local Air Quality Districts generally prohibit non-emergency operation of these units, limiting them to true emergencies (e.g., CAISO Energy Emergency Alerts) and short testing periods. In this modeling, we relax that constraint as a sensitivity to explore the potential role of limited generator dispatch in complementing different technology stacks. This should be viewed as a hypothetical case, not representative of current California permitting practice.

2.2 Results

Flexibility and Generator Sweep

In the 'All Technologies' scenario (**figure 4**), minimal cost increase was observed, even at low emergency generator runtime. This reflects the cost difference in the selected tariff and the fuel costs of technologies like natural gas reciprocating engines. As flexibility requirements increase, cost increases are minimal due to the variable costs per kWh of running the on-site generation being less than the peak energy cost, therefore enabling some cost recovery.

Figure 4: Increase in $/MWh by scenario and emergency generator allowance (10GW new DC Load)



The Low Air Pollution scenario displays a far steeper cost slope. Curtailment requirements are met by fuel cells and some storage in addition to emergency generators when enabled. This scenario displays some of the benefits of limited support from emergency generators. Just 2–10 hours per year of emergency runtime materially flattens the curve. Those scarce hours are targeted at the most expensive high-power periods, cutting required fuel-cell/BESS build. Panels 2 and 3 more closely represent the California reality: costs are driven by capex of compliant firm resources.

Key Findings from Flexibility Scenarios

1. **High power, low energy**: the fundamental challenge with data center flexibility is often not sustained energy delivery but the ability to deliver very high near-instantaneous BTM power over short windows, a task that few technologies are well suited for at the minuscule capacity factor often required.

2. **Backup generator impact**: limited usage of backup generators drastically limits capacity buildout needed for the 10 GW runs, as backup generators can produce high power for a limited amount of time, a strong match to take advantage of the steepest part of the CAISO LDC. Similarly as nameplate capacity decreases, the portion of flexibility that limited emergency generator hours enable increases, as high power but low energy is required for the early LDC slice.

3. **Cost linearity**: after emergency generator headroom, cost increases are relatively linear. This is a result of using a single data center nameplate capacity for this run. A relatively constant cost increase can be expected from firm generation capacity, since increased peak avoidance demands corresponding BTM capacity. For storage however, the trend is more nuanced. Due to the two-period TOU used, 8-hour storage in particular has strong cost recovery ability through arbitrage. This, coupled with the fact that at 10 GW, curtailment requirements rarely extend longer than 8 hours consecutively, means that 8-hour storage capacity scales with required flexibility, just as firm generation does. This trend changes at larger nameplates from longer duration needs, as will be observed in the following section.

4. **Shifting realities**: varying cost and lead times for turbines and natural gas fueled reciprocating engines, the complexities of the California permitting environment, and corporate emission targets can add substantial costs and complications that are not captured in the broad permitting assumptions incorporated into the model. This means that cost-effectively executing the natural gas fueled flexibility strategies in the present day, without usage of emergency backup generation, would likely incorporate a degree of storage to serve shorter intraday shifting while firm generation handles longer duration curtailment requirements (**figure 4**).

