

Industrial Flexibility: Turning Industrial Load into Grid Assets

Part 1: The Case for Industrial Flexibility

Ben Mucsi

mucsiben@stanford.edu

Faculty Advisor: Steve Westly

Stanford Graduate School of Business - Independent Study (GSBGEN390)

Spring 2026



Table of Contents

Acknowledgements	3
Executive Summary	4
1. The Grid Capacity Problem	7
1.1 U.S. Demand Growth	7
1.1.1 Data Centers	8
1.1.2 Industrial Electrification	10
1.1.3 Other Demand Drivers	10
1.2 Peak Planning and Grid Infrastructure	10
2. Flexibility as a Solution to Grid Constraints	13
2.1 How Flexibility Works	13
2.2 Sources of Flexibility	15
2.2.1 Data Center Flexibility	15
2.2.2 Aggregated Residential Flexibility	16
2.2.3 Commercial Flexibility	17
2.2.4 Industrial Flexibility	18
2.3 The Case for Industrial Demand Flexibility	18
2.4 Emerging Roles of Industrial Flexibility	20
2.4.1 Accelerate Speed to Power	20
2.4.2 Relieve Near-term Constraints	21
2.4.3 Hedge Uncertain Load Growth	21
2.4.4 Build a More Efficient Grid	21
3. From Context to Action	23
References for Parts 1 and 2	24

Acknowledgements

I conducted this research through an Independent Study under the guidance of Faculty Lecturer Steve Westly at the Stanford Graduate School of Business. I am grateful for the utilities, system operators, regulators, technology vendors, investors, and industrial operators who generously shared their time and perspectives with me. Their generosity, candor, and patience made this project possible.

A special thanks to Stephanie Carstairs (PG&E), Peng Wang (PG&E), Quinn Nakayama (PG&E), Luke Nickerman (PG&E), John Hernandez (PG&E), Maria Pope (PGE), Franco Albi (PGE), Isaac Barrow (PGE), Adam Rogerson (NRCan), Chris Karwick (Silicon Valley Power), Tyler Norris (Google), Daniel Padilla (Emerald AI), Dana Guernsey (Voltus), Matthew Plante (Voltus), Pete Edmunds (Voltus), Nathan Case (Camus Energy), Jigar Shah (Deploy Action), Arnab Pal (Deploy Action), Long Lam (The Brattle Group), Maria Belenky (PJM), Yvonne Huang (NYISO), Raleigh Moore (SPP), Anna Lafoyiannis (EPRI), Irene Lopez (EPRI), Juan Carlos Heras (BCG), Jay Lipman (Resilience), Khaled Dahdouh (Danovo Energy Solutions), Theo Ludwick (NosTerra Venture Capital), Cody Taylor (CEC), Olivia Wang (Sightline Climate), George Georgiev (NextEra), Aaron Panzer (National Grid Partners), Prof. Joshua Rauh (Stanford), Prof. Meagan Mauter (Stanford), Rajanie Prabha (Stanford), Fletcher Chapin (Stanford), David Edimo (Stanford), and Gauraang Biyani (Stanford) for their thoughtful conversations, introductions, and guidance throughout the project. All views, and any errors, are my own.

I am also grateful to the Stanford Climate Ventures community, including the teaching team of David McColl (Echelon), Jane Woodward (WovenEarth Ventures), Joel Moxley (Echelon), Dave Danielson (Breakthrough Energy), and Mayank Giridhar (Moonshot Factory), and my teammates Sam Peinado, Erin Musabandesu, Jason Chen, Carolyn He, Simon Damm, Sophie Tollmann, and Akshay Rao, for giving me the space and encouragement to explore the questions that ultimately led to this research.

Executive Summary

Industrial demand flexibility can help the grid serve more load before new infrastructure arrives, but only when plant-level flexibility is converted into capacity that planners can count. The U.S. power system still needs new generation, transmission, distribution, and storage. It also needs a faster way to use existing infrastructure during the specific hours and at the specific locations where the grid is constrained. This Part 1 report establishes the case for industrial flexibility as one of the most promising underused tools for that near-term capacity problem and argues that it will be critical over the long term. Part 2 turns from context to execution, examining monetization pathways and adoption barriers, while Part 3 sets out the institutional changes required to make industrial flexibility planning-grade.

The central argument is that the grid constraint is increasingly a problem of deliverable peak capacity, not annual electricity. A region can have enough energy for most hours of the year and still be unable to serve a new large load at a particular substation, feeder, or transmission interface during peak conditions. This creates a paradox: the grid is underused on average, yet locally constrained when demand is highest. Flexibility is valuable because it converts time and location into capacity. If a load can reduce, shift, or self-supply during the few constrained hours that drive grid planning, the same physical network can serve more total demand.

Five findings frame the case.

1. Electricity demand growth has broken from two decades of near-flat load. U.S. electricity demand grew by roughly 1.7 percent annually from 2020 through 2025, compared with about 0.1 percent annually from 2005 through 2019 [1]. Five-year peak-demand growth forecasts have risen from 24 GW in the 2022 forecast to 166 GW in the 2025 forecast [2]. NERC projects that summer and winter peak demand across North America will rise by more than 200 GW over the next decade, with 13 of 23 assessment areas facing resource-adequacy challenges [3]. Data centers are the most visible source of this growth, but industrial electrification, advanced manufacturing, electric vehicles, and building electrification are adding pressure in many of the same constrained regions.
2. The binding constraint is usually grid capacity, not aggregate energy. Utilities and system operators plan around the hours in which failure would be unacceptable, namely the hottest summer afternoons, coldest winter mornings, and contingency conditions that follow the loss of a line or transformer. A new large load may trigger an infrastructure upgrade not because the system lacks energy in every hour, but because it cannot deliver firm megawatts at that location during those stressed conditions. A 2025 Duke Nicholas Institute analysis found that the 22 largest U.S. balancing authorities had an average load factor of 53 percent [9]. The same study found that allowing new large loads to curtail for only 0.25 percent of annual hours, roughly 20 hours per year, could unlock 76 GW of headroom [9]. Small, bounded flexibility commitments can therefore create large capacity value when they occur in the right hours and locations.
3. Physical buildout remains essential, but it is too slow to solve the near-term problem on its own. Berkeley Lab reported roughly 10,300 active interconnection requests at the end of 2024, representing about 1,400 GW of generation and 890 GW of storage, while only 13 percent of capacity requesting interconnection from 2000 through 2019 had reached commercial operation by the end of 2024 [11]. DOE reported that newly built high-voltage transmission fell from an average of 2,500 miles per year in 2013 to 2015 to 56 miles in 2023, while major transmission projects often take about a decade to develop

[12]. Flexibility creates an adjustable layer between load growth that is arriving now and infrastructure that may arrive years later.

4. Industrial load is highly relevant because it combines scale, concentration, metering, and contractability. A single industrial facility can place several to tens of megawatts behind one interconnection point. Many facilities also contain sub-loads that can move without shutting down the plant, including pumps, compressors, refrigeration, thermal buffers, storage tanks, batch processes, and on-site generation. FERC reported that industrial customers accounted for 13,840 MW of potential retail peak demand savings in 2023, 45 percent of the national total, and had 75 percent advanced-metering penetration [18]. Industrial electricity sales were about 1,035 TWh in 2024, roughly 26 percent of U.S. retail sales [19]. The industrial base is therefore large enough for small percentage shifts to matter.
5. The core barrier is institutional conversion rather than technical potential. Industrial flexibility is useful to the grid when it is visible, measurable, dispatchable, contractually bounded, and credible enough to affect planning. Many industrial facilities already contain movable load, but the grid can use only the portion that planners can count, operators can dispatch, customers can manage safely, and capital providers can finance. The task is therefore not simply to find theoretical megawatts of flexibility. It is to convert plant-level operating capability into a planning-grade capacity resource.

Industrial flexibility has advantages over other demand-side resources for local capacity relief. Residential virtual power plants can provide broad system value, but their resources are diffuse and harder to target at a specific grid bottleneck. Commercial flexibility is useful, but the megawatt yield per site is usually smaller. Data center flexibility is important, but data centers typically have high opportunity costs of curtailment, and many software-based approaches to flexibility are still early and difficult for planners to credit broadly. Industrial facilities are often easier to contract with, easier to meter, and large enough to affect a local constraint directly. Their flexibility is not universal, but where it exists, it is closer to a planning-grade resource than the public debate often recognizes.

The economic logic is to flex the lowest-opportunity-cost load that can solve the constraint reliably. A data center may have the highest willingness to pay for firm power, especially during high-value compute, but the cheapest capacity relief may come from a nearby industrial process that can shift a few megawatts with limited disruption. In that case, industrial flexibility complements new large-load growth. The data center remains served, the industrial facility is compensated, and the grid avoids treating every megawatt as fully firm in every hour.

This report identifies four emerging roles for industrial flexibility beyond traditional emergency grid relief.

1. Industrial flexibility can accelerate speed to power. New or expanding loads may be able to energize before full grid upgrades are complete if a defined portion of demand can operate conditionally during constrained hours.
2. It can relieve near-term local constraints. When a flexible facility sits inside a constrained electrical pocket, a few megawatts of dependable reduction can create meaningful headroom at the feeder, substation, or transmission interface that is actually binding.
3. It can hedge uncertain load growth. Demand forecasts are moving by gigawatts between planning cycles, while grid infrastructure moves on multi-year timelines. Flexibility gives planners an option while forecasts evolve and physical upgrades lag, but its value should not be treated as temporary. As electrification continues and new constraints

keep emerging, flexible load remains an operating layer for a grid that can absorb more demand while preserving reliability margins.

