

Data centres for grids: the potential impact of flexibility

Data centres coming online have the potential to utilise flexibility measures to reduce power grid stress events and new grid investments.

Published date: 29 October 2026

Authors: Elisabeth Cremona

About

This report explores how much data centres could alleviate stress on power grids during peak events through two mechanisms: shifting consumption to other hours/locations or switching to onsite generators. With limited data on data centres, the analysis draws on assumptions from available literature. An interactive calculator lets readers test these ranges.

Summary

Data centre capacity is set to at least double in most regions by 2030, but capacity alone does not determine grid stress. What matters is how much data centre load will be on the grid during peak demand, and how much of that load can be reduced when the grid needs relief most.

The two main ways data centres can ease that pressure deliver different results. Shifting consumption to other hours or location – the flexibility mechanism most often discussed – appears to have limited potential because most data centre workloads cannot easily be disrupted. Switching to onsite generation – known as bring-your-own-capacity (BYOC) – offers far greater grid relief, in some cases – but whether it benefits the wider system depends on how it is regulated.

01 **Data centres could account for 25% of peak demand by 2030 in certain markets**

In Ireland, an established data centre hub, they already accounted for 19% of peak demand, a share expected to rise to 24% by 2030. In Malaysia, an emerging market, their share could rise from just 2% today to as high as 25% by 2030.

02 **Only 3% of data centre capacity is estimated to be available for load shifting in 2030**

Around 87% of capacity will be used to run workloads such as cloud services and AI inference, which are highly latency sensitive. As a result, load shifting is estimated to cut peak demand by just 1.5% in Ireland and Malaysia and between 0.1–0.6% in the other seven examined markets.

03

Onsite generation could offer nearly five times more grid relief than load shifting – but this flexibility mechanism also brings risks

A Texas case study in this analysis finds that two large data centres switching to onsite generation could take up to 2.7 GW off the grid during a stress event, compared with 0.55 GW through load shifting. Without effective regulation, this risks fossil fuel lock-in; well regulated, with effective regulation, it could reduce the need to build new firm capacity on the grid.

The two mechanisms are not mutually exclusive: they solve different problems and can work together. As data centre demand grows, regulators and grid planners need to look beyond the size of the load and understand what kind of work is running on it. The type of workload will determine which forms of flexibility are possible and how data centres can best support the grid.

As data centre demand grows, flexibility is often presented as the answer. But the majority of data centres are unlikely to be able to shift their consumption. Switching to clean onsite generation is a far more powerful lever, easing grid pressure without adding to consumer bills. That makes what regulation allows to be built behind-the-meter crucial.

Elisabeth Cremona

Energy Infrastructures Lead, Ember



Rapid data centre growth is driving flexibility – but load shifting falls short without clean onsite generation

Data centres are being built faster than the grid infrastructure needed to serve them.

A large data centre typically draws 200–500 MW of power from the grid, equivalent to 200,000–500,000 households. This represents a huge new load that can come online much faster than the grid can expand. A new data centre can be built between [18–24 months](#) while traditional grid expansion takes [4 to 10 years](#). This mismatch can exacerbate grid constraints and prompt the need for costly upgrades, particularly because data centres tend to [cluster in specific regions](#).

