2025 White Paper

Grid Flexibility Needs and Data Center Characteristics

TABLE OF CONTENTS

INTRODUCTION

The concept of flexibility in the bulk power system represents a system's ability to adapt to changing conditions. Flexibility is crucial for maintaining a reliable electric grid by improving the ability to balance electricity supply and demand from second to second while adapting to societal needs for energy over time. Traditionally, power plants provided this flexibility by adjusting their output to follow demand, and new infrastructure was built to enable both load and supply to grow. As load growth returns, electrification increases, and variable energy resources are added to the grid, ensuring that the system has the flexibility it needs has become crucial. Flexibility is critical to determining the essential characteristics of a regional electric system:

1. **Cost of providing energy:** More flexible systems avoid energy or ancillary service price spikes and negative price conditions, net the expense of investment in enabling the capability to be flexible.
2. **Power system reliability:** More flexible systems can rapidly align supply and demand over short-term operating windows of hours to seconds as well as longer term horizons of decades to years.
3. **Network planning and operation:** More flexible networks can rapidly adapt to changing grid conditions within operating limits, lessening grid congestion and improving asset utilization.
4. **Environmental impacts:** More flexible systems improve the efficiency of plant dispatch, avoiding the need to run assets for extended periods at suboptimal heat rates—minimizing spilling water and curtailing renewables.
5. **Grid access:** Flexible systems may accelerate opportunities for grid access for both generation and demand alike, accelerating time to connect for new projects.

Data centers are becoming significant components in the bulk electric system, with some individual facilities approaching the scale of these largest generation facilities. Two primary challenges in accommodating data center demand are

1. facilitating data center grid connection in a timely fashion
2. holistically integrating data centers into daily operations of modern and emerging power systems.

The Data Center Flexible Load Initiative (DCFlex) project addresses both challenges. To do so, common terminology is a prerequisite to broad implementation of flexibility. This white paper takes a first step toward developing a common understanding of flexibility between power providers, grid operators, and data center stakeholders across four sections:

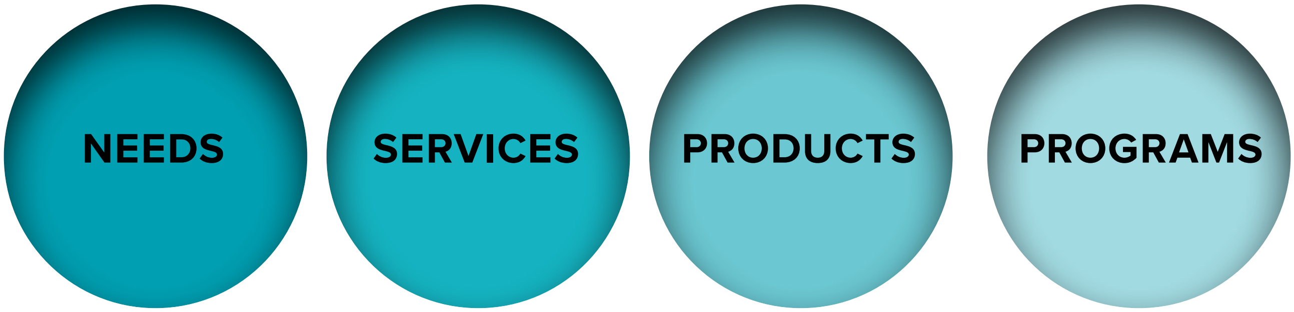
- Section 1 contextualizes flexibility in terms of flexibility needs, services, products, and programs. It concludes with a brief description of grid needs for flexibility over various time scales.
- Section 2 outlines various data center types and how they differ and can be classified to inform grid integration.

- Section 3 explores where flexibility may arise in a data center and factors that may influence its flexibility potential.
- Section 4 considers how the capabilities of data center–related flexibility and grid needs for flexibility services may be matched to the benefit of both.

FLEXIBILITY FRAMEWORK

Flexibility is a common language term that has been invoked in power system vernacular to refer to a wide range of conditions, challenges, or characteristics. Without a common taxonomy and clear definitions, it becomes challenging for stakeholders to work together efficiently, often resulting in misaligned goals and expectations. Furthermore, although power system operators have an intuition about what flexibility may refer to, the term may be interpreted differently in the data center community. By establishing a framework with well-defined terms, stakeholders can develop a common understanding of flexibility—enabling faster, more effective collaboration.