4. It can improve peak-driven grid utilization. A grid built for rare peaks runs far below its maximum capability in most hours. Controllable demand can help the system serve more useful load through existing infrastructure while preserving reliability margins.

The thesis has limits. Industrial flexibility will not replace generation, transmission, distribution, storage, or other demand-side resources. Some processes are too continuous, safety-critical, or production-sensitive to participate. Some facilities can offer only a narrow flexible band that is economically viable and becomes prohibitively expensive to regularly cut deeper into their load. Some local constraints will require physical upgrades regardless of demand-side capability. Industrial flexibility is most valuable under specific conditions, when the flexible load is electrically relevant to the constraint, when the operating terms are bounded, when performance can be measured, and when compensation exceeds the facility's opportunity cost.

Part 1 therefore makes the case for industrial flexibility as a near-term grid-capacity resource, and a necessary asset in the long-term future of the grid. The United States is short of planning-grade, deliverable capacity in specific places and hours more than it is short of electricity in aggregate. Industrial facilities are one of the largest existing pools of potentially flexible demand on the system, and their scale, concentration, metering, and contractual maturity make them especially relevant to local capacity relief. The next challenge is turning that potential into a usable product. Part 2 addresses that conversion problem, examining how industrial flexibility is monetized today and what barriers prevent adoption. Part 3 sets out what utilities, regulators, industrial operators, investors, and innovators can do to make industrial flexibility a resource the grid can trust.

1. The Grid Capacity Problem

The contemporary U.S. power debate is often compressed into a single phrase: a power shortage. This description is too blunt to guide planning. In many of the places where new data centers, factories, and electrified processes are seeking service, the relevant problem is a shortage of firm, deliverable capacity at particular grid nodes and during the small number of hours that determine system peak, rather than an absence of annual electricity supply in the abstract. A region can have adequate energy over the year and still be unable to connect a new large customer on acceptable timelines if the substation, feeder, transmission path, or local reliability margin is already constrained.

This distinction shapes both diagnosis and remedy. If the country were simply short of electricity in aggregate, the answer would be to add generation and wait for supply to catch up. The actual problem is narrower, as utilities and system operators must absorb rapid load growth while preserving reliability, even though transmission, substation upgrades, and interconnection move far more slowly than commercial demand. The result is a grid-capacity problem rather than a generic energy shortfall.

1.1 U.S. Demand Growth

For most of the 2000s and 2010s, U.S. electricity demand appeared mature. Efficiency gains, slower industrial growth, and the aftereffects of the financial crisis produced a long period of broadly flat national load. The U.S. Energy Information Administration now shows a clear break: demand grew by roughly 1.7 percent annually from 2020 through 2025, against about 0.1 percent annually from 2005 through 2019 [1]. This is a regime change, with new end uses arriving faster than efficiency improvements can offset them.

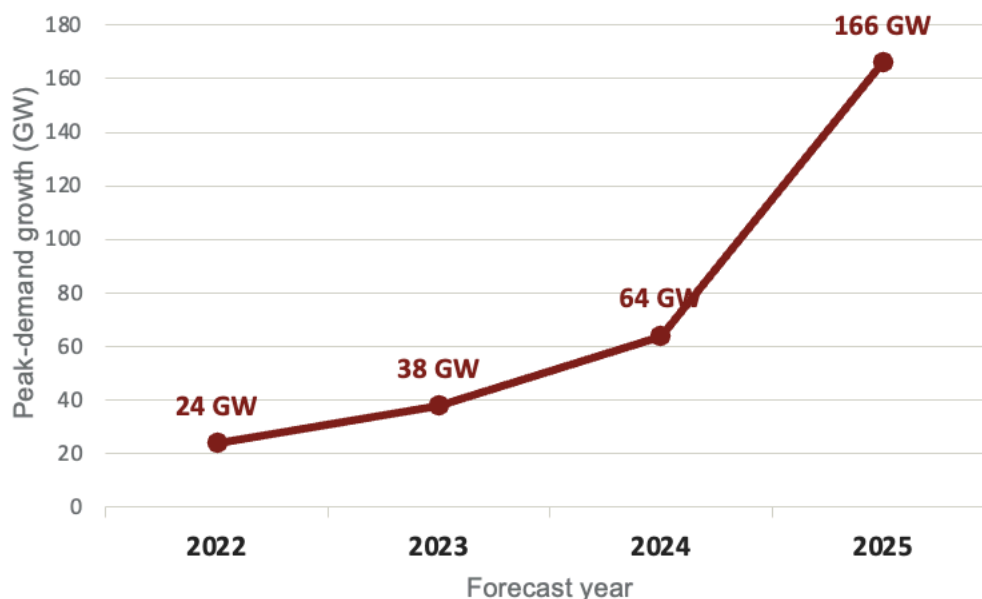


Exhibit 1: The five-year U.S. peak demand growth forecast has risen more than seven-fold in two years, from 24 GW in 2022 to 166 GW in 2025, marking a clear break from two decades of near-flat load [2].

The reliability implications are already visible in planning forecasts. The North American Electric Reliability Corporation (NERC) projects in its 2025 Long-Term Reliability Assessment that summer and winter peak demand across North America will rise by more than 200 GW over the next decade, with 13 of 23 assessment areas facing resource-adequacy challenges [3]. This demand grows in concentrated blocks, tied to specific geographies and short development timelines. Rather than arriving as a smooth national trend line, data centers, advanced manufacturing, electrified industrial processes, electric vehicles, and electrified buildings show up as concentrated additions, which is what makes them difficult for the grid to absorb.

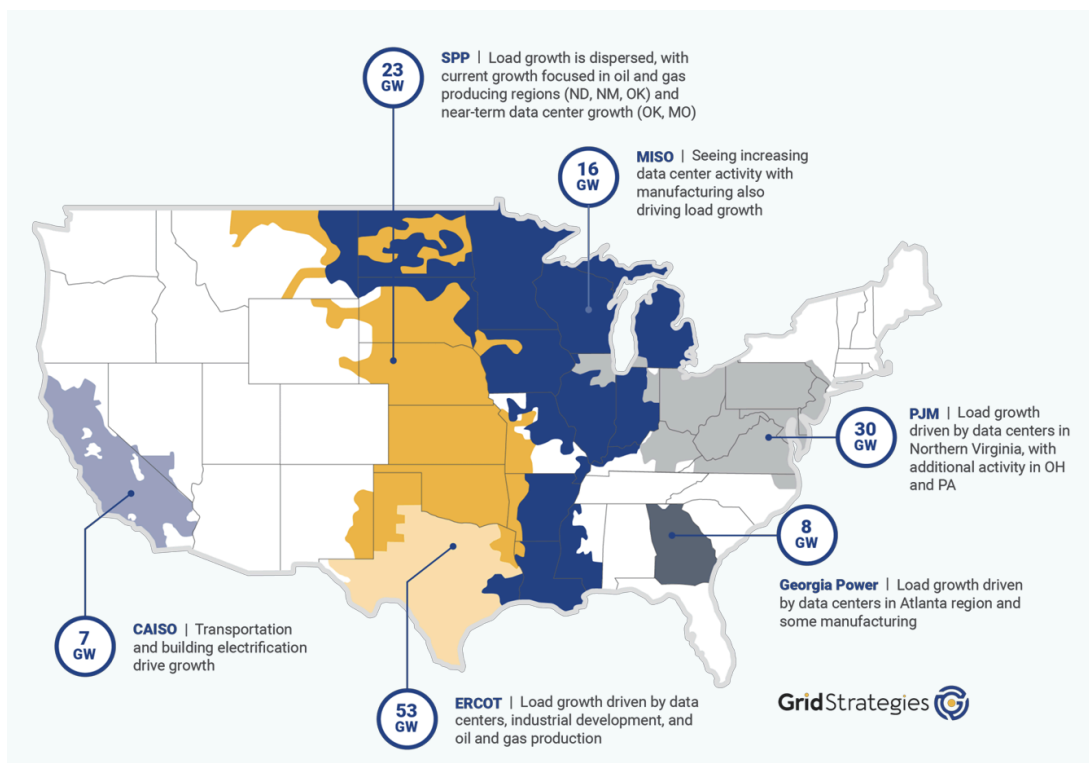
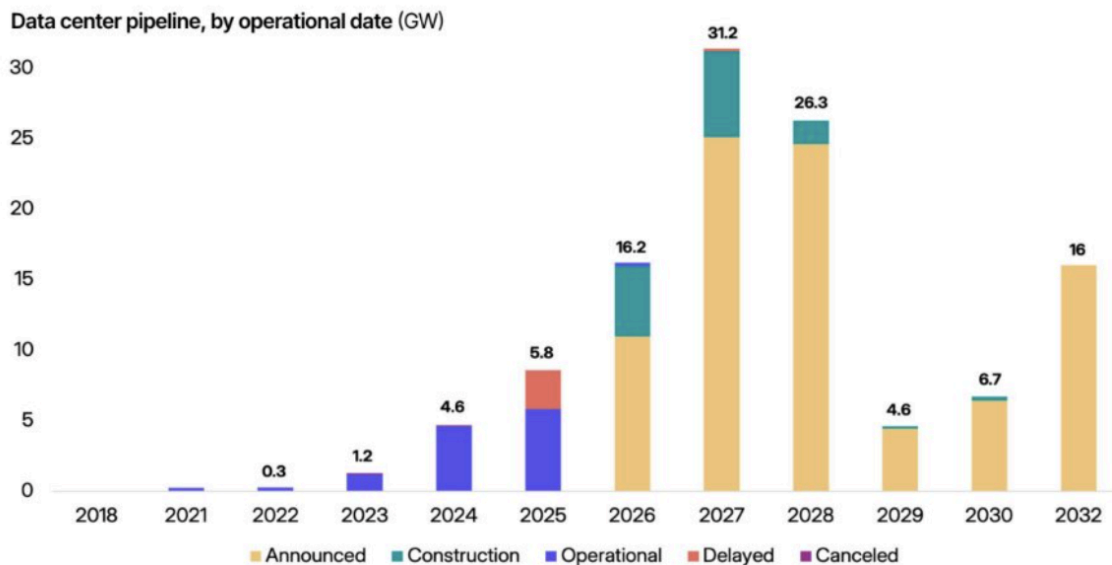


Exhibit 2: Six regions account for almost all near-term U.S. load growth, with the Electric Reliability Council of Texas (ERCOT) at 53 GW and PJM Interconnection (PJM) at 30 GW dominating, illustrating how concentrated the load surge is geographically [2]. The coloring is purely to distinguish regions.