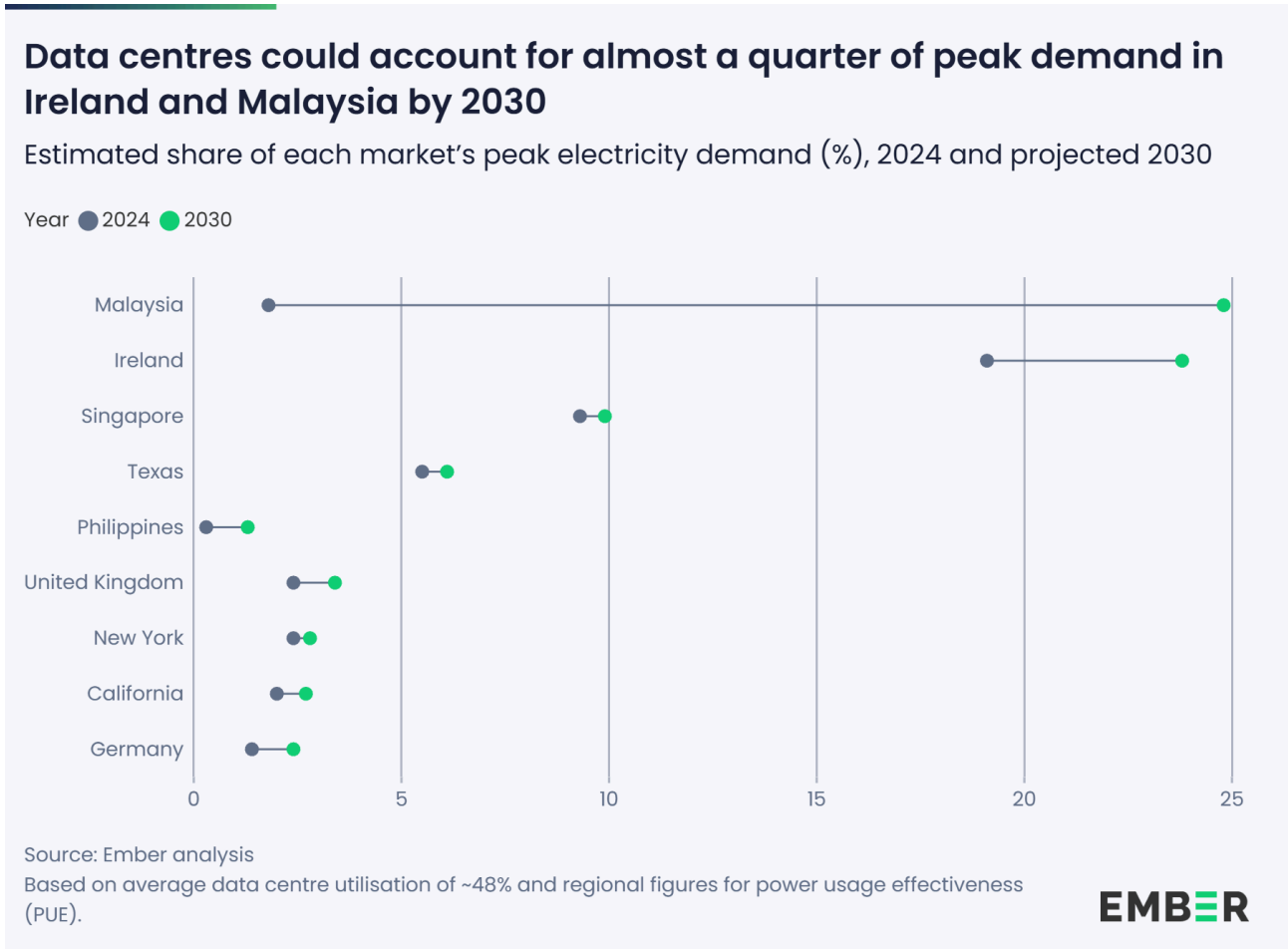
Thrust into the spotlight, data centre developers are [increasingly championing](#) data centre flexibility – adjusting how much power they draw from the grid when it is under stress – as a solution to these challenges. This chapter sets out how much of peak demand data centres could account for by 2030, and the two ways they could reduce it.

1.1 Data centres could account for 25% of peak demand in some markets by 2030

Data centre capacity is set to [at least double](#) in most regions between 2025 and 2030. But capacity alone does not tell us what this means for the grid. Grids are built to handle their single highest-demand hour, not the average one. What matters, therefore, is how much of this growth contributes to peak demand.