The flexibility framework proposed here combines four building blocks to address system needs arising from mismatches between supply and demand in the power system over a variety of time frames: needs, services, products, and programs.



Needs

A need represents a scarcity of some technical capability that is required to ensure the reliable, economic operations of the power system. These are the core objectives of bulk power system operation and design. An understanding of the relevant needs is crucial for developing relevant and effective solutions.

Needs associated with flexibility include an element of managing the requirement for the power system to change dynamically over **time**, or across **locations**, in intervals ranging from milliseconds to decades.

Relevant flexibility-related needs may include management of transient imbalances between generation or network capacity and demand requests, congestion management during periods of asset outages, net load variability or uncertainty management, real-time balancing, voltage management, or frequency control.

Services

A service refers to the process dedicated to fulfilling or solving one or more identified needs. For this report, a service can be understood as a process to access, trade, activate a response, validate performance, and settle among participants.

In the context of electricity markets, an example of service may be an intraday energy market, which serves the need to balance supply and demand for the hours ahead, using information that is available close to real time as well as optimally allocating ancillary services and resolving grid congestion issues.

Products

A product is the tradable requirement or obligation that adheres to precise specification, which is developed to meet a set of identified needs and orchestrated through a service. Several products may be exchanged through a single service, and, in some limited conditions, the same product may be exchanged through several services.

In the case of the example above, the primary product being traded in intraday energy markets is energy, measured in megawatt-hours (MWh) per interval.

Programs

A program is a specific implementation structure that enables a provider to deliver the product offered. These programs are typically designed by utilities, aggregators, or regulators and define eligibility criteria, compensation mechanisms, and participation obligations for participants. These programs may reduce the complexity of compliance with service interfaces and product specifications into simplified incentives or requirements for end-use customers.

For instance, demand and large energy users often have the opportunity to participate in demand response programs, critical peak pricing programs, or flexible grid interconnection programs in a variety of systems. These programs transform tradable products exchanged through services into customer-facing offerings. Examples of needs, services, products, and programs are listed in Table 1.

Table 1. *Examples of Needs, Services, Products, and Programs*

Flexibility Need	Service	Product	Program
Manage net load variability and uncertainty	Capacity markets	Capacity obligation	Critical Peak Power
	Day-ahead energy market	Day-ahead energy	Distributed energy resources (DER) programs
	Intraday balancing markets	Real-time energy	Demand response programs
Frequency control	Ancillary services markets	Automatic frequency restoration reserve	Interruptible load

Dimensions of Flexibility

To give structure to characterize needs, services, and products, a common language is needed to describe them. By defining the following dimensions, products and services can be effectively assessed:

- *Lead time* refers to the time ahead of the delivery period when a decision is made to activate a product. For example, in the case of frequency regulation services, a product or service may be procured in months to minutes ahead of delivery intervals. In the case of structural congestion relief, the lead time to develop network reinforcements may be decades ahead.
- *Quantity/volume* is the magnitude of the product being requested or delivered. For example, this could represent the megawatts (MW) held for ancillary service provision by a resource providing reserve.
- *Duration* measures the length of a product's delivery time interval. For instance, a battery energy storage system may be suitable for imbalance management needs of several hours. However, if the duration of the imbalances are longer, thermal storage systems or reservoirs from hydropower plants are needed. Real-time markets in U.S. independent system operators typically operate with 5-minute market intervals.
- *Direction* indicates whether the aim is to increase (upward), decrease (downward) supply or demand, or both. An example of an upward directed product is the energy provided by peak thermal plants when additional power is needed because of an imbalance. Conversely, a downward directed product can be the forced curtailment of solar energy plants when demand is not high enough to meet the supply.
- *Location* specifies the geographical scope of the service or product—if it applies, such as nodal, zonal. For instance, distributed energy resources (DER) can provide nodal products by addressing local imbalances. Over broader areas, imbalances may rely on larger power plants, for example, one country using neighboring countries' thermal plants to balance their power system.
- *Availability* represents the degree to which the need occurs continuously or in discrete events or in specific conditions. This in turn informs the design of the service and types of resources that may be able to respond to those needs.

- *Certainty* refers to the accuracy with which the need can be forecasted and the behavior of the responding resources anticipated. Certain needs have high uncertainty, such as forced outages, necessitating services that specify products requiring high availability. Others may be better anticipated, such as day-ahead congestion, which may value resources being available during specific conditions.