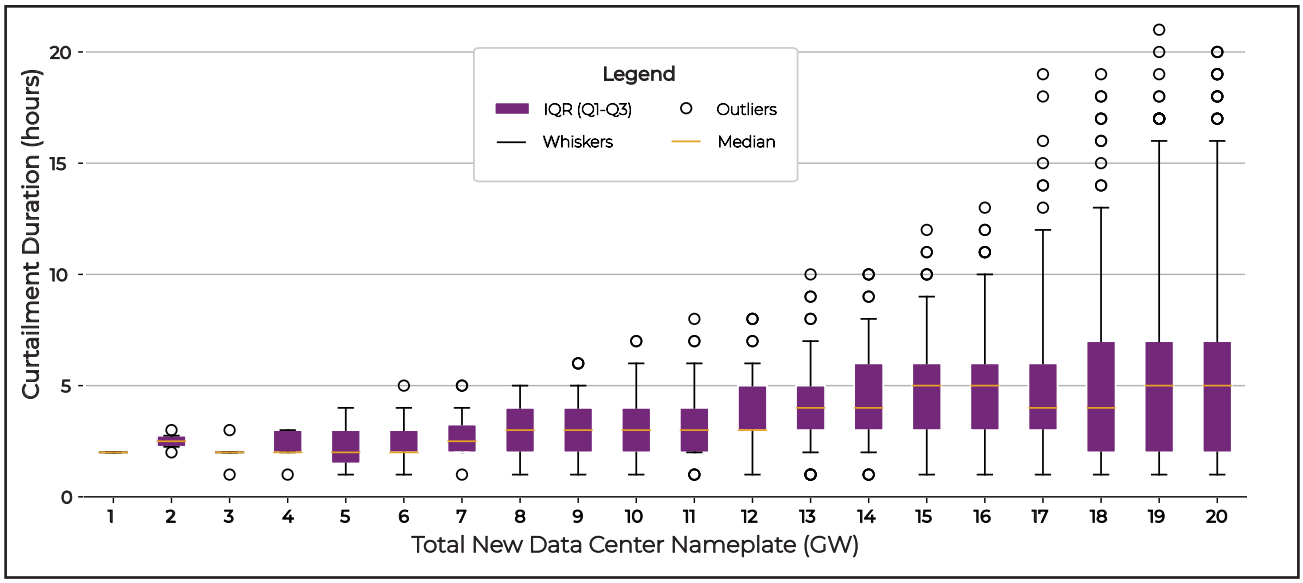
Nameplate Sweep

Each nameplate scenario was run across a range of data center nameplate capacity inputs, ranging from 1 to 20 GW. While the majority of this range may be unreasonably large, it is still valuable in demonstrating flexibility relationships, particularly within solar + storage scenarios. Each scenario holds flexibility fixed at 100%, meaning that there can be no increase in peak load from the pre-data center peak.

Perhaps unsurprisingly, in the "All Technologies" scenario the capacity mix increases at a near-constant rate with nameplate capacity growth. This is because appropriately sized firm generation can serve all hours where flexibility is demanded, and the modeled data centers must reduce their entire nameplate capacity during peak. As a result, they require at minimum an equivalent amount of BTM generation if emergency generators are not used. In the 'Low Air Pollution' scenario, the same trend remains, except the increase in operator cost is substantially higher, hovering around $15 as opposed to $4.40 for the 'All Technologies' scenario when zero backup generator hours are enabled. A more interesting trend is in the increase in costs observed in the solar and storage scenario. While the increase in observed costs is far greater than the previous two scenarios (~$25), it remains relatively constant until very high nameplate capacity (12GW). This can be attributed to a couple of factors:

- 8-hour batteries enable high cost recovery under the CB-6 TOU rate schedule, since the schedule is a two-period TOU with the higher cost "on-peak" time being 9 hours in length.
- Given the cost effectiveness of 8-hour batteries, nameplate capacity must be very large for flexibility mandated shifting to be longer than the 8 hours they provide.
- This trend deteriorates quickly at very large data center deployments, where curtailments are needed across longer periods of time (**figure 5**), and low RTE long duration energy storage is built. As nameplate capacity increases, required curtailments become greater in number and for longer periods. The cost implications of this are clear in **figure 6**.

Figure 5: Distribution of curtailment durations by new data center nameplate capacity (GW)*