This analysis examines nine markets across the US, Europe and ASEAN: Texas, California and New York in the US; Ireland, the UK and Germany in Europe; and Malaysia, the Philippines and Singapore in ASEAN. For each market, we estimate the share of peak electricity demand that is expected to come from data centres by 2030.

We compare projected data centre demand with each market’s peak demand, assuming data centres operate as an inflexible load. This establishes a baseline for measuring how much data centre flexibility could alleviate grid stress during times of peak demand.



The estimated share of data centre demand during peak hours varies considerably across the examined markets. In Ireland, a legacy data centre hub, data centres already constitute an estimated 19% of peak demand. When demand hit a record [6 GW in January 2025](#), data centres likely accounted for 1.1

GW of that total. With data centre demand expected to grow faster than overall demand, their share is projected to reach 24% by 2030, with demand of 1.6 GW against a system peak of 6.5 GW. Singapore and Texas show a similar pattern on a smaller scale, with data centres at an estimated 9.5% and 6% of peak demand, respectively. These shares are expected to remain steady through 2030.

Malaysia, an emerging data centre market, tells a different story. Its 0.5 GW data centre fleet is projected to reach [1.1 GW or 7.7 GW](#) or even [13 GW](#) by 2030, reflecting the [uncertainty of a new market](#). At the upper end of these projections, data centres could account for 25% of Malaysia's peak demand by 2030, up from just 2% in 2024.

1.2 Data centre flexibility can ease pressure on the grid in two ways

Data centre flexibility covers [various ways](#) a data centre can interact with the power system. This report focuses on how data centres can [reduce demand](#) on the grid when it is under greatest stress. There are two ways to do this: load shifting, where a data centre reschedules or relocates its computing workloads during periods of grid stress, and switching to onsite generation also known as bring-your-own-capacity (BYOC), where it continues operating but draws on its own generation or storage instead of the grid.

This report estimates how much each mechanism could reduce peak demand across the nine markets. Given the limited data available on data centre flexibility, the results should be treated as estimates of potential impact rather than definitive forecasts. The analysis draws on available literature and our own assumptions, reflecting uncertainty around data centre workloads, utilisation and future capacity. Readers can use [the interactive calculator](#) to explore how the impact changes under different assumptions.

Only 3% of data centre capacity is estimated to be available for load shifting

Load shifting can help data centres [tap existing headroom](#) on the grid, allowing new capacity to connect faster and at lower system cost by [reducing the need](#) for new generation and transmission infrastructure. But these benefits depend on scale: [one study](#) estimates that 21–49% of data centres would need to act flexibly. This chapter shows how far the 2030 data centre fleet is likely to fall short of that threshold.

2.1 Most data centre workloads cannot easily be shifted

The potential for data centres to participate in load shifting is heavily dependent on the workloads they run. There is also [considerable uncertainty](#) around the potential depth of this flexibility, particularly because data centre flexibility has so far been demonstrated largely through a small number of [pilot projects](#).