Grid Flexibility Needs

The DCFlex project explores how a range of potential flexibility solutions can accelerate the integration of new large loads into power systems and enhance system reliability. Although each region and power system is unique and may face different needs, the most common needs today include the following:

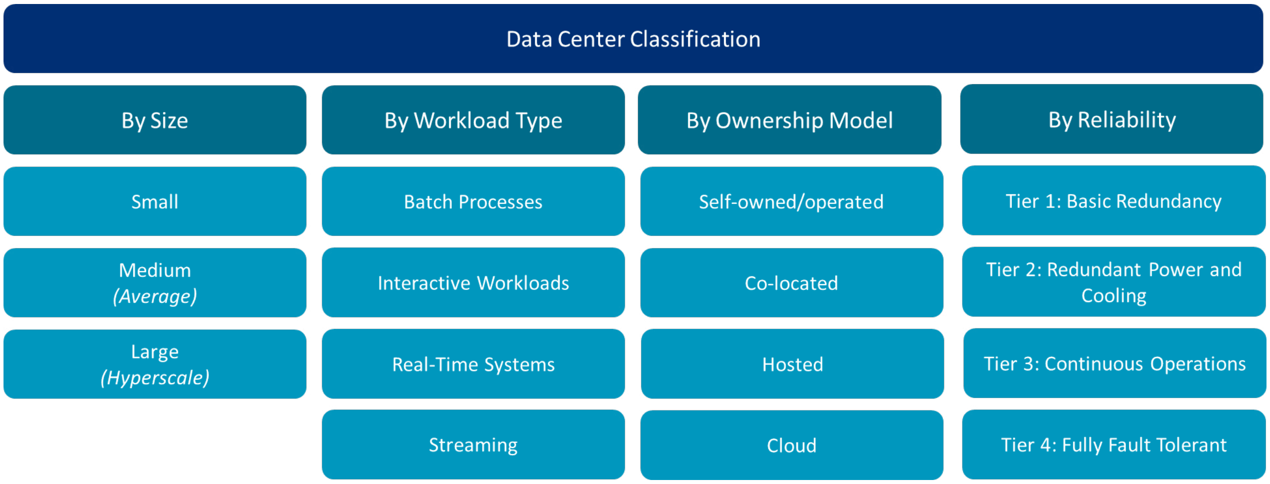
- Addressing the mismatch in timing between when new load wishes to connect and operate and the time frame for enabling network investments to allow those loads to run unrestricted. This issue arises in the planning time frame.
- The mismatch in time between new data center deployment and new generation investment, resulting in a need to address resource adequacy during scarcity conditions. This issue also arises in the planning time frame.
- The management of forecasted net load variability in the day-ahead time frame giving rise to a need to schedule generators and loads in forward markets and to manage forecast uncertainty and congestion in real-time dispatch and balancing actions. This issue is managed in the operational time frame.
- The need to support reliable operation during faults or abnormal system conditions through ancillary services such as frequency control operating reserve, load curtailment by participating in remedial action schemes, or managing harmonic distortion on the network. This issue is similarly operational in nature.

The following sections establish the general terms to describe data centers and potential sources of flexibility as well as how these may address the grid flexibility needs described in this section.

DATA CENTERS: NOT ALL THE SAME

To understand and evaluate the relationship between a data center and the host power system, it is helpful to differentiate between classes of data centers, which can behave and interact with the grid differently.

This section outlines a taxonomy of data center characteristics that are meaningful in the context of grid integration. Four categories are selected, based on their influence on data center flexibility and potential power system impacts: size, reliability, workload, and ownership model. A summary of these is presented in Figure 1 and further explained in the following text.



Data Center Classification

By Size	By Workload Type	By Ownership Model	By Reliability
Small	Batch Processes	Self-owned/operated	Tier 1: Basic Redundancy
Medium *(Average)*	Interactive Workloads	Co-located	Tier 2: Redundant Power and Cooling
Large *(Hyperscale)*	Real-Time Systems	Hosted	Tier 3: Continuous Operations
	Streaming	Cloud	Tier 4: Fully Fault Tolerant

Figure 1. *Data center characteristic classification*

Classification by Size

Data center size is a significant determinant of their importance and potential impact on the grid. First, larger data centers have a greater impact on systemwide demand and flexibility needs. Larger data centers, or concentrated clusters of many small data centers, bring greater infrastructural needs while having the potential to support power systems. Data center size also directly influences the ways in which they may participate in providing flexibility services (for example, direct wholesale market participation can be possible for larger data centers, while aggregation may be needed for smaller ones).