1.1.1 Data Centers

Data centers are the most visible and, in current planning discussions, the most consequential source of incremental load, combining unusually large sites with rapid deployment. The joint Department of Energy (DOE) and Berkeley Lab report on U.S. data center energy use estimates that data centers consumed 176 TWh in 2023, or 4.4 percent of total U.S. electricity. It projects a range of 325 to 580 TWh by 2028, roughly 6.7 to 12 percent of national use [4]. NERC's 2025 assessment states that new data centers account for most of the projected increase in North American demand over the next decade [3]. The effect is even more concentrated in exposed regions: some Western Electricity Coordinating Council (WECC) balancing authorities reported planned data center load equal to as much as 40 percent of forecast demand [3].

Utility and energy-system interviewees emphasized that data center growth is both large and uncertain. California's data center forecast had moved by roughly 1.2 gigawatts in a single year, which is roughly the equivalent of a city.



Source: Sightline Climate

Exhibit 3: U.S. data center project pipeline by expected operational year rises from 4.6 GW in 2024 to a peak of 31.2 GW in 2027, but a growing share of the outer-year pipeline is only Announced rather than under Construction or already Operational, reinforcing the need to haircut headline forecasts for realization risk [5].

Data center forecasts require careful interpretation. Announced projects, utility service requests, contracted capacity, energized capacity, and realized metered load are not the same thing. ERCOT in Texas faces an interconnection queue containing nearly 400 GW of requests. Many of these projects will not materialize, but they still have an impact. Grid planning teams cannot rely on a single long-run forecast. NERC notes that Texas planners are reducing some new data center projections by nearly 50 percent relative to requested load to reflect observed behavior, while the California Independent System Operator (CAISO) uses utilization factors, confidence levels, and ramping assumptions to separate speculative requests from credible ones [3]. Even after planners haircut the most optimistic requests, the projects that do materialize are large enough to stress local infrastructure, especially large training data centers that require GW-scale deployments. It is extremely difficult to make infrastructure commitments worth billions, for decades ahead, amidst this kind of uncertainty.

Over the longer term, data center growth may become more distributed as the sector shifts from a small number of very large training campuses toward a broader mix that includes more inference capacity. A sector expert explained that most future data center demand could fit into smaller, even already existing interconnections, including deployments in existing commercial buildings rather than always requiring new gigawatt-scale campuses. They argued that the projects already completed or under construction would be sufficient to meet the demand for training centers, and this GW-scale frenzy would not last.

This would reduce but not remove the grid challenge. It would spread it across more substations, feeders, and utility territories. The planning problem would shift from whether a single massive site can be served at the transmission level to whether many smaller sites can be connected without creating hidden distribution constraints, cost shifts, or operational risks. That shift reinforces the need for better distribution-level visibility, flexible-service rules, and locationally relevant flexibility resources that can help absorb demand where it actually appears.

1.1.2 Industrial Electrification

Data centers are only part of the story. Industrial electrification is becoming a second major source of demand growth, driven by decarbonization and industrial policy. The U.S. Energy Information Administration (EIA) projected that industrial electricity demand would grow faster than any other major sector, rising 2 percent in 2025 and 3 percent in 2026 as semiconductor and battery manufacturing came online [1]. The longer-run decarbonization agenda for steel, chemicals, cement, and other hard-to-abate sectors points toward greater use of electric motors, furnaces, electrolysis, and other electrically driven equipment.

Industrial load is often large and commercially time-sensitive. New factories are built to business schedules measured in quarters, while the network reinforcements to serve them can take years, and reshoring compounds the pressure. Incentives under the CHIPS and Science Act, the Inflation Reduction Act, and related state programs are drawing semiconductor, battery, and advanced-manufacturing investment into the same regions already contending with data center growth and transport electrification. Industrial electrification is therefore a core reason the grid is being asked, and will continue to be asked, to deliver more firm capacity, more quickly, in places not built for this pace of expansion.

1.1.3 Other Demand Drivers

Other demand drivers add a more diffuse layer of pressure. Residential and commercial electrification are gradually reshaping system load, especially through electric vehicles, heat pumps, and other end uses previously served by liquid fuels or combustion. NERC now explicitly includes household-appliance electrification and projected EV growth in its demand and energy estimates [3]. Individually small, these loads can in aggregate raise evening, overnight, or winter peaks depending on charging behavior, weather, and tariff design.

Several overlapping trends pressure the grid at once. Warehousing electrification, commercial fleet charging, and a wide range of new manufacturing loads may each look manageable in isolation, yet they compete for the same substations, feeders, and transmission corridors. This raises the stakes of forecast error and forces a slow-moving, risk-averse sector to adapt. A utility may absorb one category arriving below expectations; absorbing several in the same local area at once is far harder. For planning, data centers, industrial electrification, and other electrified end uses are best viewed as cumulative pressures on peak capacity rather than as separate problems solved one by one.

1.2 Peak Planning and Grid Infrastructure

These growth trends translate quickly into grid pain because the power system is planned around the hours in which failure would be unacceptable, rather than the average hour: the hottest summer afternoons, the coldest winter mornings, and the contingencies that follow the loss of a line or transformer. NERC's Facilities Design (FAC) and Transmission Planning (TPL) standards reflect this logic [7], [8]. When a utility studies a new large load, it asks whether the network can deliver firm megawatts at that location while staying within thermal limits, maintaining voltage, preserving protection coordination, and remaining stable under normal conditions and contingencies. A customer with modest annual energy use but a high coincident peak can therefore trigger infrastructure upgrades. This approach is by design, to ensure that power is always available and consumers can always rely on their service.

The logical impact of this is that the grid is “underutilized” in most hours. A 2025 Duke Nicholas Institute analysis of the 22 largest U.S. balancing authorities found load factors ranging from 43 to 61 percent, with an average and median of 53 percent [9]. Much of the system is therefore built to serve peaks that occur infrequently, under contingency conditions that are rare. That off-peak headroom disappears exactly when the system is most stressed, and it may be absent at the specific feeder, substation, or deliverability area where a new customer wants service. Low average utilization therefore does not disprove the capacity problem. Appearing on the Volts podcast, Ian Magruder of the Utilize Coalition likened it to a freeway engineered for rush-hour capacity rather than average traffic, built so that nobody slows down even at peak, which leaves it largely underused the rest of the year [10]. The reality is slightly more nuanced because real-time grid operations are rarely “perfect,” but the message still holds, and the existing infrastructure may have more to offer if it is managed optimally.

Peak planning is critical because those are the hours when the risks are highest. A grid-modernization lead at a large California utility described a constrained heat event by saying the utility had only a narrow window before it would have to start dropping load, which is the last thing the utility wants. The risk of dropping critical loads that have major safety or financial implications ensures that the grid is planned with safety factors above what would be needed at a theoretical peak efficiency. That is the operational meaning of underutilized infrastructure, since most hours have spare capacity, but the few binding hours force utilities to build more capacity, curtail customers, or find controllable demand that can ride through the constraint.

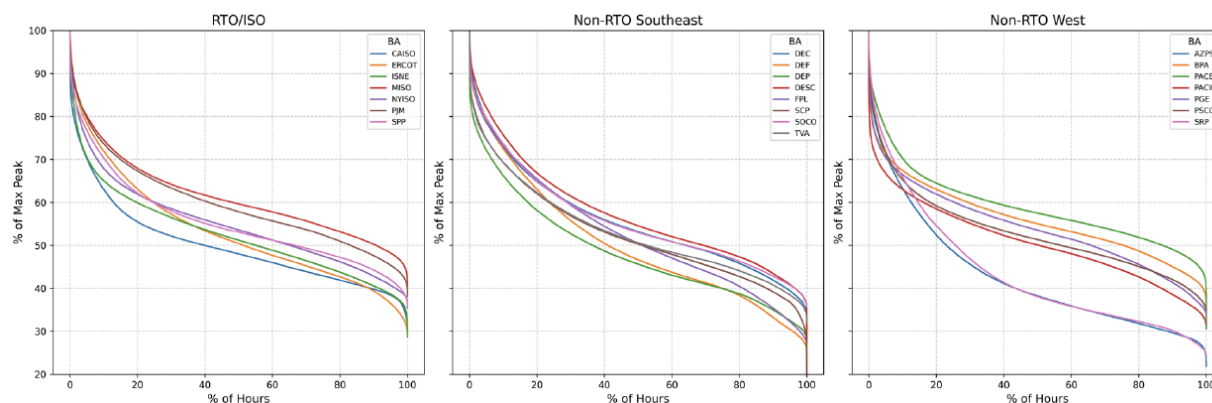


Exhibit 4: Load duration curves across U.S. RTOs and independent system operators (ISOs) show that even the peakiest systems run at well below maximum demand for most hours of the

year, illustrating why infrastructure built for the top few percent of hours is structurally underused the rest of the time [9].