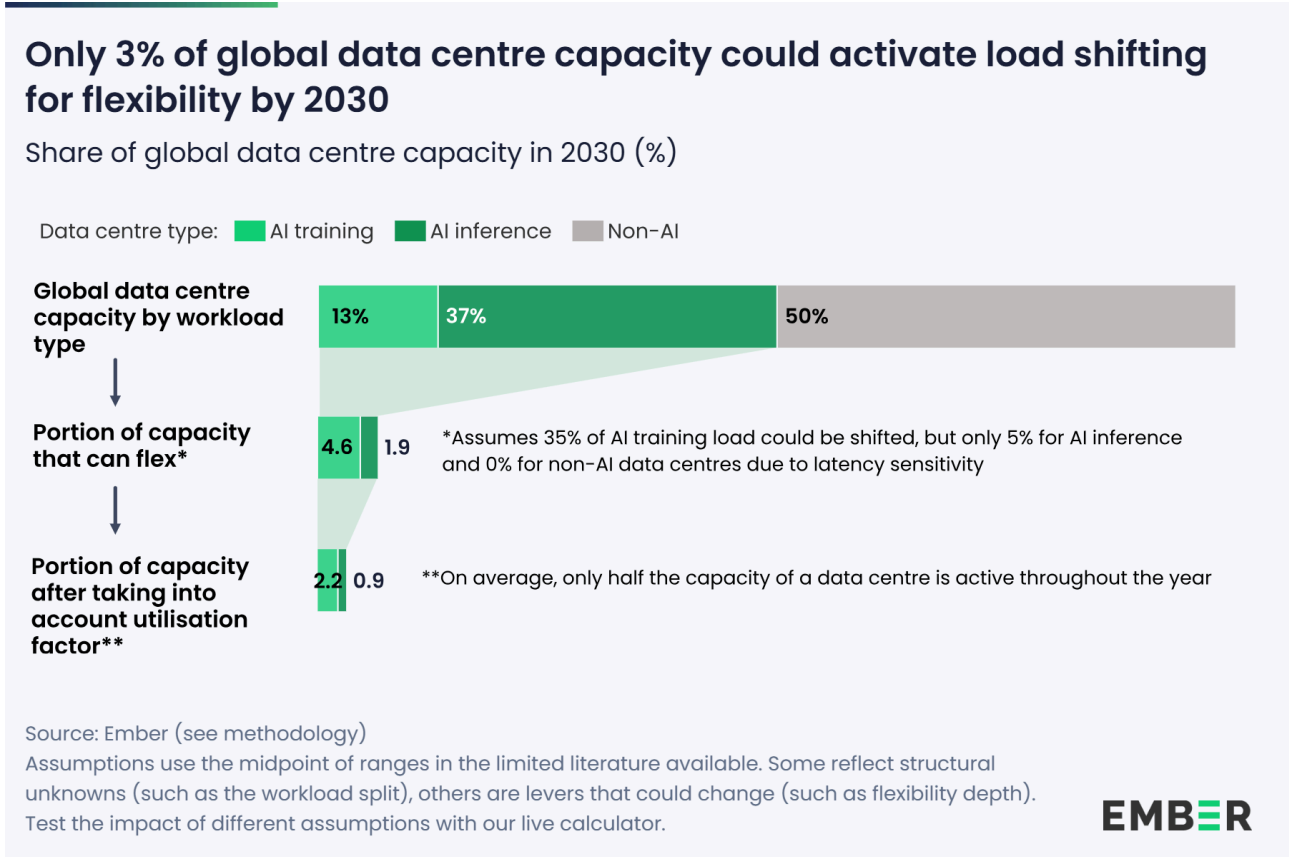
Traditional data centres, which run cloud computing, enterprise IT and web hosting, need to respond to users instantly, which limits their ability to operate flexibly. AI inference data centres, which run trained AI models to answer user requests, are similarly sensitive to delay and unlikely to cut demand without significant [financial or reputational risk](#). Together, these two categories made up an estimated 86% of data centre capacity in 2024. Despite AI's rapid growth, they are still expected to account for 87% in 2030, as inference rather than training drives most of the expansion.

AI training data centres make up the remaining 13% and have the greatest potential for flexibility, typically 25–55% of their power demand, because training can be paused without major consequences. Real-world demonstrations by

Emerald AI, part of Electric Power Research Institute’s (EPRI) [DCFlex](#) initiative, have cut power consumption by up to 40% within a minute and sustained the reduction for several hours [without compromising](#) the highest-priority jobs.

To estimate the share of capacity available for load shifting, we apply different assumptions to each workload: 35% of AI training capacity can flex, 5% of AI inference capacity and none of traditional data centre capacity. These are modelling assumptions informed by the available evidence (see Methodology), not observed participation rates. Applied to the projected 2030 workload mix, they suggest that 6.5% of global data centre capacity could theoretically be available for load shifting.