Historically, the term *data center* was used to refer to a range of computing facilities, from small server rooms and server closets to large, stand-alone buildings specially designed to house and support information technology (IT) equipment. Over the past ~15 years, businesses have shifted more of their IT applications to environments that are hosted off-premises, both in the cloud and at co-location data centers. This shift has reduced the size of local on-premises data centers and increased demand for larger cloud data centers—changes that have driven growth in large data centers in terms of both size and number. Several classifications of data centers by size exist and are evolving. A recent definition of data center size, server count, and power needs by Dgtl Infra is shown in Table 2.[1]

Table 2. *Modern-Scale Large Data Centers*

Size	Small	Medium (Average)	Large (Hyperscale)
Building size	5,000–20,000 square feet (ft^2) (465–1,858 square meters [m^2])	20,000–100,000 ft^2 (1,858–9,200 m^2)	100,000 ft^2–1M+ ft^2 (9,200–92,903+ m^2)
Server count	500–2,000	2,000–10,000	10,000–100,000+
Power consumption (peak load and annual consumption)	1–5 MW (~0.01–0.05 terawatt-hours [TWh])	5–20 MW (~0.05–0.2 TWh)	20+ MW (~0.2–1.0+ TWh)

Classification by Workload

Data center classification by workload type, in addition to size, improves the understanding of a specific asset's potential for flexibility. As the data center industry has matured, several variations of business models have evolved that blur the lines between data center types; however, several broad categories are proposed next. Note that data center workload categories can be nonbinary, with the same facility conducting workloads of several categories with a combination of hardware, in certain cases tailored to specific applications.

- **Batch processes.** Batch processes are discrete workloads that may combine retrieval and writing of data from memory and data processing or computation. Although batch processes may not require user interaction and may operate autonomously, they may also be triggered by certain conditions or depend on other tasks. Workloads may be micro-sized (for example, patch updates) or very large (for example, artificial intelligence [AI] model training run).

  Electric demand during such a process depends on the type and complexity of the operation or workflow that is executed. Batch workloads offer some potential to be scheduled (for example, backup operations) or triggered (for example, queued tasks), making them potentially tolerant to delay. Cloud service providers may offer variable spot pricing to induce such applications to execute during preferred periods of lower demand for the data center's assets.

- **Interactive workloads.** These processes are characterized by their user-driven nature. Examples of these workloads include web services and user-driven database queries involving iterative sessions between user and back-end processes. Examples of this workload may include e-commerce websites and enterprise systems.

  Interactive workloads may follow a daily usage cycle to some extent. They are typically subject to service-level agreement requirements and on-demand usage

---

[1] https://dgtlinfra.com/data-center-power/.

patterns and require high responsiveness. Interactive workloads are less likely to be delay tolerant, given their higher opportunity cost for users.

- **Real-time systems.** Real-time workloads include always-on services such as Internet of Things (IoT)-based control or analytic systems, online productivity or collaboration tools, and online gaming servers, which combine user interaction with computational workloads. Such workloads combine rapid read/write, data processing, and messaging actions to support high uptime, on-demand, or always-on applications. Performance is managed to ensure availability, reliability, and performance against critical time metrics.

  Electric demand for this type of application may remain relatively constant over time, with periodic bursts of demand triggered by use-specific situations. These are mission-critical applications, are more sensitive to delay, and have limited potential to move across data center locations.
- **Streaming.** Streaming workloads involve high-throughput, continuous data flows with a requirement for low-latency performance. Examples include video conferencing, media streaming, logging, monitoring systems, high-frequency trading systems, and content delivery networks.

  Like the real-time use cases, streaming loads are on-demand and may be 24/7 systems that also benefit from proximity to load. Streaming loads are highly likely to be delay sensitive but potentially amenable to movement across adjacent locations.

Classification by Reliability

Data center workloads represent a range of applications that have differing impacts in the event of unavailability, ranging from low to no consequence to mission critical. Standards such as the Uptime Institute data center tiers or ANSI/BICSI 002 are defined to delineate between data center designs that enable various classes of expected availability performance. Unlike power system design where one common, almost-always-on reliability standard is applied for all users, data center operators may make a design choice that best suits the workload.