The infrastructure build cycle tightens the constraint further. Berkeley Lab's Queued Up 2025 reports roughly 10,300 active projects seeking interconnection, about 1,400 GW of generation and 890 GW of storage, yet only 13 percent of capacity requesting interconnection from 2000 through 2019 had reached commercial operation by the end of 2024 [11]. The Department of Energy's Transmission Impact Assessment adds a second warning. Newly built high-voltage transmission fell from an average of 2,500 miles per year in 2013-2015 to just 56 miles in 2023, with major projects averaging development timelines of about a decade [12]. Even where supply is proposed, the path from proposal to deliverable capacity is slow, uncertain, and constrained by grid infrastructure. An interconnection planner at a large utility noted that at this point if you are applying to interconnect a large load facility in today's reality, you are likely looking beyond 2032 to be energized.

This is the central framing for the study. The United States is short of planning-grade, deliverable capacity in specific places and hours, more than it is short of electricity in aggregate. That explains why queue volumes can be enormous and new loads wait, why flexible and conditional service structures are gaining appeal, and why the value of demand-side flexibility is increasing. If a small number of high-stress hours creates the binding constraint, a resource that can reliably move load out of those hours can unlock materially more headroom for the rest of the grid.

2. Flexibility as a Solution to Grid Constraints

If Chapter 1 establishes that the grid bottleneck is primarily a peak-capacity problem, the next question is how much additional demand can be accommodated without waiting years for new transmission, substation, or distribution upgrades. The answer explored in this paper is flexibility, the ability of loads to change when and how they consume electricity, so the system need not treat every megawatt as a fully firm, all-hours obligation. Utilities build for stressed hours but recover those costs over the entire year. Any resource that reliably reduces demand in the binding hours raises utilization of existing infrastructure and can postpone capital that short-duration peaks would otherwise trigger.

Not all flexibility is equally valuable for grid planning. A planner does not reward an abstract willingness to use less power; the relevant questions are:

- how many megawatts can be adjusted
- at what location
- with what notice
- for how long
- how frequently
- with what consequences if performance fails.

This chapter compares the main demand-side resource classes through the lens most relevant to grid-capacity relief, namely whether they can provide bounded, auditable, and locationally relevant reductions that utilities and system operators can count in planning and operations.

2.1 How Flexibility Works

Flexibility works by converting time into capacity. If a customer can move some consumption away from the constrained hours, by shifting production, pre-cooling, drawing on storage, using on-site generation, or accepting a temporary operating limit, the same physical network can serve more load overall. The Duke Nicholas Institute's 2025 headroom analysis shows the payoff: 76 GW of new large load could be integrated with average annual curtailment of just 0.25 percent, rising to 98 GW at 0.5 percent and 126 GW at 1.0 percent [9]. The 0.25 percent means that the load would have to deploy some form of flexible solution in roughly 20 hours every year.

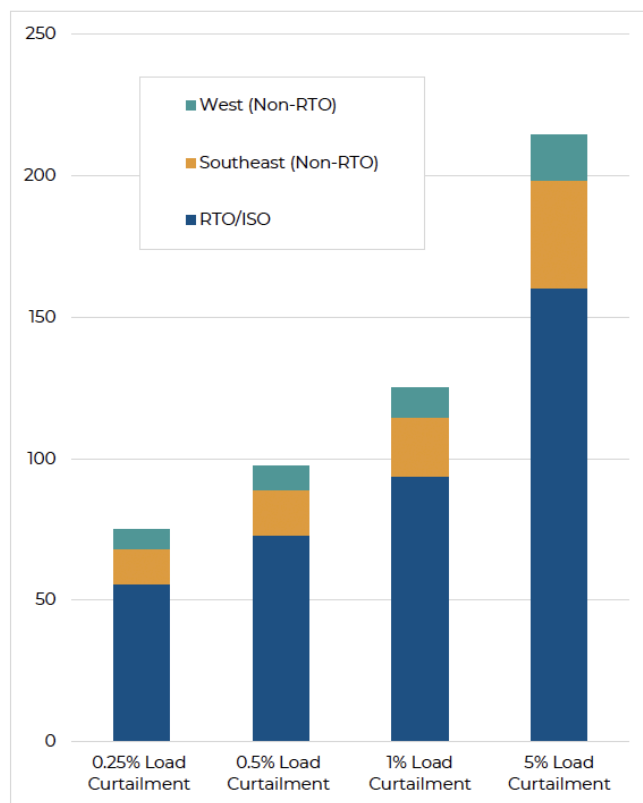


Exhibit 5: Across the 22 largest U.S. balancing authorities, accepting just 0.25 percent annual curtailment of new load unlocks 76 GW of headroom, rising to 126 GW at 1.0 percent and 215 GW at 5.0 percent, showing the convex relationship between small flexibility commitments and large capacity gains [9].

The grid value of flexibility comes from two related sources: deferring investments driven by peaks and allowing additional load to connect faster than it otherwise could. When the binding constraint is a transformer, feeder, substation, or local transmission interface that overloads only during a few hours, shaping demand can be cheaper than building new hardware. That logic is now explicit in utility and policy discussions. Recent state-level work on grid utilization and affordability argues that better use of existing assets lowers the cost of accommodating new demand, especially when the alternative is a peak-driven expansion of infrastructure that sits underused most of the year. Pacific Gas and Electric’s (PG&E) Flex Connect pilot illustrates the principle: in its published example, a site requesting 2 MW that would otherwise be limited to 0.5 MW year-round can access the full 2 MW most of the time if it responds to scheduled or real-time power limits during constrained hours [13].

The second value stream is speed-to-power. In the conventional model, scarce grid capacity is rationed through time, since a large customer that cannot be served until a substation, feeder, or transmission upgrade is completed simply waits. The economic value lost during that waiting period is mostly invisible to the planning process. The study may identify the cost of the physical upgrade, but it rarely prices the customer’s opportunity cost of delayed operations, delayed revenue, lost market timing, or foregone local economic activity.

Flexibility changes that by creating an intermediate service model between “serve the full load immediately” and “make the customer wait for years.” A customer may be able to connect earlier

if some portion of its demand is conditional during the limited hours when the grid is constrained. A high-value new load may also be willing to fund flexible capacity elsewhere on the system if doing so creates reliable headroom for earlier service. In both cases, the value extends beyond avoided infrastructure cost to the conversion of waiting time into a compensable capacity product.

Flexibility for capacity relief is distinct from flexibility for energy arbitrage or ancillary services. Energy arbitrage captures value by shifting consumption away from expensive hours, yet a customer can save money without solving a local grid constraint. Ancillary services reward very fast response for frequency or reserves but are usually smaller in scale and governed by different performance rules. This paper concerns planning-grade flexibility, meaning load adjustments that are locationally relevant, expressed in bounded megawatts, supported by telemetry or interval metering, and governed by a durable contract or tariff. PG&E's Base Interruptible Program shows what that looks like operationally, with a firm service level, a minimum load reduction, event caps, advance-notice requirements, and penalties for non-performance [14]. That standard is more demanding than generic smart load management, and it is what is required if flexibility is to substitute for infrastructure or accelerate interconnection.

Flexibility is valuable to the grid only when it is controlled, observable, contractually bounded, and integrated into planning and operations. Otherwise it can pose a risk to the grid, since uncontrolled tripping, sudden transfer to backup power, or oscillatory behavior in large loads can look to the grid like a disturbance rather than a dispatchable resource [48]. For industrial flexibility, the same plant action can be an asset or a problem depending on whether the system can anticipate, verify, and rely on it.

2.2 Sources of Flexibility

Demand-side flexibility is a family of resource types rather than a single market, differing in load-block size, control authority, contractual structure, and measurement difficulty. A hyperscale data center managing workload placement, a residential virtual power plant of thousands of thermostats and batteries, a portfolio of commercial buildings with pre-cooling, and an industrial site with interruptible motors or thermal storage may all be called flexible. To a utility planner, they behave very differently. The number of counterparties, dispatch risk, locational relevance, and maturity of program rules all vary.

The useful comparison is which category offers the best combination of load size, dispatchability, duration, scalability, commercial maturity, and regulatory readiness. Some resources are excellent for broad system balancing yet too diffuse for a local capacity problem; others offer large, measurable megawatt blocks but only in narrow sectors or under bespoke contracts. The sections below compare the four categories dominating the current conversation, data center, aggregated residential, commercial, and industrial flexibility, to identify which are most promising for grid-capacity relief rather than to crown a universal winner.