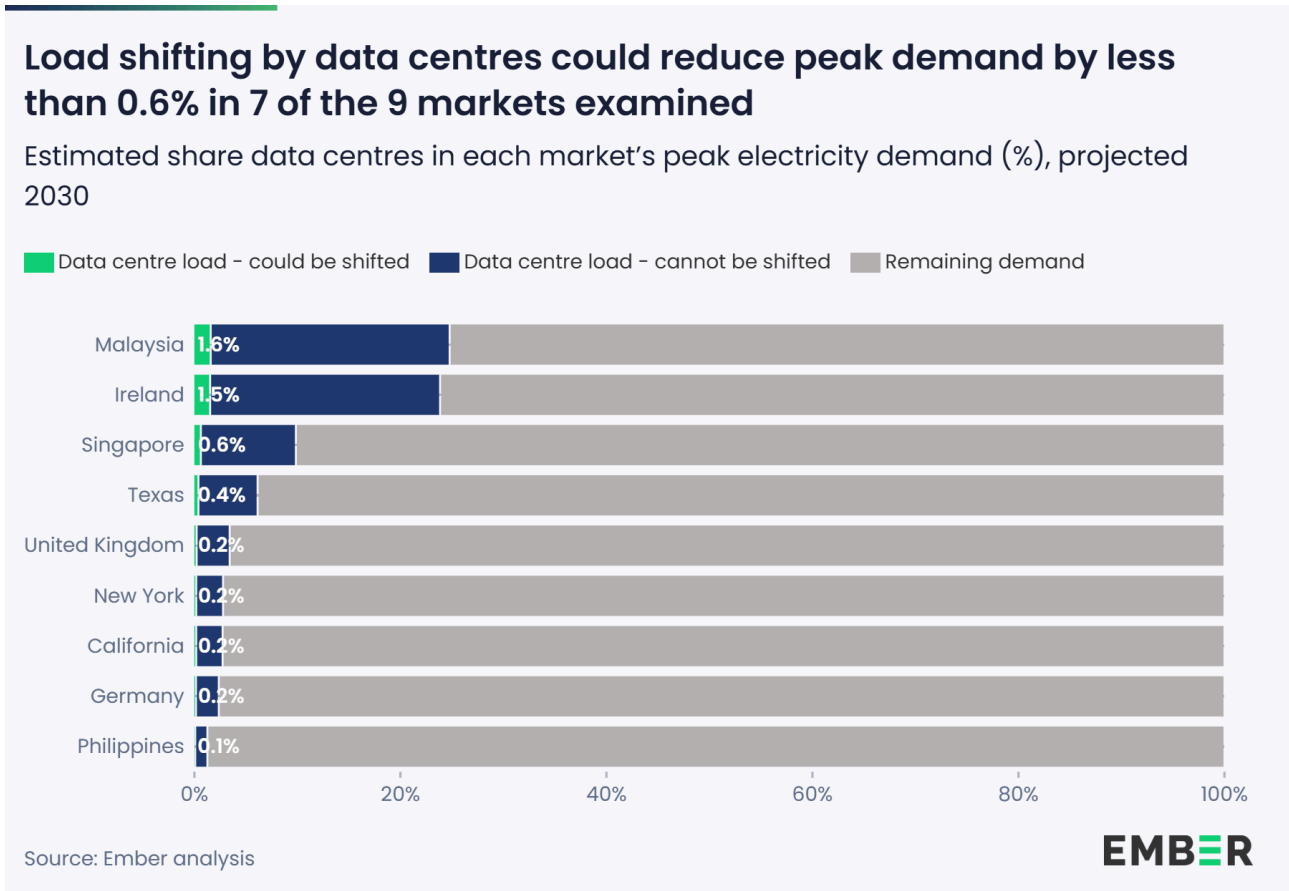
However, data centres do not run at full capacity all the time. Adjusting for an average utilisation rate of about 48%, and assuming this rate holds when the grid needs flexibility, the share of capacity estimated to be available for load shifting falls to 3%.