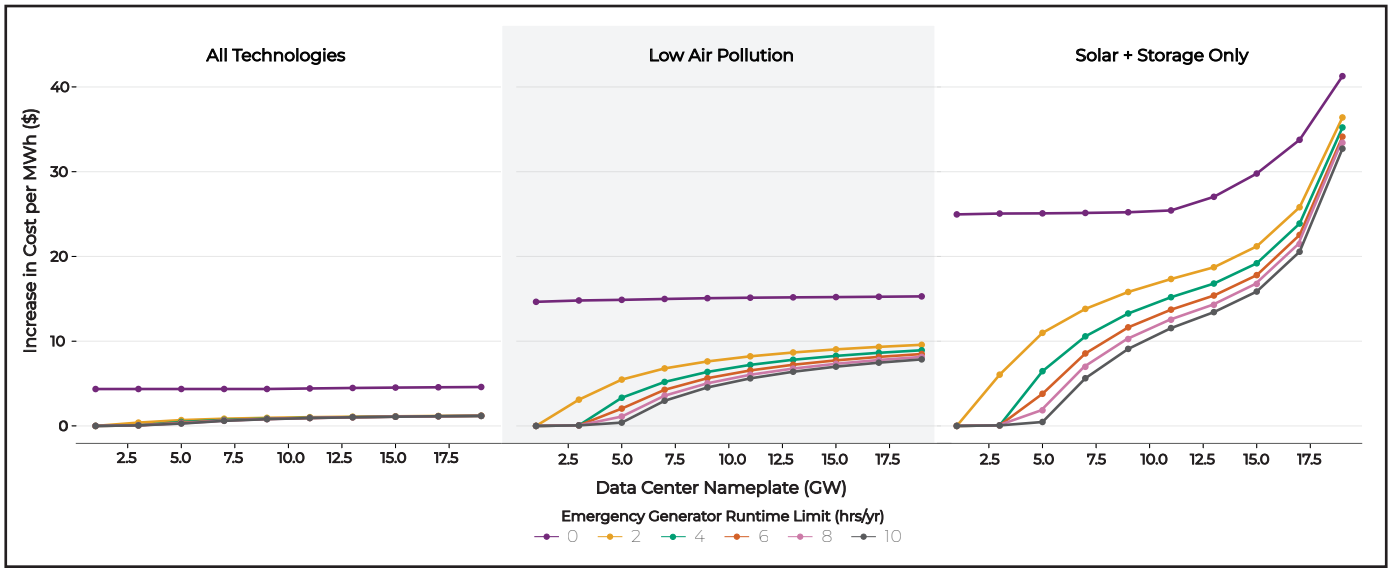


*Note: One outlier (20 GW, 43-hour curtailment duration) was removed for reasonable chart scale.

The increase in curtailment durations translates directly to cost increases in solar + storage scenarios. At low data center deployment, emergency generators can reduce costs. When the model is provided with zero emergency generator hours, the high power, low energy phenomenon is seen once more (**figure 6**). Full reduction of load is required for no more than a handful of hours, yet with no generator deployment, BESS or generation capacity must be scaled to meet the peak. Past these hours, the capacity is only used for intraday arbitrage on the selected rate schedule, leading to the substantial cost delta between the zero emergency generator hour scenarios and the scenarios with it enabled.

As workload shifting potential becomes better understood, it must be incorporated into future modeling. If it proves inexpensive relative to on-site generation and storage, it could displace much of the capacity otherwise required. Even if limited to shorter durations (<3 hours), workload flexibility may still be critical for handling curtailment that would otherwise demand extremely low–capacity-factor generation or storage. A central question is whether it will be more cost effective to assign workload shifting to cover all curtailment below a given share of nameplate capacity, leaving storage and generation for the remainder, or the reverse. Essentially, which option is more suitable for more consistent curtailment requirements over time, and which is preferable for short, large curtailments?

Figure 6: Increase in $/MWh by scenario, generator allowance, and Data Center Nameplate (GW)



Key Findings from Nameplate Scenarios

1. **Curtailment duration**: the sweep indicates a trend of increasingly long curtailment periods at higher data center deployment. With all technologies and limited policy burden, reciprocating engines are used as a low cost way to cover curtailment periods of any duration, resulting in a nearly flat cost increase line. In the low air pollution scenario, fuel cells are switched in, resulting in more significant costs from the higher capex investment. The solar + storage scenario exhibits the most drastic cost increases as curtailment durations demand less efficient, longer duration storage.

2. **Emergency generator impact on cost**: minimal cost increase is seen with usage of backup generators until 3 GW of data center capacity. Additional generator runtime can keep the costs low until 5 GW. This illustrates the impact of limited high-power generation, even when constrained to few hours of yearly runtime. As greater data center nameplate capacity is brought online, the value of backup generators is diminished, since high BTM power is required for more hours than the allocated generator runtime.

3. **The real solution space**: The high-power, short-duration reductions currently provided by emergency generators could, in some cases, be substituted or supplemented by other flexibility options, such as demand response or workload shifting (temporal or spatial). While emergency generators are the most nimble and well-suited resource for rapid curtailment, these alternative pathways should be explored to determine under what conditions they can deliver equivalent value, namely, when they can respond within similar timeframes, operate at comparable power levels, and align with the same peak-load periods. If these conditions cannot be met, and curtailment continues to rely solely on new or underutilized T&D or storage assets (BTM), total system costs are likely to increase.

Conclusion

Data center flexibility offers a near-term path to unlock headroom on constrained grids, but it is not costless. The analysis shows that while high levels of curtailment quickly become expensive if operators are limited to storage and renewables, small allowances of dispatchable on-site generation, especially existing emergency generators, can dramatically reduce costs. Importantly, these generators do not need to run often: even a handful of hours per year provides much of the system benefit until ~10 GW data center deployment. At lower data center penetrations, minimal runtime can almost entirely remove data center contribution to peak, while producing limited air pollution as compared to allotted maintenance and testing runtime. This highlights a core lesson of flexibility: high-power over a few hours goes incredibly far in grids with steep LDCs.

Alternatively, flexibility at low data center buildout can be met by higher CAPEX generation or storage, of which the required curtailments imply a minuscule capacity factor. While some cost recovery is possible for high RTE storage and natural gas fueled generation, depending on permitting processes these options can be costly for data center operators. A rising alternative may be to use a cost minimizing combination of storage and workload flexibility, however this strategy may still struggle with peak events that require near 100% load curtailment and longer duration curtailment requirements. While flexibility in the future may come from a combination of spatial load shifting, temporal load shifting, and on-site generation/storage, it is essential to consider the possibility of high reliance on on-site generation and storage and the costs associated with it.