- **Tier 1: Basic redundancy (N).** This design tier provides basic protection against normal run of business uncertainties such as voltage sags or human error. Uninterruptible power supply (UPS) and standby generation are present as well as cooling, but these are not redundant and may be taken out of service during maintenance and completely unavailable during unplanned outages. A single power supply path is used, and standby generators are available to provide backup generation. Standby generators are not intended for continuous operation for prolonged periods (that is, beyond several hours).
- **Tier 2: Redundant Power and Cooling (N+1).** This design tier provides some redundancy in both power equipment (UPS, generators, fuel tanks, fuel cells, and cooling, such as pumps or chillers) but may become exposed to the impact of unexpected equipment failure during periods of maintenance. Such designs include multiple standby generators, as in Tier 1.
- **Tier 3: Continuous Operations (2N).** Redundant power supplies across two separate paths and concurrent maintainability for critical systems allow any one asset to be removed from service for maintenance while maintaining tolerance to faults. Multiple on-site generators can run continuously for extended periods but are not required to be able to maintain production for a period beyond several days. These generators are classified as prime power sources.
- **Tier 4: Fully Fault Tolerant (2N+1)+.** This tier includes redundancy on all systems as well as external interfaces. Designs ensure high tolerance to faults without any asset constituting a single point of failure, and concurrent maintainability is ensured. Multiple on-site power generation resources are capable of continuous long-term operation.

Classification by Ownership Model

A final consideration, as it pertains to assessing data center capabilities for flexibility provision, is the data center ownership model. Ownership models are important because of the various roles and responsibilities that end users of the data center—and the data center owners—play under each model that can influence the requirements for availability, service usage, and reliability. The main ownership models are as follows and listed in Table 3:

- **Self-owned/operated.** The user of the data center also owns and operates it. In this case, the data center responds to the business needs from a single entity, which can facilitate flexibility provision to the grid from the unit.

- **Co-location data centers.** Space is leased by data center owners to multiple companies in a co-located environment. These arrangements can involve the provision of power, cooling, networking, redundancy, and physical security as services provided to the lessees. Some co-location providers offer to build, own, and operate a data center for a single customer, which can be a hyperscale facility for a cloud provider. Sharing data center space with multiple companies—each with different business goals—may limit available flexibility compared to self-owned/-operated data centers.[2]
- **Hosted.** The equipment and the data center itself are owned by the provider, which leases the servers to the customer. The services, storage space, and servers of the customer are allocated in the provider's racks. Hosting clients are responsible for orchestrating their workloads with the contracted capability operating within a defined service-level agreement (SLA).
- **Cloud.** The data center is typically owned and operated by hyperscale service providers who offer access to specific services to third parties using SLAs, end-user license agreements, or similar. Cloud services may comprise any of the workload types mentioned previously.

Table 3. *IT User and Data Center Operator Responsibilities Under Each Ownership Model*

Ownership Model	Owns DC	Owns Racks	Arranges Maintenance	Has Access
In-House	IT user	IT user	IT user	IT user
Co-located	DC	IT user	IT user	IT user
Hosted	DC	DC	IT user	IT user
Cloud	DC	DC	DC	IT user[3]

HOW MIGHT DATA CENTERS PROVIDE FLEXIBILITY?

Once a data center has been characterized, the potential capabilities to provide flexibility become clearer. Data centers comprise common components and systems that may each be flexible; however, similar to the grid, the constraints within which they operate may influence their potential to respond. Nevertheless, flexibility potential may exist to some degree for each data center subsystem that can be described using similar characteristics to those used to describe grid needs:

- **Potential capability** refers to the power modulation capability an asset or subsystem can deliver or reduce.
- **Notification time** indicates the notification period required by the operator to allow the asset or subsystem to react to the request to change the power profile.
- **Duration sustained** measures how long an asset or subsystem can sustain an instruction to deviate from its baseline power profile.
- **Availability** indicates the portion of time the asset or subsystem is likely to be available to respond to a request to modulate.

The purpose of Workstream 1 of the DCFlex project is to be able to assess how the combination of each of these characteristics for a range of data center assets and subsystems may be potential sources of flexibility to alleviate both short- and long-term needs. There are three main sources of flexibiltiy within the data center: compute assets, balance of plant, and power assets.

Compute

A data center's core computing infrastructure consists of memory, processors, and networking devices, which are responsible for processing and storing vast amounts of data—

---

[2] Uptime Institute Global Data Center Survey Results 2024 - Uptime Institute.

[3] Nonphysical access only.

enabling communications and information security. Power consumption varies depending on the job type and related workload, with high-performance computing (HPC) or AI training compute tasks requiring significantly more energy than that required for storage, for example.