This estimate is sensitive to the assumptions used, particularly the share of AI training and inference capacity that can flex and the utilisation rate during periods of grid stress. It should therefore be interpreted as an indicative estimate, rather than a direct measure of the flexibility that will be available in 2030.

2.2 Despite their large load, data centres could reduce peak demand by less than 0.6% in seven of the nine markets examined

Although data centres could account for a substantial share of peak demand in some markets, only a portion of their capacity is estimated to be available for load shifting. Based on the assumptions outlined in Section 2.1, this translates into a relatively modest reduction in peak demand across most of the nine markets examined. The limit is structural: most of the fleet runs traditional or AI inference workloads that cannot easily be paused, while flexible AI training makes up only a small share of capacity.



In Ireland and Malaysia, where data centres could reach around a quarter of peak demand by 2030, load shifting is estimated to reduce peak demand by 100 MW and 523 MW respectively, equivalent to about 1.5% of projected peak demand in each market. In the other seven markets, the reduction falls between just 0.1–0.6%.

While 1.5% reduction may appear small, it is not insignificant particularly when combined with other sources of demand flexibility. In Ireland, grid operator EirGrid [expects](#) household demand flexibility and smart electric vehicle charging together to cut 420 MW from the 2030 peak day. Data centre load shifting could add another 100 MW, bringing the total potential reduction to around 520 MW, covering 8% of the projected 2030 peak demand. Yet EirGrid does [not yet factor](#) data centre flexibility into system planning, which risks building more infrastructure than needed.

Although load shifting at scale needs technical and commercial arrangements that are [not yet fully developed](#), some arrangements are already emerging. EirGrid has confirmed that a small number of data centre sites can be called on to [reduce demand](#) to prevent system alerts. In Texas, [Senate Bill 6](#) requires large loads of 75 MW or more that connect after 31 December 2025 and have onsite backup generation to be curtailable during a grid emergency.

Our [live calculator](#) tool allows users to explore how different assumptions change the potential impact of load shifting from data centres on peak demand. Some of these assumptions relate to structural unknowns – such as the split of data centre workload types – while others represent levers that could be changed – such as the flexibility depth of different data centres.

Onsite generation offers more grid relief – but its value depends on what is built

Load shifting is limited because most data centre workloads cannot easily be moved to different times. An alternative is to keep workloads running while switching their electricity source from the grid to the data centre's own generation or storage, where these can be built. Many developers are already pursuing this option in response to multi-year waiting times for a grid connection.

Nowhere is this trend more visible than in Texas. The state is set to [overtake Virginia](#) as the largest data centre market in the US. Meanwhile, its grid operator, ERCOT, faces almost [400 GW](#) of connection requests from data centres, nearly five times the state's record peak demand of 85.5 GW. In response to this queue, developers have announced [40 GW](#) of onsite generation for data centres in Texas.

Onsite generation could take significant pressure off the grid, but the benefits and risks depend on what is built behind-the-meter. This chapter uses two Texas data centres to estimate how much grid relief onsite generation could offer compared with load shifting, then looks at what would be needed to deliver that relief from clean sources.

3.1 Switching to onsite generation would provide almost five times more grid relief than load shifting

The two data centres, [Stargate Abilene](#) and [Comanche Circle](#), are both expected to have an ERCOT grid connection and onsite generation. Together they represent about 5.6 GW of new load due online by 2030. At the typical US data centre utilisation rate of 48%, they would draw about 2.7 GW on average and likely more, as AI-specific facilities tend to run at higher utilisation.

Using the same assumptions as Section 2.1, load shifting at these facilities could relieve an estimated 0.55 GW during a grid stress event. Switching to onsite generation could reduce grid demand up to 2.7 GW, almost five times as much, without changing how the data centres operate. Against ERCOT's projected 2030 peak demand, this is the difference between relieving 0.4% of peak demand and 2%.