2.2.1 Data Center Flexibility

Data centers dominate the new flexibility discussion because they are both large loads and one of the fastest-growing demand segments. In principle, some of their demand is highly adaptable. Delay-tolerant AI training workloads can be shifted in time, slowed, or moved across geographies, and facility controls can optimize cooling, battery dispatch, and backup systems to reduce grid draw during stressed hours without interrupting computing tasks. An increasing number of companies are serving this space, seeing significant value creation if such solutions

can expedite the interconnection of new data centers, or mitigate the financial hit from curtailments during peak times.

Data center flexibility is strongest where the customer has full control, concentration, and contractual freedom, conditions most common in hyperscale environments and least common in colocation. Public evidence remains thin on how broadly utilities will credit software-based workload flexibility in planning studies, especially when the claimed reduction rests on a counterfactual baseline rather than a directly metered asset. Training and inference differ in ways that reinforce this limit. Training is more delay-tolerant and concentrated at very large sites, while inference is latency-sensitive and geographically distributed, which makes temporal shifting harder and raises the importance of always-available local service. Data center flexibility is therefore strategically important, yet today less standardized and less institutionally mature than many legacy demand response programs.

Companies such as Emerald AI are making this flex capability operational. Emerald AI's public announcements with National Grid Partners, the Electric Power Research Institute (EPRI), Nebius, and NVIDIA describe software that mediates between grid signals and AI infrastructure so selected workloads respond to grid conditions without disrupting mission-critical work [15]. Rather than treating the data center as an unpredictable block load, the software layer distinguishes workloads by urgency, moves or delays selected tasks, and exposes a bounded reduction to the grid.

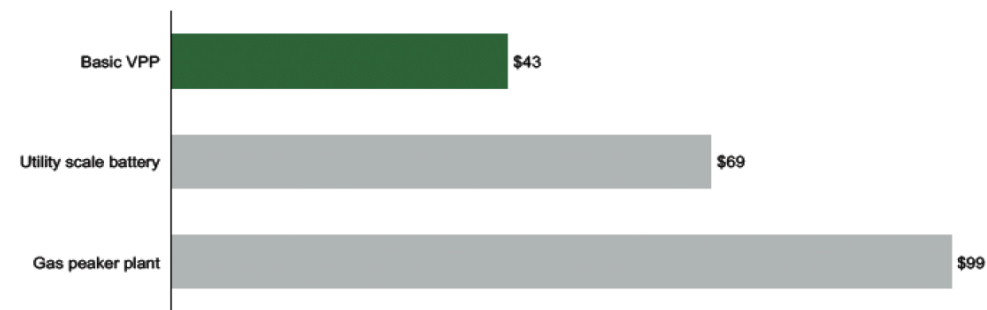
The evidence base is still early. Many utilities, regulators, and planners remain skeptical about how much data center load will actually flex once customer contracts, latency requirements, service-level agreements, and reputational risk are accounted for. A data center may be technically able to move some computation, yet that does not mean a utility can count the reduction in a planning study.

Data center flexibility also has a power-quality dimension distinct from simple load shifting. An executive at a power-electronics and storage firm cautioned that AI load is continuously changing with no rhythm to it as opposed to being uniform. Some behind-the-meter storage at data centers may be needed to smooth fast fluctuations and protect equipment rather than to provide grid capacity. Gauraang Biyani in the Stanford MBA program recently published two reports on data center flexibility, diving into the details of how these work [16]. He separates three concepts that are sometimes conflated, shifting compute in time or geography, using batteries to manage intra-site volatility, and using flexible loads elsewhere on the grid to support data center interconnection.

2.2.2 Aggregated Residential Flexibility

Aggregated residential flexibility, usually discussed as virtual power plants, is the broadest and most publicly popular model. It combines thousands or millions of small distributed devices, including smart thermostats, electric water heaters, EV chargers, and rooftop solar paired with batteries, into a coordinated portfolio that can act like a power plant. The scale potential is real. DOE projects that expanding VPP capacity could address a meaningful share of peak load by 2030 while lowering system costs [17]. DOE also finds that a VPP of residential thermostats, water heaters, EV chargers, and batteries can provide peaking capacity at materially lower net cost than a utility-scale battery or gas peaker. Residential VPPs are compelling because they reduce local peak demand, lower customer bills, and create a visible consumer-facing path to grid modernization.

Comparison of net cost to an example utility of providing 400 MW resource adequacy across three options, Net cost per kW-year



Note: Values for 400 MW of peaking capacity are shown in \$/kW-yr. The VPP analyzed consists of smart thermostats, smart water heating, home EV managed charging, and BTM battery demand response. Modeled equipment subsidy costs to utility are \$75 for smart thermostats, \$315 for smart water heaters, and \$0 for EV charging and BTM batteries. Marketing costs assumed at \$50 per device. Utility studied is assumed to have 50% renewable generation mix, with resource adequacy needs in summer and winter. 8760 hours were considered, and resources must be able to operate in 63 peak hours (when top 400 MW are needed) spanning 7 months, for 7 consecutive hours at a time. Benefits of emissions reduction and resilience are not shown; when included, VPP net cost is lower, though actual emissions impact will vary by local grid mix.¹¹⁶

Source: Brattle Group's Real Reliability: The Value of Virtual Power Report

Exhibit 6: Net cost to a utility of procuring 400 MW of resource adequacy is roughly \$43 per kW-year for a basic virtual power plant, \$69 for a utility-scale battery, and \$99 for a gas peaker, illustrating why aggregated demand-side capacity can substitute for supply-side build at materially lower cost [17].

Residential aggregation also exposes the gap between theoretical scale and planning-grade deliverability. A utility facing a local substation or feeder constraint usually needs dependable relief from a specific part of the network, rather than a probabilistic portfolio of small devices spread across a territory. Residential resources require very large customer counts, sustained acquisition and retention effort, and consistent telemetry and control across heterogeneous devices and vendors. Even when pilots perform well, utilities stay cautious about accrediting the full claimed value, because participation, communications reliability, and local deliverability vary. For broad peak shaving and system flexibility, residential VPPs are likely indispensable; for large, site-specific capacity problems, they are often too fragmented to substitute cleanly for a few large, contractually controlled loads.

2.2.3 Commercial Flexibility

Commercial flexibility occupies the middle ground between residential aggregation and industrial demand response. The resources are familiar, including heating, ventilation, and air conditioning (HVAC) in large buildings, thermal mass for pre-cooling, building automation, and behind-the-meter batteries on commercial campuses. Compared with residential demand response, commercial sites generally have better controls, fewer counterparties, and more predictable operating patterns, which makes them easier to enroll and dispatch. Many can deliver useful one-to-four-hour flexibility with limited disruption, especially when the strategy shifts load rather than shutting equipment off.

However, most commercial sites are materially smaller than large industrial facilities, so the megawatt yield per contract is lower. Split incentives between building owners and tenants complicate investment in controls, storage, or program participation. Although commercial loads are often easier to curtail than industrial process loads, they are more likely to be bound by occupancy and comfort constraints that narrow dispatch windows on the hottest or coldest days. Commercial flexibility is useful, yet it rarely combines the load concentration, production-linked value, and multi-year commitments that define the best industrial opportunities. It is an important

middle layer that is more reliable and scalable than purely residential aggregation, but usually less consequential per site than industrial load.

2.2.4 Industrial Flexibility

Industrial flexibility stands out for the size, variability, and concentration of its loads. A single facility may represent several to tens of megawatts at one point on the grid, so even modest percentage adjustments create meaningful local relief. The customer base differs too. Industrial operators are already accustomed to interval metering, complex tariffs, maintenance planning, and multi-year commitments. The Federal Energy Regulatory Commission (FERC)'s 2025 Assessment of Demand Response and Advanced Metering offers a useful reality check. Industrial customers accounted for 13,840 MW of potential retail peak demand savings in 2023, 45 percent of the national total, exceeding both residential and commercial classes. They also reported a 75 percent advanced-metering penetration rate, so the telemetry foundation for auditable programs is more developed than often assumed [18].

This does not make industrial flexibility simple or universal. Many industrial processes are continuous, safety-critical, or tightly tied to product quality, and not every large plant is a viable demand response resource. But rather than shutting a full factory down, the best industrial flexibility can come from specific sub-loads, such as motors and pumps, compressed air, refrigeration, thermal buffers, and on-site generation that changes net grid import during constrained hours. Because the resource is concentrated in fewer, larger sites and can be expressed in operationally bounded terms, industrial flexibility is well suited to the local, auditable capacity relief utilities value, though this depends on the flexible plant sitting at the same node as the constraint, which may not hold in fast-growing, constrained urban networks. That mix of scale and credibility makes it the focus of this paper.

2.3 The Case for Industrial Demand Flexibility

The case for industrial demand flexibility rests on several factors, and scale is where it starts. Industrial loads are large enough to matter at the circuit, substation, and system levels, and demand-response data show they already supply a meaningful share of enrolled dispatchable demand in many regions. A single facility can offer megawatts of relief at one interconnection point, and a portfolio of such facilities becomes a planning-relevant resource once the reductions are measured, contractually committed, and locationally deliverable. Scale also frames the data center debate. Although data centers dominate the load-growth conversation (see §1.1.1), industrial retail electricity sales were about 1,035 TWh in 2024, roughly 26 percent of U.S. retail sales and several times larger than the data center base [19]. That gap is expected to persist even as it narrows over the coming decade, so the industrial base remains one of the largest existing demand segments on the grid, far from a niche source of flexibility.

