INTELLIGENCE UPDATE

Grid flexibility classification gives structure to demand response

Dr. Tomas Rahkonen 23 Jun 2026

In several regions around the world, data center loads represent the fastest-growing source of electricity demand in absolute terms. The rapid pace of data center capacity growth is increasing the strain on electricity grids, particularly during periods of peak demand, while IT load volatility raises additional reliability and power quality concerns. These challenges are being amplified by the growth of AI training supercomputers. In the US, more than half of the forecasted increase in peak load is attributed to data centers, according to calculations by power sector consultancy Grid Strategies, which estimates that an additional 90 GW of peak grid power will be required. In areas with high concentrations of data centers, such as Northern Virginia, this share could approach 100%.

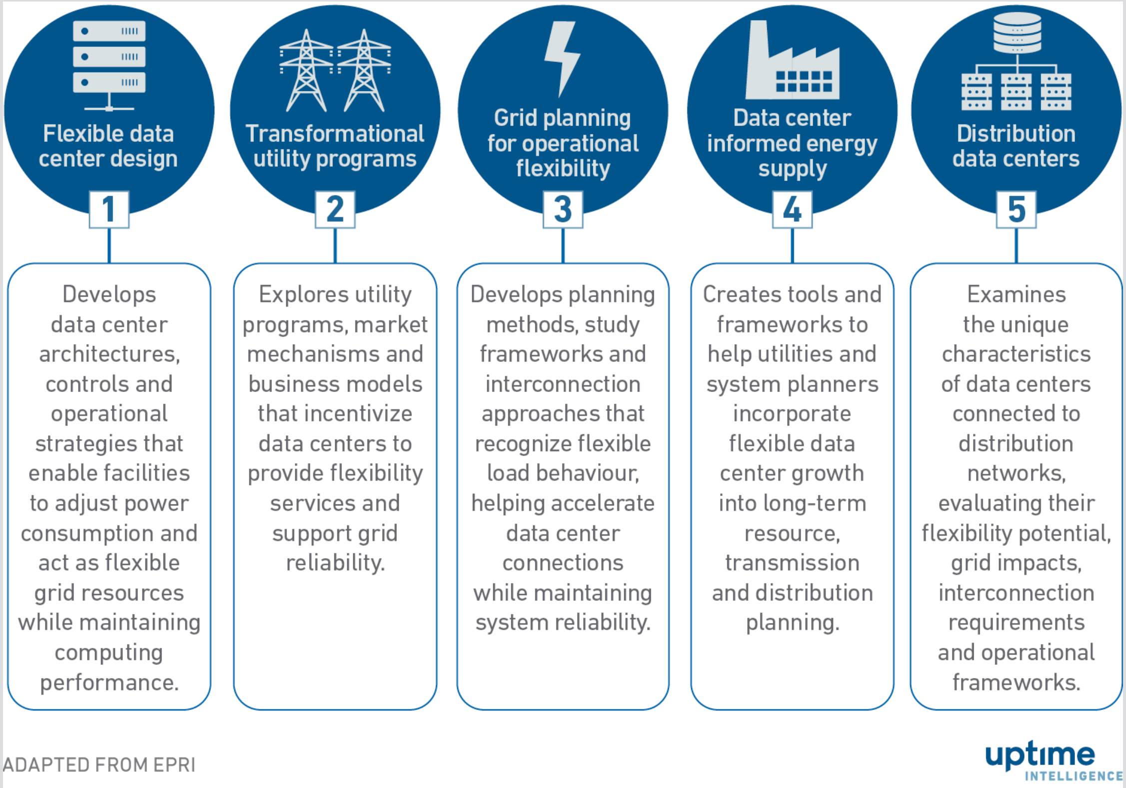
At the same time, grid operators are seeking new sources of flexibility to maintain reliability, integrate increasing amounts of renewable energy, and make more efficient use of existing infrastructure. While current interconnection and planning processes often assume that large loads operate near their maximum demand, some data centers may be capable of providing varying degrees of flexibility. Recognizing and standardizing these capabilities could help accelerate grid connections without compromising reliability or increasing costs, supporting the continued growth of both digital infrastructure and the electric grid (see *Is demand response a viable accelerator for grid interconnects?*).