From a grid perspective, switching to onsite generation offers greater potential to reduce demand at peak times, although it needs careful coordination. If many facilities switch to batteries or backup generation at the same time during a grid disturbance, the sudden drop in demand could itself [affect system stability](#).

Switching also appears to be the data centre sector's preferred route. In Ireland, the energy regulator [initially proposed](#) mandatory flexibility requirements for data centres. Following public consultation, these were [replaced](#) by requirements to provide generation or storage, now mandatory for data centres of 10 MW and above.

3.2 The risks of onsite generation could be turned into an opportunity for a more cost-efficient power system

Without appropriate regulation, onsite generation carries two risks. The first is fossil lock-in: as the US trend shows, developers seeking reliable power are turning to gas even as grids decarbonise, and uncertainty over future data centre growth adds the risk of [stranded assets](#).

The second is paying twice for the same generation capacity. As power demand grows, grid operators seek to secure additional firm generation capacity to ensure system reliability, often through capacity auctions. Many data centres are seeking a grid connection while also building onsite capacity. If grid operators do not count the onsite generation a data centre already has, they risk contracting additional capacity to cover load that could largely supply itself at peak as two sets of assets would be built for the same load – with the cost of the grid-side generation capacity shared by all consumers, not just the data centre. These are not insignificant costs: in the capacity auction organised by PJM, the largest regional grid operator in the US, for the years 2028–2029, data centre load alone was responsible for [\\$6.3 billion of the \\$16.4 billion](#) in capacity charges.

The US pattern is not inevitable. In Malaysia, developers are pairing new data centres with large clean power projects. DayOne Data Centres has signed agreements for around [1.5 GW of solar and 2.2 GWh of battery storage](#) to supply its Malaysian operations, and is exploring up to 1.5 GW of onsite generation and storage with national utility Tenaga Nasional Berhad (TNB).

Overbuilding can be avoided too. One of the few available studies [modelled six large US data centre sites](#) combining onsite resources with a flexible grid connection. Grid power remained available for more than 99% of hours, with onsite resources needed for only 40–70 hours a year. By using their own capacity only when needed, data centres can reduce the need for new system build and [shift costs](#) of meeting their demand onto themselves onto the data centre, substantially lowering the costs that fall on other consumers. Whether onsite generation benefits or harms the wider system therefore depends on how it is regulated.

Load shifting and onsite generation work best together

Load shifting and onsite generation offer data centres different ways to ease pressure on the grid. Onsite generation offers more relief at peak while load shifting gets [more use out of the existing grid and generation assets](#). This [benefits consumers](#) by avoiding infrastructure that would otherwise be rarely used.

The right mix depends on the workload. Traditional and AI inference data centres, which are expected to make up 87% of capacity in 2030, cannot easily shift demand. For these facilities, clean onsite generation and storage may be the main way to reduce reliance on the grid at peak. AI training data centres have greater potential to shift computing across time or locations.

Rather than seeking a single mechanism that works for every data centre, the aim should be to combine workload flexibility with behind-the-meter flexibility. This also gives developers what they value most: faster access to power, a far stronger [incentive](#) than traditional demand response payments. Onsite assets need to be considered alongside wider grid infrastructure, not simply as a substitute for it and built from clean sources, as some developers in Malaysia are [beginning to show](#).

How much of this flexibility the grid can actually rely on will depend on how regulators design connection rules. Making flexibility and cost responsibility part of grid-connection frameworks for large data centres, with [clear terms](#) on when they can be [curtailed](#), and on who pays for duplicated capacity, would let planners count on data centre flexibility rather than [build around it](#).

Supporting materials

Methodology

This report estimates the flexibility potential of data centres in the US, Europe and ASEAN. It combines projections of data centre capacity with assumptions on workload mix, power usage effectiveness and utilisation, and compares the results against peak electricity demand in each country.