Size alone does not decide which loads should flex first. Flexibility is best dispatched from the lowest-opportunity-cost load that can solve the relevant constraint, rather than from the largest or most visible new one. Data centers often derive very high economic value from each megawatt-hour, because electricity supports high-margin computing, so they tend to have a higher willingness to pay for firm service than other customers. When the goal is to maximize total economic output while maintaining reliability, curtailing a data center first can be inefficient if an industrial, commercial, or municipal load can provide the same relief at lower cost. This points to a flexibility merit order in which the first resources called have low production losses, clear operating buffers, reliable metering, and acceptable customer risk, while higher-opportunity-cost loads participate at compensation that reflects the value they forgo. A well-designed program sorts flexible resources by locational usefulness, performance confidence, and opportunity cost, rather than by sector.

Even data centers benefit when industrial flexibility is available. An executive at a grid-software vendor frames a data center as a changing stack of compute work rather than one monolithic load, noting that “every moment in time, they have a different stack of compute loads,” with the lowest-priority tier representing work that is theoretically movable. The difficulty is timing, because a grid constraint may arrive when the facility is running mostly high-value compute, so “it’s not guaranteed that those two timings match up.” In those hours the data center wants a backstop, today usually on-site batteries. An incumbent industrial facility in the same constrained transmission area can serve that function at lower marginal cost than a dedicated battery, provided its response is contractually firm and dispatchable on the same notice horizon. With industry estimates putting hyperscaler capacity value in the tens of millions per MW for accelerated interconnections, well above the cost of an industrial controls or buffer upgrade, industrial flexibility is one of the few resources that can meet a data center’s reliability bar, potentially at the right location and notice, at a cost the data center can underwrite.

This flexibility also comes in the durations grid planners care about. According to a Western utility, some industrial chemical plants in its territory can regularly shed three to ten megawatts for three to four hours, and a few can turn compressors off for up to 24 hours, especially where the process has storage or redundancy built in. That range may stretch further as grids lean more heavily on renewable supply. Most of the flexibility discussed in this paper operates over hours, but a smaller and more exploratory line of research points to value at seasonal horizons, where demand and renewable output can drift out of step for weeks at a time. Modeling of aluminum smelting in China illustrates the possibility, finding that energy-intensive plants carrying spare capacity could concentrate production when power is plentiful and idle whole lines during prolonged scarcity, at a system saving large enough to offset the added storage and maintenance cost [43]. A load that can stand down for weeks would act as seasonal storage on the demand side, the kind of resource a renewable grid finds hardest to build, and it recasts overcapacity, usually written off as a stranded liability, as something a grid could choose to retain. The use of industrial capacity as this type of long-duration seasonal storage is still novel, and while it is not the core focus of this paper, it is a worthwhile area for further investigation.

The second pillar is institutional credibility. Industrial operators are better positioned than residential customers to negotiate multi-year commitments, invest in controls and metering, and accept explicit performance obligations. By leaning on those contractual commitments, utilities can plan with more certainty than probabilistic models allow. The limitation is that industrial flexibility must be designed around plant economics and operational intrusiveness. The same 5 MW reduction may be cheap from a chilled-water buffer or compressed-air optimization, but costly if it changes staffing, workflow, product quality risk, or production-line operation. Separating those cases before dispatch is the core program-design task. Industrial resources are rarely the fastest, and many need day-ahead or hour-ahead notice rather than sub-minute dispatch, which is usually acceptable for recurring peak events. For planning, the relevant question is whether the resource can deliver a bounded reduction under repeatable conditions, with commercial consequences for non-performance.

Regulatory and commercial maturity reinforces this credibility. Residential virtual power plants are scaling, but their rules, protocols, and valuation methods remain uneven across jurisdictions. Data center flexibility is receiving intense attention, yet many of its most important pathways, especially flexible interconnection and third-party capacity-offset structures, are still emerging through pilots, bespoke agreements, and early tariff reforms. Industrial demand response sits on a thicker institutional foundation, since existing programs already assign large customers firm service levels, event caps, telemetry requirements, baselines, penalties, and settlement rules. That foundation leaves the sector-specific challenges discussed later

unresolved, yet industrial flexibility still starts closer to planning-grade status at scale than the public conversation often recognizes.

A final reason is timing. Flexibility is becoming a condition of grid connection rather than an optional feature, and the shift is already visible. Utilities such as Silicon Valley Power and Portland General Electric now typically require new data centers to connect flexibly (see Part 2, §1.3.2), and that expectation is moving outward. New and electrifying industrial loads will likely meet the same constrained grids, interconnection queues, and speed-to-power pressures that made flexibility a prerequisite for data centers, which makes them the natural next candidates to be designed flexible from the start, saving millions per facility compared with flexibility retrofits. The existing industrial base, which already holds firm service, is the harder and later case, yet it is also the largest pool of latent flexibility on the system. Industrial flexibility is therefore the next focus area of a shift already underway because of the data center boom.

This remains a conditional thesis. Industrial flexibility will not replace residential virtual power plants, nor eliminate the need for transmission and distribution upgrades, and some sectors will stay too bespoke, risky, or operationally constrained to scale. The opportunity in industrial flexibility can be described as tranches: if sub-loads are properly defined, there may be some that can yield capacity that is significantly more economically viable than the alternative of building more infrastructure or curtailing a data center. Comparative evidence suggests industrial demand flexibility is one of the more underexplored near-term tools for grid-capacity relief, because it combines large site loads, relatively mature measurement infrastructure, and a more legible path to contractual accountability than competing resource classes.

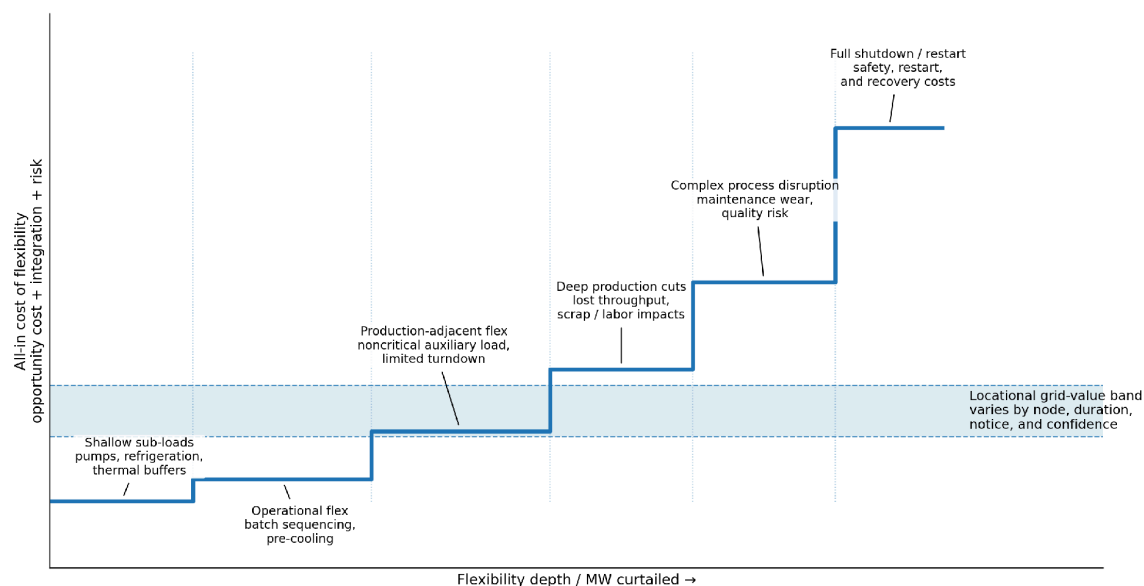


Exhibit 7: Illustrative depiction of the flexibility cost curve. Industrial flexibility is a tranche-based merit order, not a single resource. Low-disruption sub-loads such as pumps, compressors, refrigeration, thermal buffers, and batch sequencing may provide capacity below the locational grid-value band, while deeper production cuts and shutdowns can become uneconomic once throughput, quality, and restart costs are included. Source: Author illustration. Costs, tranches, and grid-value band are illustrative and do not represent a specific facility, tariff, or market price.

2.4 Emerging Roles of Industrial Flexibility

Industrial flexibility as a capability is decades old, since large plants have long shifted load, or shut down, to manage their bills and to support the grid in emergencies. What has changed is the value the system now places on it. As new load arrives faster than the network can be rebuilt, the same operating flexibility takes on four distinct roles, each drawing on a different source of value.

It can accelerate the arrival of new power, relieve the specific places where the grid is already tight, hedge a load forecast that no one can pin down, and lift the long-run efficiency of the system as a whole. The roles overlap in practice, yet each makes its own case, and together they explain why industrial flexibility deserves a central place in the response to the coming demand wave.

2.4.1 Accelerate Speed to Power

For a growing class of grid customers, the scarcest resource is time. A data center that cannot energize for four years, or a factory that cannot run a new electric furnace until a substation is rebuilt, loses revenue, market position, and sometimes the project itself. The conventional interconnection process offers these customers only full firm service with a wait measured in years. Flexibility breaks this paradigm. By accepting a defined limit on its draw during the handful of constrained hours in a year, a customer can begin operating well before the grid is rebuilt around it, and the utility avoids reserving scarce capacity for hours when the customer would not use it.

For data centers, the value of this accelerated speed to power is immense. Power has become the number one constraint on new data center developments, because costly data center hardware depreciates quickly. This incentive has driven the renewed attention on load flexibility; however data centers still face challenges. The technical know-how to make data centers flexible is still nascent, and they generally cannot risk being unexpectedly curtailed.