Computing infrastructure can use the following strategies to provide flexibility services:

- **Scaling.** Adjustments to the computing power or availability of compute for tasks may be possible when the use cases allow it. Examples of this may include reducing the number of central processing units (CPUs) or graphics processing units (GPUs) available to tasks such as AI training, avoiding power demand at the expense of performance. Modern processors and servers may also support power-saving and performance modes that may slow computation; for example, dynamic voltage and frequency scaling adjust processor clock speed to allow voltages (and power consumption) to be reduced. Other techniques offer the possibility to increase power demand.
- **Shifting.** If the power requirements to complete the task cannot be scaled down, batch-processed tasks can be postponed or moved to another location.
  - **Time.** If the task is non-time-sensitive, it can be scheduled or delayed during grid stress periods to non-peak hours.
  - **Location.** If the task must be completed at that time (for example, real-time) but is not location-sensitive (for example, financial transactions or emergency services), compatible workloads may be relocated to other sites in regions experiencing less grid stress at that time.

Balance of Plant

- **Cooling systems** regulate the temperature within data centers to prevent overheating of IT equipment, enhancing energy efficiency and ensuring performance to SLAs. These systems typically consist of air conditioning units, chilled water loops, and economizers, which can require a significant portion of the electricity used in a conventional data center. In certain circumstances, the power consumption of cooling systems may be modulated, offering temporary reductions in energy use, or may be augmented with thermal storage.
- **Electrical network configuration** may be designed to enhance reliability by including redundant network connections with multiple connections to independent substations. This configuration allows data centers to alter the flow of power in response to grid conditions. Similarly, the potential exists to provide system services such as reactive power or dynamic response in the case of grid disturbance.

Power

On-site power refers to all the power generating auxiliary systems, power generation capability, and energy storage located behind the meter or in direct proximity to the data center or group of data centers. Following are the most relevant components:

- **Uninterruptible power supply.** UPS systems act as the first line of defense against power disruptions by providing instantaneous backup power and helping smooth transitions to backup generation sources, explained next. Their consumption represents only 1% of the total.[4] However, even a brief power loss of 20 milliseconds can lead to IT system failures, while 1-minute outages could require several hours of recovery processes to restore the affected applications and systems. UPS systems are crucial to maintain secure operation.

  In a default design configuration, UPS systems are designed to supply the minimum uninterruptible load until backup generators can start providing power. However, alternative designs may extend the time through which the UPS can supply power by adding battery capacity.
- **Standby generators** provide a fast-start, longer term energy security capability to data centers during extended grid outages. Standby generators are typically powered by diesel. Natural gas, biogas, linear generators, and hydrogen fuel cells are emerging in some cases. Depending on the data center reliability tier, backup generation capacity may be equivalent to or significantly greater than the IT demand. Standby generators require seconds to minutes to fully activate, necessitating UPS as a bridging measure.
- **Customer-sited generation.** Data centers with on-site generation—such as gas turbines, renewables, or other similar resources—may provide power production capability: continuous, prime power, or peak power. These may be located behind or in front of the meter. The potential capability of on-site generation can be as

---

[4] A new, efficient UPS can help achieve sustainability goals at the edge.

high as the capacity installed, with its availability and duration depending on weather conditions if renewable assets are being used.

- **Customer-sited energy storage** allows data centers to leverage grid power or on-site generation to charge behind- or front-of-meter storage to provide power later. Storage may be used as part of a UPS or as a stand-alone asset.
- **Contracted generation** involves a generation plant not directly connected to the data center or owned by the same owner but that provides power or services through a power purchase agreement, or similar agreement, that provides a data center owner with a way to hedge energy procurement costs and manage emissions over time.

Table 4 shows a stylistic representation of the potential capability of each component or data center subsystem to be flexible along with a response time that may be possible in which to activate the resource, a duration, and a high-level estimation of availability. This type of assessment is helpful to systematically relate asset classes to needs.

Table 4. *Flexibility Potential from Data Center Assets and Subsystems*

Underlying Asset	Potential Capability	Response Time	Duration Sustained	Availability
On-site generation	▮▮▮▮	▮▮▮▯	▮▮▮▮	▮▮▮▯
On-site battery storage	▮▮▮▮	▮▮▮▮	▮▮▯▯	▮▮▮▮
On-site thermal storage	▮▮▮▮	▮▮▯▯	▮▮▮▮	▮▮▮▯
Cooling systems	▮▮▯▯	▮▮▮▮	▮▮▯▯	▮▮▯▯
Standby generation	▮▮▯▯	▮▮▯▯	▮▮▮▮	▮▮▮▮
UPS	▮▯▯▯	▮▮▮▮	▮▮▮▮	▮▮▮▮
Compute	▮▮▮▯	▮▮▯▯	▮▯▯▯	▮▮▯▯

What Factors Affect Flexibility Potential?