The Electric Power Research Institute (EPRI), a US non-profit organization, launched the Data Center Flexible Load Initiative (DCFlex) in 2024 to help electrical energy suppliers and grid operators address the interconnection, reliability and demand management challenges associated with the recent surge in large load data center projects. The initiative has more than 60 participants from the US and Europe, including data center operators and infrastructure suppliers, utilities and transmission system operators. EPRI recently announced a collaboration with the Open Compute Project (OCP) to help advance the project and broaden participation among data center operators, including colocation providers.

EPRI has categorized data center work

EPRI is a global research institute that collaborates with more than 450 organizations across 45 countries — including utilities, governments and academic institutions — to tackle complex energy challenges. It conducts research and development to find solutions to issues affecting the generation, decarbonization, delivery and use of electricity. EPRI organizes its data center work into five workstreams for its DCFlex initiative (see **Figure 1**).

Figure 1 The five DCFlex workstreams



Flex Mosaic demand response categories

In March 2026, the DCFlex working group produced a framework to establish common terminology and standardize approaches to demand response services (see *Is demand response a viable accelerator grid interconnects?*). Called the Flex Mosaic, the framework presents five service classes (A to E) in increasing order of grid flexibility (see **Figure 2**).

The technology-neutral framework aims to make grid flexibility easier to define and apply by clearly detailing requirements for each service level. This can streamline interconnection studies and support faster, more informed siting decisions by aligning flexible loads with available capacity. Clear service categories can also help developers and operators design projects more effectively, reducing risk and delays. The framework is being developed and refined iteratively. Two sprints have been completed so far, and a third has recently begun.

Figure 2 EPRI's Flex Mosaic classification framework for grid flexibility

	CLASS A	CLASS B	CLASS C	CLASS D	CLASS E
	Critical peaking	Peaking	Prolonged	Fast	Fully grid responsive
Notification window	Day ahead	Day ahead	Day ahead	>5 min	>5 min
Activation period	<5 hours	<5 hours	>24 hours	<2 hours	>24 hours
Annual utilization	~1%	~>5%	<10%	>8%	30%
	Responds to rare scarcity events in 5 hours or less	Respond to frequent scarcity events in 5 hours or less	A&B, plus respond to prolonged events of 24 hours or more	A&B, plus provide fast response with short notice	Fully grid responsive

ADAPTED FROM EPRI

The framework defines flexibility in terms of performance characteristics, which currently include the following (shown in **Figure 2**):

- **Notification window**: the minimum time before activation that a flexibility service must be called upon, ranging from 1 day ahead to 5 minutes.
- **Activation period**: the duration for which an activated flexibility service can be provided, ranging from up to 2 hours to more than 24 hours. For the latter (24+ hours), activation periods ranging from 8 hours to 2 days have been discussed. This may be clarified further in the current sprint, which aims to refine and finalize the framework.
- **Annual utilization**: indicates how much a flexibility service is utilized, expressed as a percentage of yearly hours.

Several other performance characteristics are being discussed for Flex Mosaic, although limited guidance is currently available. These include depth-of-load adjustment, ramp behaviour and flexibility in service availability. More information may emerge from upcoming framework development sprints.

Flex Mosaic defines three areas in which data centers can provide flexibility:

- **Computational flexibility**: the ability of a data center to temporarily reduce IT load. This can involve shifting the IT load to other data centers, a practice currently limited to advanced IT operators, including hyperscalers. In March 2026, Google announced that it had integrated a total of 1 GW of demand response capacity (by limiting or shifting workloads) into its long-term energy contracts with multiple utilities across the US. However, the terms of the agreements — including the response duration, the committed demand at individual sites, and whether participation is voluntary or mandatory — has not been disclosed. Because colocation data centers do not control IT, any use of computational flexibility would need to be covered in service level agreements (SLAs) with their tenants.
- **Auxiliary flexibility**: the ability to reduce the power used for cooling and other auxiliary systems. For most colocation operators (and enterprise internal facility teams), cooling temperatures are governed by SLAs that specify fixed IT supply temperatures, limiting flexibility to reduce cooling power use. Notably, ASHRAE standards (used to define IT warranties) allow for temperature excursions that could support auxiliary flexibility if

permitted by SLAs.
Hyperscalers that design their own IT hardware sometimes allow supply temperatures to drift to maximize the use of free cooling while accounting for the effects of premature IT equipment aging — an approach that could also enhance flexibility.
Demand response events are commonly triggered during periods of cold or hot weather. During cold-weather events, data centers are typically cooled using economizers (free-cooling systems), leaving little or no mechanical cooling load to reduce. During hot-weather events, mechanical cooling systems often run close to full capacity, meaning that reducing cooling system output can (within minutes) cause IT systems to slow or shut down. In general, there is very little available demand reduction in the facilities area.