Industrial sites are unusually well placed to take this route, especially when they electrify or expand. A new increment of electric load, an added boiler, furnace, compressor train, or process line, can often be carried on conditional terms and kept separate from the plant's existing firm service, so the core operation never bears the risk. The payoff is counted in years of earlier output, set against transmission projects that routinely take close to a decade to complete. The long-term buildout is still required, and flexibility simply lets productive load start earning while that buildout proceeds.

2.4.2 Relieve Near-term Constraints

Scarcity on the grid is local before it is general. A region can hold ample energy across the year and still be unable to serve a new customer at one substation during a dozen stressed hours, because the binding limit sits on a particular transformer, feeder, or transmission interface. A resource helps with that bottleneck only when it sits electrically close to it, which is what gives industrial flexibility an edge here that diffuse, system-wide resources lack.

A single plant can place several to tens of megawatts behind one interconnection, so even a partial reduction can free real headroom at the asset that constrains it. A few megawatts of dependable reduction in the constrained window can substitute for a transformer or feeder upgrade that would otherwise cost millions and take years to build. The role is strongest for incumbent plants already embedded in the tight pocket, since they bring service history, interval data, and a fixed location, and a plant inside the affected area can release capacity that a new project elsewhere cannot. Everything here rests on performance a planner can rely on, because

no utility will defer an upgrade against a reduction it cannot measure, enforce, and count when the constraint binds.

2.4.3 Hedge Uncertain Load Growth

Planners now work from forecasts they cannot trust. Demand projections swing by gigawatts between cycles as data centers and other loads appear in concentrated and often speculative blocks, while much of the announced load will move location, slip in time, or never reach its requested scale. The hard question has shifted from how much load will grow to which of these projects will actually appear, where, and when.

Major infrastructure is not an ideal solution in such uncertain conditions. These assets are slow to build and nearly impossible to unwind once approved, so a planner who builds for the high forecast risks stranding capital on load that never arrives, while one who builds for the low forecast risks turning customers away. Flexibility offers a way through, because it can be procured in small increments, aimed at the places under pressure, and adjusted as real load reveals itself. A utility can extend a flexible industrial resource modest credit at first and revise it as performance data accumulate, or lean on it to bridge the gap between load that has arrived and infrastructure that has not. Its value behaves like an option. The grid secures the right to call on a reduction during uncertain peaks while keeping the freedom to build firm capacity once the picture clears, and that option grows more valuable as the forecast grows more volatile.

2.4.4 Build a More Efficient Grid

The first three roles answer the present scramble for capacity. The fourth concerns the kind of grid that scramble leaves behind. A system sized for its rare peaks runs far below its capability in most hours, as the load-factor evidence in Section 1.2 showed. Of course, some of that slack is deliberate margin for reliability and to allow headroom for further growth, and to maintain maneuverability for grid operators. The difficulty is that a grid which keeps meeting growth by building for ever-higher peaks risks growing steadily more expensive, especially if some of that growth does not materialize and customers carry the cost of capacity that sits idle the rest of the year.

Controllable demand can help alleviate this concern. When an industrial load can stand down during the narrow windows that set the peak, it raises how much useful work the existing network performs without eroding the margin that keeps the system safe. The gain is also an affordability strategy. A deferred transformer, smaller substation upgrade, or more load served behind existing wires reduces the capital that must be recovered through rates. This is the most durable of the four roles and the slowest to realize, because it depends on regulators and utilities treating utilization as a planning objective instead of defaulting to construction. It also marks the limit of what flexibility can do, making existing transmission, distribution, generation, and storage go further without standing in for them.

These four roles show the potential value that load flexibility can bring, with much of it being untapped today. Speed to power and local relief act on the immediate bottleneck, hedging buys time against an uncertain future, and structural efficiency reshapes the long-run cost of the system. All four rest on the same precondition. Each depends on turning a plant-level operating capability into a resource that planners can count, operators can dispatch, and capital providers can finance.

3. From Context to Action

Part 1 has established the premise that industrial flexibility matters only if it can be made planning-grade. The opportunity is not simply to identify loads that can move; many industrial facilities already contain some movable load. The harder question is when that movement is reliable, locationally useful, commercially worthwhile, and credible enough to change grid-planning decisions.

Parts 2 and 3 turn to that execution problem. Part 2 examines how industrial flexibility can be paid for and why adoption remains difficult, and Part 3 sets out what must change to convert plant-level capability into trusted capacity. The next 12 to 24 months are important because data center interconnection pressure is forcing utilities, regulators, and market participants to define new flexibility models now. The question is whether those models remain bespoke solutions for new large loads or become the foundation for a broader market in industrial flexibility.

References for Parts 1, 2, and 3

- [1] U.S. Energy Information Administration. *Short-Term Energy Outlook*. January 2025.
<https://www.eia.gov/outlooks/steo/archives/jan25.pdf>
- [2] Grid Strategies. *Power Demand Forecasts Revised Up for Third Year Running, Led by Data Centers*. November 2025.
<https://gridstrategiesllc.com/wp-content/uploads/Grid-Strategies-National-Load-Growth-Report-2025.pdf>
- [3] North American Electric Reliability Corporation. *2025 Long-Term Reliability Assessment*. 2025. https://prod.nerc.com/globalassets/our-work/assessments/nerc_ltra_2025.pdf
- [4] Shehabi, Arman, et al. Lawrence Berkeley National Laboratory and U.S. Department of Energy.
2024 United States Data Center Energy Usage Report
. December 2024.
https://eta-publications.lbl.gov/sites/default/files/2024-12/us_data_center_energy_usage_report_lbnl-2001637_0.pdf
- [5] Sightline Climate data, cited by Axios. Harder, Amy.
Why the AI Data Center Boom Is Stalling Worldwide
. Axios, February 24, 2026. <https://www.axios.com/2026/02/24/ai-data-center-boom-projects-numbers>
- [6] U.S. Department of Energy, Alternative Fuels Data Center.
U.S. Plug-in Electric Vehicle Registrations by State
. <https://afdc.energy.gov/vehicle-registration>
- [7] North American Electric Reliability Corporation.
Reliability Standard FAC-002-4: Facility Interconnection Studies
. <https://www.nerc.com/pa/Stand/Reliability%20Standards/FAC-002-4.pdf>
- [8] North American Electric Reliability Corporation.
Reliability Standard TPL-001-5.1: Transmission System Planning Performance Requirements
. <https://www.nerc.com/pa/Stand/Reliability%20Standards/TPL-001-5.1.pdf>
- [9] Norris, Tyler H., Tim Profeta, Dalia Patino-Echeverri, and Adam Cowie-Haskell. Nicholas Institute for Energy, Environment & Sustainability, Duke University.
Rethinking Load Growth: Assessing the Potential for Integration of Large Flexible Loads in U.S. Power Systems
. February 2025.
<https://nicholasinstitute.duke.edu/sites/default/files/publications/rethinking-load-growth.pdf>
- [10] Volts. Episode featuring Ian Magruder of the Utilize Coalition on grid utilization and large-load growth. <https://www.volts.wtf/>
- [11] Lawrence Berkeley National Laboratory.
Queued Up: 2025 Edition, Characteristics of Power Plants Seeking Transmission Interconnection as of the End of 2024
. 2025. <https://emp.lbl.gov/queues>
- [12] U.S. Department of Energy, Office of Policy.
Transmission Impact Assessment: Power Sector Infrastructure Deployment and Investment
. October 2024.
https://www.energy.gov/sites/default/files/2024-10/DOE_OP_2024_Report-Transmission_Impact_Assessment.pdf
- [13] Pacific Gas and Electric Company.
Flexible Service Connection Pilot Overview
. <https://www.pge.com/assets/pge/docs/clean-energy/electric-vehicles/flexible-service-connection->

pilot-overview.pdf

[14] Pacific Gas and Electric Company.

Electric Schedule E-BIP: Base Interruptible Program

and related Base Interruptible Program customer materials.

https://www.pge.com/tariffs/assets/pdf/tariffbook/ELEC_SCHEDS_E-BIP.pdf

[15] NVIDIA, Emerald AI, Electric Power Research Institute, and DCFlex. Public materials on flexible AI factories and data center demand flexibility.

<https://nvidianews.nvidia.com/news/nvidia-and-emerald-ai-join-leading-energy-companies-to-pioneer-flexible-ai-factories-as-grid-assets> ; <https://dcflex.epri.com/>

[16] Biyani, Gauraang.

Flexible AI Data Centers: Four Pathways to Power AI Sustainably

. Stanford draft, 2026.

[17] U.S. Department of Energy. *Pathways to Commercial Liftoff: Virtual Power Plants*. 2023.

<https://www.energy.gov/oced/virtual-power-plants>

[18] Federal Energy Regulatory Commission.

2025 Assessment of Demand Response and Advanced Metering

. 2025.

https://www.ferc.gov/sites/default/files/2025-12/25_Annual%20Assessment%20of%20Demand%20Response_1212.pdf

[19] U.S. Energy Information Administration.

Electric Power Monthly

, Table 5.1: Sales of Electricity to Ultimate Customers by End-Use Sector.

<https://www.eia.gov/electricity/monthly/>

[20] California Public Utilities Commission. Emergency Load Reduction Program (ELRP).

<https://www.cpuc.ca.gov/elrp/>

[21] California Energy Commission. Demand Side Grid Support (DSGS) Program.