The broader lesson is that flexibility is both real and finite. Early movers, given alignment with utility tariff design and policy, can capitalize on relatively cheap opportunities to reduce time-to-power and defer capacity upgrades, but as penetration grows in strict policy environments, the cost of flexibility rises. Understanding where backup generation can provide outsized value—without becoming a crutch for long-term planning—should be central to both utility tariff design and data center development strategy. The effective usage of on-site generation can bridge the gap between time-to-power needs and short to medium term constraints on grid expansion, enabling the potential for new loads and reduced ratepayer costs.

It is our hope that this paper is a starting point for the type of flexibility analysis that is increasingly necessary for data center demand growth. If load flexibility is to be taken up by utilities as the primary strategy to integrate new data centers, analysis of the costs associated with providing flexibility can be used to inform the development of rate designs, explore lower cost capacity expansion scenarios that incorporate data center ability to pay more for faster service, and enable decision making for optimal flexibility strategies.

Appendix: Model Overview, Inputs, and Assumptions

A.1 Model Overview

The model developed for this cost analysis is a linear optimization that produces the cost minimizing buildout and dispatch of BTM technologies across a specified period of time. The model operates with hourly granularity, and outputs the $/kWh costs faced by data center operators under the unique set of inputs. The objective function minimizes total cost to the modeled data centers. This model operates similarly to many capacity expansion models that support planning for capacity investments under a set of constraints, if a bit simplified for this exercise. The more novel contribution from this model is that it captures the impacts of flexibility needs, data center reliability requirements, and unique tariff designs, all within particular grid load profiles. This model uses a series of representative inputs for each run, given that it does not capture a spatial dimension.

Simplified Objective Function (C = cost in 2024 USD)

$$\min \quad C_{\text{total}} = C_{\text{capex}} + C_{\text{opex}} + C_{\text{fuel}} + C_{\text{grid\_energy}} + C_{\text{demand}} + C_{\text{shift}} + C_{\text{reserve\_capacity}}$$

The parameterized inputs to the model are the fixed, user-defined settings that shape each scenario. These include the hourly load shape for the modeled balancing authority, total data center nameplate capacity, maximum allowable generator runtime, technology cost assumptions, and tariff design parameters such as coincident demand charges and time-of-use energy rates. Certain technologies are further constrained by exogenous limits which are incorporated as nameplate-proportional capacity limits. These limits can include constraints such as rooftop space for solar PV or permitting restrictions for natural gas-fueled generation. Together, these parameters define the economic, operational, and regulatory boundaries within which the model must operate.

The endogenous variables are the decisions the model makes to minimize total cost while satisfying flexibility and reliability requirements. They include the optimal capacity buildout of each technology, dispatch schedules for those technologies, the level of imports from the bulk energy system, and the extent of cooling and compute load shifting.

The constraints applied ensure that the optimization produces results consistent with real-world technical, operational, and policy conditions. They include bounds on battery state of charge, dispatch limitations, reliability requirements, tariff compliance rules, and load shift feasibility limits, among others. These constraints prevent the model from selecting least-cost solutions that would be technically impractical or impermissible under modeled operational and regulatory frameworks.

A.2 Model Inputs

Determining the Technology Set

The first key model input is the technologies. The baseline technology set reflects BTM options that are either commercially mature or realistically deployable by California data centers within a time horizon of one to five years. Technologies were excluded if they had significant uncertainty in deployment timelines or commercial readiness, such as small modular nuclear reactors (SMRs) with widely varying estimates for initial deployment. Technologies were also excluded if they were too spatially constrained for meaningful buildout at typical hyperscale or co-location sites without major land acquisition, such as pumped storage hydropower and utility-scale wind.

The remaining baseline set includes commercial scale battery energy storage systems, emergency diesel generators, rooftop solar, prime diesel generators, natural gas-fueled reciprocating engines, fuel

cells, and some long duration energy storage (LDES). These technologies represent the main classes of dispatchable thermal, variable renewable, and storage resources that have some indication of BTM deployment feasibility. Gas turbines were removed due to current lead time delays. The degree to which each of these technologies are deployment constrained is further discussed in the sensitivity parameters section. The costs for the included set of technologies and associated fuels were gathered from a range of sources including the NREL 2024 ATB [23], the NREL REopt manual baseline estimates [24], the Energy Information Administration [25], and California Public Utilities Commission 2024-2026 IRP inputs [26].