- **Alternative power generation and storage**: this includes backup engine generators, battery energy storage systems (BESS), on-site or shared gas turbines for primary power, thermal generators and storage, and virtual power plants.

A complication for colocation operators across all Flex Mosaic classes is that flexibility and service utilization requirements must be considered in tenant agreements, including notification provisions that allow tenants to plan their IT operations accordingly.

DCFlex provides a basic demonstration tool on its website to illustrate the impact of different flexibility strategies. For example, a data center designed with either 20% computational flexibility or 20% generation and storage flexibility can achieve Class A status. Auxiliary flexibility, as expected, has less leverage; according to the tool, 80% of the load is required to be available for demand response to achieve Class A.

The following sections take a closer look at the Flex Mosaic classes and discuss the technical solutions.

Class A (Critical Peaking) and Class B (Peaking)

Class A requires data centers to participate in grid power reserve schemes and occasionally take capacity offline for up to 5 hours with a day-ahead notice. Annual utilization of data center flexibility is limited to rare grid power scarcity events (about 1% annually, corresponding to 88 hours per year). Class B is the same as Class A but is designed for more frequent scarcity events, resulting in higher data center flexibility utilization of up to 5% annually.

One approach proposed by some hyperscalers (chiefly Google) to meet Class A and B requirements is to deploy computational flexibility to manage part of the data center IT load. The day-ahead notice allows time for IT workload planning.

Batteries can provide backup power for up to 5 hours, especially for smaller facilities and where only partial grid disconnection is required. A key limiting factor is capital cost, given the limited revenues from an infrequently used flexibility service (up to 5% of annual hours for Class B). Other limiting factors include the availability of on-site space and battery recharging times, which can affect how frequently the service can be used.

Given the infrequent use of data center flexibility, relying on backup generators can be a realistic option for data center operators to meet Class A requirements in some geographies. One limiting factor is the number of annual engine-generator operating hours allowed under environmental permits. To operate generators as a Class B service, they will need to be permitted for extended operating hours (438 hours or more per year), which typically requires Tier IV diesel generators (with particle matter and nitrogen oxide emissions controls) or natural gas engines or turbines equipped with nitrogen oxide emissions controls. This can make the use of backup generators more challenging, as environmental permits must allow around 438 operational hours (5% of the year). Tier IV generators also have capital and operating costs that are 25-50% higher than those of standard generators.

Class C (Prolonged)

Class C requires data centers to (in addition to meeting Class A and B requirements) go offline for prolonged periods that exceed 24 hours, with a day-ahead notice. Annual data center flexibility utilization can reach 10%, corresponding to 876 hours per year.

The Class C requirement is particularly relevant in Ireland and Texas, where new data centers must bring their own power generation under regulations introduced to address the sector's high electricity use, which can account for about 20% of total grid demand.

Computational flexibility may still be an option for periods exceeding 24 hours for certain workloads (mainly hyperscalers) that can be throttled or shifted to other data centers for extended durations. Current battery systems are generally insufficient to provide power for the prolonged period required for Class C.

With high flexibility utilization of up to 876 hours per year, diesel generators become a less attractive — or in some cases unrealistic — option for meeting Class C requirements in many markets due to environmental permitting constraints and operational costs. For new sites, on-site primary power (such as gas engines or turbines) can be deployed, making compliance with Class C requirements more achievable.

Class D (Fast)

In addition to meeting Class A and B requirements, Class D requires data centers to go offline for durations up to 2 hours with short notice (5 minutes). Annual data center flexibility utilization is up to 8%, corresponding to 701 hours per year.

Adding relatively frequent service activations with a short duration of up to 2 hours (to Class A and B requirements) makes batteries an attractive option. However, given the frequency of service activations, battery recharging times will need to be factored into service availability.