<https://www.energy.ca.gov/programs-and-topics/programs/demand-side-grid-support-program>

[22] Southwest Power Pool.

Revision Request 703

materials and related stakeholder or Market Monitoring Unit comments on demand response accreditation and aggregation. <https://www.spp.org/>

[23] Southwest Power Pool.

Conditional High Impact Large Load Service

materials, including CHILLS and HILLGA filing materials. 2026.

https://www.spp.org/documents/75954/20260210_revisions%20to%20add%20the%20conditional%20high%20impact%20large%20load%20service_er26-1323-000.pdf

[24] Electric Reliability Council of Texas.

Large Load Integration

materials, including PGRR134, NPRR1284, PGRR145, NPRR1325, Batch Zero, Provisional Controllable Load Resource materials, and large-load interconnection Q&A.

<https://www.ercot.com/services/rq/large-load-integration> ;

<https://www.ercot.com/mktrules/issues/PGRR134> ;

<https://www.ercot.com/mktrules/issues/NPRR1284>

[25] Camus Energy, encoord, and Princeton ZERO Lab. *Flexible Data Centers: A Faster, More Affordable Path to Power*. 2025.

[26] Lawrence Berkeley National Laboratory and The Brattle Group.

Electricity Rate Designs for Large Loads: Evolving Practices and Opportunities

. January 2025.

https://eta-publications.lbl.gov/sites/default/files/2025-01/electricity_rate_designs_for_large_loads_evolving_practices_and_opportunities_final.pdf

[27] Federal Energy Regulatory Commission. Docket RM26-4,

Interconnection of Large Loads to the Interstate Transmission System

; U.S. Department of Energy Section 403 large-load interconnection materials, including Secretary Wright's October 23, 2025 direction and proposed ANOPR on large-load interconnection. <https://www.ferc.gov/rm26-4> ;

<https://www.energy.gov/sites/default/files/2025-10/403%20Large%20Loads%20Letter.pdf>

[28] National Association of Regulatory Utility Commissioners.

Performance-Based Regulation

guidance and related valuation and ratemaking materials.

<https://www.naruc.org/core-sectors/energy-resources-and-the-environment/valuation-and-ratemaking/>

[29] Prabha, Rajanie. Stanford report on Western U.S. grid utilization. Stanford University, 2026.

[30] Commonwealth of Virginia. House Bill 434 (identical companion Senate Bill 621), Electric utilities; electric grid utilization metrics; State Corporation Commission. 2026.

<https://lis.virginia.gov/bill-details/20261/HB434>

[31] Energy Systems Integration Group, Large Loads Task Force.

Resource Adequacy with Large Loads: Planning for Flexibility to Accelerate Integration

. April 2026.

[32] Electric Power Research Institute. *DCFflex Flex MOSAIC Framework*. EPRI, 2026.

<https://restservice.epri.com/publicattachment/97236>

[33] Electric Power Research Institute.

Flex MOSAIC

. DCFlex, 2026. <https://dcflex.epri.com/flex-mosaic>

[34] Utility Dive. "Grid Operators, Ratepayers Shouldn't Fear Flexible Data Centers: GridCARE." December 16, 2025.

<https://www.utilitydive.com/news/grid-operators-ratepayers-shouldnt-fear-flexible-data-centers-gridcare/808032/>

[35] Federal Energy Regulatory Commission and Southwest Power Pool. *Order Accepting Tariff Revisions, Subject to Condition* (HILL and HILLGA processes), Docket ER26-247-000, January 14, 2026, including Commissioner Rosner's concurrence.

https://www.spp.org/documents/75725/20260114_order%20-%20revisions%20to%20add%20the%20high%20impact%20large%20load%20processes%20and%20generation%20assessment_er26-247-000.pdf

[36] McGuireWoods. "Texas Senate Bill 6 Significantly Expands Regulatory Oversight Over Large Loads in ERCOT." July 2025.

<https://www.mcguirewoods.com/client-resources/alerts/2025/7/texas-senate-bill-6-significantly-expands-regulatory-oversight-over-large-loads-in-ercot/>

[37] Canary Media. "Oregon Utility Tries Out a Faster, Cheaper Way to Power Data Centers." November 4, 2025. <https://www.canarymedia.com/articles/data-centers/oregon-utility-tries-out-a-faster-cheaper-way-to-power-data-centers>

<https://www.canarymedia.com/articles/data-centers/oregon-utility-tries-out-a-faster-cheaper-way-to-power-data-centers>

[38] PV Magazine USA. "Avangrid Aims to Expand Its Flexible Interconnection Pilot in New York State." October 28, 2025.

<https://pv-magazine-usa.com/2025/10/28/avangrid-aims-to-expand-its-flexible-interconnection-pilot-in-new-york-state/>

[39] PV Magazine. "U.S. State Orders Utility to Enable Flexible Grid Connections for Community Solar." December 31, 2025.

<https://www.pv-magazine.com/2025/12/31/us-state-orders-utility-to-enable-flexible-grid-connections-for-community-solar/>

[40] LandGate. "Power-First Data Centers in 2025: How Grid Constraints Are Repricing Land, Leases, and Revenue." 2025. <https://www.landgate.com/news/power-first-data-centers-in-2025-how-grid-constraints-are-repricing-land-leases-and-revenue>

[41] Hitachi Ventures. "Flexible Interconnection: Rewiring the Grid for the AI Era." Medium,

2026.

<https://medium.com/@HitachiVentures/flexible-interconnection-rewiring-the-grid-for-the-ai-era-66fdb5fc8d8c>

[42] Volts (David Roberts). “Are utilities making too much money?” Conversation with Joe Daniel, RMI. February 11, 2026. <https://www.volts.wtf/>

[43] Lyu, R., Li, A., Wang, J., Luo, H., Shen, Y., Guo, H., Du, E., Kang, C., and Jenkins, J. “Can Industrial Overcapacity Enable Seasonal Flexibility in Electricity Use? A Case Study of Aluminum Smelting in China.” arXiv preprint 2511.22839, 2025.

[44] U.S. Department of Energy, Federal Energy Management Program. “Energy Savings Performance Contracts (ESPCs).”

<https://www.energy.gov/cmei/femp/articles/energy-savings-performance-contracts-espcs>

[45] Voltus. “Voltus and Google Announce Bring Your Own Capacity Agreement.” June 2, 2026.

<https://www.voltus.co/press/voltus-google-bring-your-own-capacity-pjm>

[46] Utility Dive. “Google to Fund 100-MW Virtual Power Plant in PJM in ‘First-of-its-Kind’ Deal.” June 3, 2026.

<https://www.utilitydive.com/news/google-virtual-power-plant-vpp-pjm-voltus/821838/>

[47] Latitude Media. “Google is Voltus’s First ‘Bring-Your-Own-Capacity’ Customer.” June 2, 2026.

<https://www.latitudemedia.com/news/google-is-voltus-first-bring-your-own-capacity-customer/>

[48] Energy Systems Integration Group, Large Loads Task Force. *Interconnection Processes for Large Loads: Current Practices and Recommendations*. June 2026.

<https://www.esig.energy/reports-briefs/large-load-interconnection-process>

[49] Federal Energy Regulatory Commission. *FERC Launches Aggressive Targeted Action to Speed Large Load Integration Across the Grid*. June 18, 2026.

<https://www.ferc.gov/news-events/news/ferc-launches-aggressive-targeted-action-speed-large-load-integration>

[50] Federal Energy Regulatory Commission. *Fact Sheet: FERC Takes Action to Supercharge America’s Grid for Efficiency, Reliability, and a Bold Energy Future*. June 18, 2026.

<https://www.ferc.gov/news-events/news/fact-sheet-ferc-takes-action-supercharge-americas-grid-efficiency-reliability-and>

[51] Federal Energy Regulatory Commission. *E-7 to E-12 Show Cause Orders Presentation, June Commission Meeting*, Docket Nos. EL26-67-000 et al. June 18, 2026.

<https://cms.ferc.gov/sites/default/files/2026-06/E-7%20to%20E-12%20SCO%20Presentation%20-%20June%20CM.pdf>

[52] Rosner, David. Federal Energy Regulatory Commission. *Commissioner Rosner’s Remarks on the Large Load Show Cause Orders, E-7 to E-12*. June 18, 2026.

<https://www.ferc.gov/news-events/news/commissioner-rosners-remarks-large-load-show-cause-orders-e-7-e-12-june-18-2026>

[53] U.S. Energy Information Administration. *U.S. Electric Utility Demand-Side Management 1994*. <https://www.eia.gov/electricity/data/eia861/dsm/058994.pdf>

[54] California Public Utilities Commission. *Interruptible Tariffs Background*. 2001.

https://docs.cpuc.ca.gov/publisheddocs/published/comment_decision/5757-01.htm

[55] Peak Load Management Alliance. *Defining the Evolution of Demand Response: From 1.0 to 2.0 and Beyond*.

<https://plma.memberclicks.net/assets/dr-dialogue/Evolution-of-DR-White-Paper.pdf>

[56] Federal Energy Regulatory Commission. *Order No. 719, Wholesale Competition in Regions with Organized Electric Markets*. 2008.

<https://www.ferc.gov/sites/default/files/2020-06/OrderNo.719.pdf>

[57] Federal Energy Regulatory Commission. *Order No. 745, Demand Response Compensation in Organized Wholesale Energy Markets*. 2011.

<https://www.ferc.gov/sites/default/files/2020-06/Order-745.pdf>