Model Time Horizon

The model is run over a single representative year (2024) to ground results in current grid conditions, technology costs, and regulatory constraints. A one-year horizon avoids speculative assumptions about future tariff designs, interconnection timelines, or air-quality permitting that could materially change model outcomes. Additionally, this period reflects the near-present headroom and resource availability of the California grid, including substantial recent deployment of BESS technologies. Finally, it limits model complexity and compute resources required to run scenarios. The central drawback to this is that it makes the model myopic to changing future conditions, however it can still provide significant directional value, as intended for this report.

Location Selection

Given significant potential load growth and a worsening energy affordability crisis, California was selected as a suitable location to apply the model. This modeling exercise aims to determine the cost burden shifted to data centers under a specific tariff design, considering various flexibility requirements, nameplate capacities, and backup generator regulatory standards. The subsequent analysis will encompass an overview of the baseline model inputs and their underlying assumptions, the inputs subjected to sweeps or scenarios, and the resulting outcomes.

Input Load Profile

For model runs, we used the CAISO system load net of renewable generation. Data was sourced from CAISO's OASIS API by querying hourly actual load. The hourly load profile was then sliced to the modeling time horizon and adjusted to the timezone of other inputs (PST), where it was then parameterized in the model.

Discounting and Costs

A 7% discount rate is applied across all scenarios for discounting the costs of technologies. Capital costs are calculated using a capital recovery factor that is applied to the $/kW costs and then converted to costs representative of the model run timeframe.

Sweep Parameters

To evaluate the cost and operational implications of different data center flexibility strategies, we conduct a series of sensitivity runs that vary key input parameters. These sensitivities are used to explore how costs and technology choices shift under a range of plausible conditions. The model is particularly focused on identifying inflection points — when certain technologies become cost effective, when flexibility becomes disproportionately expensive, and how constraints like emergency generator runtime influence operator costs.

The three parameters that we sweep are flexibility targets, data center nameplate capacity, and enabled runtime of emergency backup generators. Flexibility and nameplate runs are distinct, with emergency generator runtime as a nested sweep within both. As defined previously, flexibility within this model is the reduction in peak relative to the firm load counterfactual. A 10% flexibility target means that the data center will have to curtail its load through a combination of BTM generation and storage, in order to limit the new maximum peak to be 90% of its nameplate plus previous peak load. All flexibility sweeps are run in 10% increments from 0 to 100. Each progressive increment demands additional curtailments from the data centers.

Nameplate capacity represents the amount of new firm data center capacity brought onto the grid. Within the model, nameplate capacity is divided into representative data centers that are each distinct under the selected rate schedule. For nameplate sweeps, flexibility stays fixed at 100% to illustrate the changes in data center operator cost as additional grid headroom is used and no capacity expansion is undergone.

Planned generator runtime is included for illustrative, highly constrained demand-response scenarios. California air districts currently prohibit the use of emergency generators for demand response; they may operate only for testing, maintenance, or true emergencies such as PSPS events or grid outages. Nevertheless, we include a small allowance of runtime to illustrate the potential cost impact if backup generators were incorporated into a formal demand response program or dispatched during severe grid stress. In doing so, we explore how even a very small allowance of backup generation could alter least-cost technology selection and cost curves. A more thorough overview of generator permitting can be found in the "Generator Permitting and Usage" section below.

A.3 Assumptions

Representing Spatially Constrained Technologies and Tariffs

The primary technology within the model that has a capacity factor dependent on location is solar. Within California, there is limited variation in capacity factors between most locations where data centers exist today, making it more reasonable to assume a single set of capacity factors from a data center dense location. A location that has significant data center density is Santa Clara, which houses an estimated~26% of California data centers. Given this, we used solar capacity factors from Santa Clara as representative for modeling purposes.

Another factor that isn't included in the model is the availability and cost of land. Given that many data centers are located near populated areas, significant land purchase for BTM solar would likely be uneconomic. As a result, solar buildout in the model is constrained to available rooftop space for rooftop solar. To calculate this, we estimated the kW of solar capacity per kW of data center buildout, and then scaled it down to account for rooftop space that is unavailable for solar[1].

We additionally used Silicon Valley Power's CB-6 Large Combined General Service [29] as a representative rate schedule. This rate schedule includes the option for large loads to select either a flat or time of use (TOU) schedule. It also includes a demand charge and capacity reservation charge if on-site generation is to be used, reflecting the system risk if data center load falls back to the grid. The model allows for simple toggling between flat and time of use (TOU) schedules; however, for the results, a TOU schedule is used to be slightly more representative of intraday pricing. By using this type of schedule, we make a core assumption that data centers will be able to access both a large load schedule, and be of the adequate size to take advantage of industrial-scale pricing. SVP requires a minimum combined monthly electric billing demand of 5,000 kW or more, which is easily achieved given the capacity of modern-day data centers [30].