Data centre capacity

Published estimates of existing and projected data centre capacity vary widely, and there is no clear basis for preferring one over another. This report therefore uses a single source per region wherever possible, so that figures are consistent and comparable across geographies. All capacity figures refer to nominal IT capacity. This is the maximum power draw of a facility's IT equipment at full load. It excludes cooling and other supporting infrastructure.

- United States: Electric Power Research Institute (EPRI), [Powering Intelligence 2026](#). We use the “medium growth scenario” which is based on: all projects under construction, 75% of those in advanced planning, and 10% in early planning are fully operational by 2030.
- Europe: European Data Centre Association (EUDCA). [State of European Data Centres 2026](#). One set of projections is provided which combines a bottom-up analysis of announced and probable data centre projects in the EUDCA own European colocation and hyperscale database with a top-down check against international datasets and market expectations.
- ASEAN: no regional dataset was found, so country-level projections are used, taken from government sources wherever possible.
 - Singapore: our figures are based on the [Green Data Centre Roadmap](#) published in 2024 by the Infocomm Media Development Authority (IMDA).

- Malaysia: 2024 data centre capacity from Ember, [From AI to emissions: Aligning ASEAN's digital growth with energy transition goals](#). Projected 2030 capacity is based on [government forecasts of 12.89 GW](#).
- Philippines: Data centre capacity in 2024 and 2030 is based on the [Philippines AI+ Infrastructure Masterplan](#) (PAIIM) 2026–2033 which targets a 30-fold increase in AI data centre capacity from 50 MW to 1.5 GW by 2033, with around 400 MW planned for deployment by 2030.

Converting IT capacity into power demand

Two adjustments convert nominal IT capacity into the power a facility actually draws.

Power usage effectiveness (PUE) is the ratio of total facility power to IT power. It accounts for cooling and other supporting infrastructure. Regional PUE values are taken from [Energy and AI](#), International Energy Agency (IEA).

Utilisation factor reflects that facilities rarely run their IT equipment at full capacity. Estimates in the literature converge at around 50%, including figures from EPRI and the IEA. This report uses the IEA's regionally specific values from [Energy and AI](#), International Energy Agency (IEA). We acknowledge that utilisation is likely to differ between workload types. However, without data to support this, the same utilisation factor has to be applied to all workloads.

Power demand = nominal IT capacity × utilisation factor × PUE

Breakdown by workload type

Capacity is split by workload type rather than by facility type (hyperscale, colocation or enterprise), because workload type is considered to better reflect flexibility potential. Colocation facilities can host a mix of AI and non-AI workloads, and part of that mix could be flexed. Classifying these sites by facility type alone may understate their flexibility potential. Three sources provide a split by workload type: [BloombergNEF](#) (BNEF), [JLL \(2026\)](#) and [McKinsey \(2025\)](#).

Their estimates differ substantially. BNEF's forecast (which is specific to the US) gives AI workloads the smallest share – about 30% by 2030 – diverging from the other two sources – 50% (JLL) or 71% (McKinsey). This report uses the JLL split as a central estimate, as it appears to represent a middle scenario.

The split between workload types is likely to vary by country. Some countries are prioritising AI, while others have a more even mix of AI and non-AI workloads. The balance between AI training and inference also differs. BNEF is the only source with a regional breakdown, but it was not used as it diverges significantly from the other two sources. The global JLL split is applied uniformly across all regions, and the report makes no assumptions about national workload mixes.

It is likely that countries with a larger share of AI training capacity could have greater potential for load shifting, while those dominated by traditional computing and AI inference may have less. The estimates should therefore be interpreted as indicative comparisons, rather than precise forecasts for each market.

Acknowledgements

The author thanks Aditya Lolla, Rashmi Mishra, Alnie Demoral, Ardhi Arsala Rahmani, Chelsea Bruce-Lockhart and Leonard Heberer for their contributions. Thanks also goes to Taylor Gosk for data collection and the brainstorming that shaped this report, and Liam Newcombe, GreenScale Data Centres, for the insights he shared during the research.