Technical, economic, and operational factors determine the extent to which data centers or the grid may offer flexibility to enable successful integration of new loads. Here we describe the main considerations in evaluating the potential and performance of measures to increase flexibility.

Opportunity Cost

One of the fundamental constraints on data center flexibility is the opportunity costs associated with reducing power consumption from the grid. The following factors play a key role in determining economically accessible flexibility:

- **Lost revenue potential** represents the loss of income to the data center, cloud service provider, or end user because of the unavailability of computational capability. This is the most significant factor.
- **Service-level agreements** may mandate high availability and performance, limiting the extent to which power consumption can be reduced or postponed.
- **Customer preferences** for uninterrupted, highly performant services can also restrict flexibility, especially for latency-sensitive applications.
- **Competition** between data center owners, operators, and cloud service providers drives incentives to offer high uptime to new users of data center capabilities.
- **Time to power** represents the benefit of accelerating timelines to connect a new data center, potentially at partial load for a specific duration.

Economic Cost to Implement

The cost of enhancing flexibility capabilities or more flexible technologies into data center operations varies across each option—for example, software-based measures are more capital efficient than the addition of new storage or generation assets.

Notice Periods

The ability of a data center to provide flexibility depends on the notification period given by system operators. Longer lead times allow the potential for planned adjustments, while real-time or short-notice flexibility requirements may be more challenging to fulfill by, for instance, altering the cooling system, using thermal storage, or firing up gas-fueled plants.

Controllability

For certain flexibility services, grid operators require high certainty and direct controllability of the service provider. This occurs in cases of runback or remedial action schemes that operate automatically and locally through protection devices to protect against fast transient events. Such schemes may be required to ensure compliance with grid planning and operation standards. In other cases, an indirectly incentivized response may be suitable—for example, network tariff designs.

Weather

Weather influences both cooling demand for electricity and the production of customer-sited renewables. In some situations, the impact of weather may be to increase load while increasing production. In other cases, the situation may be reversed for supply. Holistic assessment of all systems in and around a data center is required to understand the net impact of the interaction on the system.

Proximity

Both the need for flexibility services and the potential capability of data centers to provide flexibility may be influenced by proximity among data centers in the grid. Large clusters of data center demands may concentrate the need for flexibility in a specific region, whereas dispersed data centers may spread the same requirement around a larger area. However, clusters may be able to leverage economies of scale to develop flexible solutions that would otherwise not be possible.

Environmental Considerations

Several data center operators have emissions reduction goals, often tied to corporate sustainability policies. Regulatory emission limits—for example, air permits—can limit the use of backup diesel generators or on-site gas-powered plants as a flexibility resource when needed by the system operator. Noise limitations also restrict the run time of on-site generation in certain areas. Combined, these requirements constrain location and power generation technology choice.

Ownership Structures and Business Models

The ownership of a data center influences its ability to provide grid services. Enterprise- and corporate-owned data centers may have flexibility priorities different from those of co-location facilities. Colocation data centers, which often serve multiple customers, can have complex contractual obligations that complicate flexibility actions. Decision-making processes, stakeholder interests, and compliance requirements further affect the potential responsiveness of data centers to grid conditions.

Fuel Supply Constraints

Finally, for data centers relying on fossil fueled plants as one of their power supply sources, fuel constraints can be a limiting factor for implementing flexibility strategies. Availability of on-site fuel storage, supply chain reliability, price volatility, and regulatory restrictions on fuel usage impact the feasibility of using that source for demand response. These factors can also apply to data centers relying on backup generators as part of their flexibility approach.

Emissions

Air quality permitting procedures typically limit the extent to which conventional standby or customer-sited diesel generation may operate.

Market Rules and Tariffs

For market and nonmarket areas alike, the rules governing the operation of the system may influence the ability of flexible solutions to be developed. The impact of such rules and the consequences of their incentives may accelerate or limit the potential for flexibility to be accessed and deployed.