Generator Permitting and Usage

The model incorporates emergency (standby) generators for data center backup power, with certain scenarios including limited runtime constraints on these generators. At this point, this is entirely theoretical, given that in California, the operation of emergency generators is strictly regulated across the 35 air districts. Emergency backup generators are typically permitted to run only for testing, maintenance, and true emergency conditions such as loss of load events or public safety power shutoff (PSPS) events, depending on the local air district's rules.

1 Estimated using California data center total from Baxtel and Santa Clara data center count from Data Center Map [27], [28]

Each California Air District sets its own permit conditions for stationary emergency generators. South Coast AQMD for instance allows a combined cap of 200 hours per year across emergency use and non-emergency testing and maintenance (50 hour cap within the 200 total) [31]. Emergency operations—including PSPS and CAISO Energy Emergency Alert (EEA) events—vary by district in being subject to an hourly cap. What differs between districts is often the emergency alert level or classification that triggers permission to operate backup generators during grid stress events. Many Air Districts rely on CAISO's Energy Emergency Alerts (EEA), which have three alert levels that are used to define at what stage of grid stress an emergency generator can legally run [17]. EEA guidelines are defined by the North American Electric Reliability Corporation (NERC) [32]. It is currently not possible for data centers to use emergency generators for non-emergency demand response within California.

Despite the structure of California regulation, emergency generator runtime is included in the model to demonstrate the impact of emergency generator usage for demand response. In this case, it is used as a planned grid asset rather than during load shedding events alone. While this is not the reality of current California regulation, it is actively being considered at the federal level. The EPA recently clarified that an EEA Level 1 (issued by BAs) counts as an emergency trigger when it's clear an escalation to Level 2 is imminent without intervention [33], [34]. That means this lower-level alert may be able to legally justify deploying generators under the 50-Hour Rule, provided local standards are met and EEA 2 is oncoming. As Kirkland and Ellis, a leading law firm note, "this interpretation regarding EEA Level 1 appears to differ from EPA's prior 2013 comment that EEA Level 1 situations cannot appropriately trigger engine dispatches under the 50-Hour Rule" [38]. This implies that this administration may find non-emergency usage of emergency generators far more palatable than the historical precedent. Overall, this lowers the federal floor for generator runtime; however, state requirements still bind if more stringent.

Strict limits on emergency generator use exist for good reason: diesel engines contribute to local air pollution and associated health impacts. However, this modeling assumes only minimal runtime for these units. In practice, the number of hours required to materially reduce flexibility costs across nearly all modeled nameplate capacities are far lower than the annual testing and maintenance hours already permitted in most California air districts. Modeling generators in this way is not meant to suggest widespread or routine dispatch, but rather to illustrate how even a handful of hours of high-power support can significantly reduce the need for costly new firm capacity. This sensitivity explores both the environmental trade-offs regulators face and the outsized system value that limited generator runtime can provide.

A.4 Limitations

There are a number of limitations inherent in a small scale model of this type, some of which have been noted briefly. The most relevant ones to establish or reiterate are the following:

1. **Single-year snapshot**. The model uses 2024 CAISO net load as a representative year. It excludes multi-year dynamics such as evolving technology costs, fuel price volatility, and long-term grid changes, making it myopic to a single year rather than reflective of data center lifetimes.

2. **Perfect foresight**. Peak hours and curtailment requirements are assumed to be known in advance. In practice, data centers would face some uncertainty in both timing and duration of curtailment events.

3. **Solar capacity factors**. The capacity factors for variable renewable energy resources are not perfectly captured through a single set, even if the capacity factor variation from location to location is relatively small in California.

4. **Flat load shape**. Data center load shapes are assumed to be flat, when in reality they have significant hourly variation and different PUEs depending on factors like ambient temperature, IT infrastructure type, cooling methods, etc.

5. **No nodal or T&D constraints**. Results assume that all CAISO system headroom is equally usable. In reality, interconnection is limited by local transmission and distribution bottlenecks that may limit the realistically useful system headroom.

6. **Technology simplifications**. Backup generators and batteries are modeled without start-time delays,

minimum run constraints, degradation, or part-load efficiency losses. This biases outcomes toward smoother, more optimistic operations.

7. **Tariff and regulatory generalization**. Results rely on a single tariff (SVP CB-6) and stylized permitting assumptions. Actual data center costs vary substantially across utilities, air districts, and regulatory regimes.

Given these limitations, the results should be interpreted as directional rather than prescriptive. They are useful for exploring how regulatory environments, flexibility requirements, and resource penetration levels shape cost trajectories, but they should not be relied upon for procurement decisions, permitting applications, or site-specific economics without a locational, multi-year framework. The analysis is intended to provide policy designers and operators with intuition about the cost trade-offs of different flexibility strategies, while grounding expectations for data center expansion in the constraints of today's grid.