Computational flexibility for a part of the data center IT load may be an option for some advanced IT operators (primarily hyperscalers) that directly manage workloads capable of responding to a 5-minute notice period. However, high flexibility utilization of up to 701 hours per year makes engine generators unrealistic in many geographies due to environmental permitting constraints and operational costs. For new sites, on-site or shared primary power (such as gas engines or turbines) can be deployed, making compliance with Class D requirements more feasible.

Class E (Fully Grid Responsive)

Class E requires data centers to go offline for prolonged periods exceeding 24 hours with short notice (5 minutes). Annual data center flexibility service utilization exceeds 30%.

To meet Class E requirements, data centers will need dedicated or shared primary power generation, such as gas turbines. Such projects have already been deployed, or are under development, in regions with congested power grids, including Ireland and parts of the US, particularly Texas. Another facility type that can operate in a similar manner is bitcoin mining, which uses computational flexibility to operate only during hours when mining is profitable.

For very large new data centers, such as gigawatt-scale AI data centers that account for a significant portion of overall grid power consumption, implementing Class E may be a practical approach, as discussed in *Is demand response a viable accelerator for grid interconnects?*.

> ## The Uptime Intelligence View
>
> Most data center operators prefer to minimize risk by securing firm grid power from day one, sufficient to meet current needs and foreseeable expansions. However, where flexibility (including new primary generation) can provide faster speed-to-power, some operators are likely to invest in the related infrastructure to compete for high-growth workloads such as AI. Another potential driver of flexibility services is the introduction of use-it-or-lose-it regulations, which penalize passive behaviour (i.e., holding unused capacity) and reward active, flexible electricity consumption.
>
> Voluntary demand response programs have long been provided by grid regions/utilities across Europe and North America. In stressed grid regions, the trend is toward making participation mandatory for large loads. Texas has already legislated mandatory participation in demand response programs for large loads of 75 MW or more, and other jurisdictions are likely to follow. Where feasible or required, data center operators need to design demand response capabilities into their new facilities. For large data centers, standby generation systems that are designed and permitted for extended operation are likely to offer the best approach.
>
> EPRI's Flex Mosaic framework can help standardize flexibility markets, starting with large hyperscaler and colocation projects in North America. However, for data center operators that do not control the IT (e.g., colocation providers), tenant requirements and service-level agreements can add significant complications when implementing flexibility services.

Other related reports published by Uptime Institute include:
NERC alert points to future of grid
Are data centers to blame for power quality issues?
Electrical considerations with large AI compute

ABOUT THE AUTHOR

Dr. Tomas Rahkonen

24 Jun 2026

Dr. Rahkonen is the Research Director Sustainability, Europe at Uptime Institute. Rahkonen has spent the last 25 years in positions within the telecommunications, mobile communications, and data center sectors globally, and most recently served as the CTO of Flexenclosure, where he managed the design and delivery of prefab data centers across four continents.

TRahkonen@uptimeinstitute.com

About Uptime Institute

Uptime Institute is the Global Digital Infrastructure Authority. With over 4,000 awards issued in over 122 countries around the globe, and over 1,100 currently active projects in 80+ countries, Uptime has helped tens of thousands of companies optimize critical IT assets while managing costs, resources, and efficiency. For over 30 years, the company has established industry-leading benchmarks for data center performance, resilience, sustainability, and efficiency, which provide customers assurance that their digital infrastructure can perform across a wide array of operating conditions at a level consistent with their individual business needs. Uptime's Tier Standard is the IT industry's most trusted and adopted global standard for the design, construction, and operation of data centers.

Offerings include the organization's Tier Standard and Certifications, Management & Operations reviews and assessments including SCIRA-FSI financial sector risk assessment, the Sustainability Assessment, and a broad range of additional risk management, performance, availability, and related offerings. Uptime Education training programs have been successfully completed by over 100,000 data center professionals, such as the much-valued ATD (Accredited Tier Designer) and AOS (Accredited Operations Specialist). The Uptime Education curriculum has been expanded by the acquisition of CNet Training Ltd. In 2023.

Uptime Institute is headquartered in New York, NY, with offices in London, Sao Paulo, Dubai, Riyadh, and Singapore, and full-time Uptime professionals based in over thirty-four countries around the world.

For more information, visit www.uptimeinstitute.com