MATCHING SYSTEM NEEDS WITH DATA CENTER FLEXIBILITY POTENTIAL

Previous sections described the current critical needs for grid operators as they pertain to data center integration and flexibility. By systematically identifying flexibility needs, the services implemented to address those needs, and the related products and functional understanding of the components and subsystems in a data center, collaborative solutions across energy and IT may emerge more rapidly.

The DCFlex project aims to build an evidence base that can guide the development of flexibility solutions that accelerate the integration of new large loads and improve reliability. The framework presented here provides a structured underpinning to project modeling, simulations, and demonstration that will occur throughout the project.

The next step in the project is to identify a set of use cases where flexibility needs may be met through innovative services, flexibility products, and new designs or operation of related technologies. Through a progressive approach of modeling and demonstration, the project can improve the evidence base to guide the adoption of successful approaches.

This collaboration between data centers and system operators will be pivotal in ensuring a stable, resilient, affordable, and sustainable power grid and the continued development of a digital society. With the right strategies and technologies in place, data centers may successfully integrate with the power system while facilitating growth in the sector.

DISCLAIMER OF WARRANTIES AND LIMITATION OF LIABILITIES

THIS DOCUMENT WAS PREPARED BY THE ORGANIZATION(S) NAMED BELOW AS AN ACCOUNT OF WORK SPONSORED OR COSPONSORED BY THE ELECTRIC POWER RESEARCH INSTITUTE, INC. (EPRI). NEITHER EPRI, ANY MEMBER OF EPRI, ANY COSPONSOR, THE ORGANIZATION(S) BELOW, NOR ANY PERSON ACTING ON BEHALF OF ANY OF THEM:

(A) MAKES ANY WARRANTY OR REPRESENTATION WHATSOEVER, EXPRESS OR IMPLIED, (I) WITH RESPECT TO THE USE OF ANY INFORMATION, APPARATUS, METHOD, PROCESS, OR SIMILAR ITEM DISCLOSED IN THIS DOCUMENT, INCLUDING MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, OR (II) THAT SUCH USE DOES NOT INFRINGE ON OR INTERFERE WITH PRIVATELY OWNED RIGHTS, INCLUDING ANY PARTY'S INTELLECTUAL PROPERTY, OR (III) THAT THIS DOCUMENT IS SUITABLE TO ANY PARTICULAR USER'S CIRCUMSTANCE; OR

(B) ASSUMES RESPONSIBILITY FOR ANY DAMAGES OR OTHER LIABILITY WHATSOEVER (INCLUDING ANY CONSEQUENTIAL DAMAGES, EVEN IF EPRI OR ANY EPRI REPRESENTATIVE HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES) RESULTING FROM YOUR SELECTION OR USE OF THIS DOCUMENT OR ANY INFORMATION, APPARATUS, METHOD, PROCESS, OR SIMILAR ITEM DISCLOSED IN THIS DOCUMENT.

REFERENCE HEREIN TO ANY SPECIFIC COMMERCIAL PRODUCT, PROCESS, OR SERVICE BY ITS TRADE NAME, TRADEMARK, MANUFACTURER, OR OTHERWISE, DOES NOT NECESSARILY CONSTITUTE OR IMPLY ITS ENDORSEMENT, RECOMMENDATION, OR FAVORING BY EPRI.

EPRI PREPARED THIS REPORT.

About EPRI

Founded in 1972, EPRI is the world's preeminent independent, nonprofit energy research and development organization, with offices around the world. EPRI's trusted experts collaborate with more than 450 companies in 45 countries, driving innovation to ensure the public has clean, safe, reliable, and affordable access to electricity across the globe. Together...Shaping the future of energy.®

EPRI CONTACTS

ANUJA RATNAYAKE,, *Technical Executive*
aratnayake@epri.com

IRENE DANTI LOPEZ, *WS1 Co-Lead*
idantilopez@epri.com

BASKAR VAIRAMOHAN, *WS1 Co-Lead*
bvairamohan@epri.com

EAMONN LANNOYE, *Senior Area Lead*
elannoye@epri.com

For more information, contact:

EPRI Customer Assistance Center
800.313.3774 • askepri@epri.com

3002031504 **June 2025**

EPRI
3420 Hillview Avenue, Palo Alto, California 94304-1338 USA • 650.855.2121 • www.epri.com

© 2025 Electric Power Research Institute (EPRI), Inc. All rights reserved. Electric Power Research Institute, EPRI, and TOGETHER...SHAPING THE FUTURE OF ENERGY are registered marks of the Electric Power Research Institute, Inc. in the U.S. and worldwide.