References

[1] M. Schipper and T. Hodge, "After more than a decade of little change, U.S. electricity consumption is rising again - U.S. Energy Information Administration (EIA)," May 13, 2025. Accessed: Aug. 11, 2025. [Online]. Available: https://www.eia.gov/todayinenergy/detail.php?id=65264

[2] K. Baranko, D. Campbell, Z. Hausfather, J. McWalter, and N. Ranshoff, "Fast, scalable, clean, and cheap enough: How off-grid solar microgrids can power the AI race," offgridai. Accessed: Aug. 11, 2025. [Online]. Available: https://www.offgridai.us/

[3] R. Jones et al., "Annual decarbonization perspective: Carbon neutral pathways for the United States 2024," Evolved Energy Research, 2024. Accessed: Aug. 11, 2025. [Online]. Available: https://www.evolved.energy/us-adp-2024

[4] J. D. Wilson and Z. Zimmerman, "The Era of Flat Power Demand is Over," Grid Strategies, Dec. 2023. Accessed: Aug. 11, 2025. [Online]. Available: https://gridstrategiesllc.com/wp-content/uploads/2023/12/National-Load-Growth-Report-2023.pdf

[5] "Queued up: Characteristics of power plants seeking transmission interconnection," Lawrence Berkeley National Laboratory – Energy Markets & Policy, Apr. 2024. Accessed: Aug. 11, 2025. [Online]. Available: https://emp.lbl.gov/queues

[6] "Utility experiences and trends regarding data centers: 2024 survey," Electric Power Research Institute (EPRI), 2024. Accessed: Aug. 11, 2025. [Online]. Available: https://www.epri.com/research/products/000000003002030643

[7] T. H. Norris, T. Profeta, D. Patino-Echeverri, and A. Cowie-Haskell, Rethinking Load Growth: Assessing the Potential for Integration of Large Flexible Loads in US Power Systems, NI R 25-01. Durham, NC: Nicholas Institute for Energy, Environment & Sustainability, Duke Univ., 2025. Accessed: Aug. 11, 2025. Available: https://nicholasinstitute.duke.edu/publications/rethinking-load-growth

[8] "Grid flexibility needs and data center characteristics," Electric Power Research Institute (EPRI), 2025. Accessed: Aug. 11, 2025. [Online]. Available: https://www.epri.com/research/programs/063638/results/3002031504

[9] C. Seiple, "Gridlock: The demand dilemma facing the U.S. power industry," Wood Mackenzie Horizons. Accessed: Aug. 11, 2025. [Online]. Available: https://www.woodmac.com/horizons/gridlock-demand-dilemma-facing-us-power-industry/

[10] "Tier classification system," Uptime Institute. Accessed: Aug. 11, 2025. [Online]. Available: https://uptimeinstitute.com/tiers

[11] M. Osman, "Beginner's guide to different types of data centers," Nexcess. Accessed: Aug. 11, 2025. [Online]. Available: https://www.nexcess.net/blog/types-of-data-centers/

[12] K. Jacobs, "Supply shortages and an inflexible market give rise to high power transformer lead times," Wood Mackenzie. Accessed: Aug. 11, 2025. [Online]. Available: https://www.woodmac.com/news/opinion/supply-shortages-and-an-inflexible-market-give-rise-to-high-power-transformer-lead-times/

[13] L. Stone, "Gas turbine supply constraints threaten grid reliability; More affordable near-term solutions can help," RMI. Accessed: Aug. 11, 2025. [Online]. Available: https://rmi.org/gas-turbine-supply-constraints-threaten-grid-reliability-more-affordable-near-term-solutions-can-help/

[14] J. Jenkins, J. Farbes, and B. Haley, "Impacts of the one big beautiful bill on the U.S. energy transition—Summary report," REPEAT Project, July 2025. doi: 10.5281/zenodo.15801701.

[15] J. Moch, "Review of transmission lines since 2005." Harvard Dataverse, July 26, 2022. doi: 10.7910/DVN/MDQ6ME.

[16] "2025 data center power report," Bloom Energy, 2025. Accessed: Aug. 11, 2025. [Online]. Available: https://www.bloomenergy.com/news/data-centers-are-turning-to-onsite-power-sources-to-address-35-gw-energy-gap-by-2030/

[17] CAISO, "Emergency Notification Fact Sheet." 2023. Accessed: Aug. 11, 2025. Available: https://www.caiso.com/documents/emergency-notifications-fact-sheet.pdf

[18] V. Mehra and R. Hasegawa, "Using demand response to reduce data center power consumption," Google Cloud Blog. Accessed: Aug. 11,

2025. [Online]. Available: https://cloud.google.com/blog/products/infrastructure/using-demand-response-to-reduce-data-center-power-consumption

[19] M. Terrell, "How we're making data centers more flexible to benefit power grids," Google, Aug. 4, 2025. Accessed: Aug. 11, 2025. [Online]. Available: https://blog.google/inside-google/infrastructure/how-were-making-data-centers-more-flexible-to-benefit-power-grids/

[20] P. Colangelo et al., "Turning AI data centers into grid-interactive assets: Results from a field demonstration in Phoenix, Arizona," July 01, 2025, arXiv:2507.00909. doi: 10.48550/arXiv.2507.00909.

[21] S. Smith, "Will Memphis pay a price for Elon Musk's xAI 'Colossus' bait & switch?," Southern Alliance for Clean Energy, Dec. 10, 2024 . Accessed: Aug. 11, 2025. [Online]. Available: https://cleanenergy.org/news/will-memphis-pay-a-price-for-elon-musks-xai-colossus-bait-switch/

[22] "Elon Musk's xAI removes controversial gas turbines from Memphis data center," Datacenter Dynamics, May 7, 2025. Accessed: Aug. 11, 2025. [Online]. Available: https://www.datacenterdynamics.com/en/news/xai-removes-some-of-controversial-gas-turbines-from-memphis-data-center/

[23] "Technologies: Electricity 2024," NREL Annual Technology Baseline, 2024. Accessed: Aug. 11, 2025. [Online]. Available: https://atb.nrel.gov/electricity/2024/technologies

[24] K. Anderson et al., "The REopt web tool user manual," NREL. Accessed: Aug. 11, 2025. [Online]. Available: https://reopt.nrel.gov/tool/reopt-user-manual.pdf

[25] "Gasoline and Diesel Fuel Update," Energy Information Administration, Sep. 7, 2025. Accessed: Aug. 11, 2025. [Online]. Available: https://www.eia.gov/petroleum/gasdiesel/index.php

[26] "Inputs & Assumptions: 2024 – 2026 Integrated Resource Planning (IRP)," California Public Utilities Commission, Feb. 2025. Accessed: Aug. 11, 2025. [Online]. Available: https://www.cpuc.ca.gov/-/media/cpuc-website/divisions/energy-division/documents/integrated-resource-plan-and-long-term-procurement-plan-irp-ltpp/2024-2026-irp-cycle-events-and-materials/draft-2025-inputs-and-assumptions-document.pdf

[27] Baxtel, "California Data Centers & Colocation." Accessed: Aug. 11, 2025. [Online]. Available: https://baxtel.com/data-center/california

[28] "Data Center Map - Colocation, Cloud and Connectivity", Data Center Map. Accessed: Aug. 11, 2025. [Online]. Available: https://www.datacentermap.com/

[29] Silicon Valley Power, "CITY OF SANTA CLARA SILICON VALLEY POWER RATE SCHEDULE CB-6 LARGE COMBINED GENERAL SERVICE," 2025. Accessed: Aug. 11, 2025. [Online]. Available: https://www.siliconvalleypower.com/home/showpublisheddocument/62467/638718504925570000

[30] "AI data center growth: Meeting the demand | McKinsey," McKinsey & Company. Accessed: Aug. 11, 2025. [Online]. Available: https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/ai-power-expanding-data-center-capacity-to-meet-growing-demand

[31] "Fact Sheet and Frequently Asked Questions: Use of Backup Generators," Retail Compliance Center, Mar. 2025. Accessed Aug. 11, 2025. [Online]. Available: https://www.rila.org/retail-compliance-center/emergency-generator-permitting-matrix

[32] "EOP-011-4 – Emergency Operations," North American Electric Reliability Corporation, Feb. 15, 2024. Accessed: Aug. 11, 2025. [Online]. Available: https://www.nerc.com/pa/Stand/Reliability%20Standards/EOP-011-4.pdf

[33] "New EPA Guidance Clarifies When Data Centers and Other Operators May Utilize Emergency Backup Generators to Support Local Power Supply," Kirkland & Ellis LLP., May 12, 2025. Accessed: Aug. 11, 2025. [Online]. Available: https://www.kirkland.com/publications/kirkland-alert/2025/05/new-epa-guidance-clarifies-when-data-centers-and-other-operators-may-utilize-emergency-backup

[34] "EPA Clarifies Rules for Backup Generator Use," Sidley, May 15, 2025. Accessed: Aug. 11, 2025. [Online]. Available: https://www.sidley.com/en/insights/newsupdates/2025/05/us-epa-issues-new-guidance-on-data-center-emergency-generator-operations