Integrating Large Flexible Loads in US Power Systems

Tyler H. Norris

NewsData 2025 Western Energy Summit

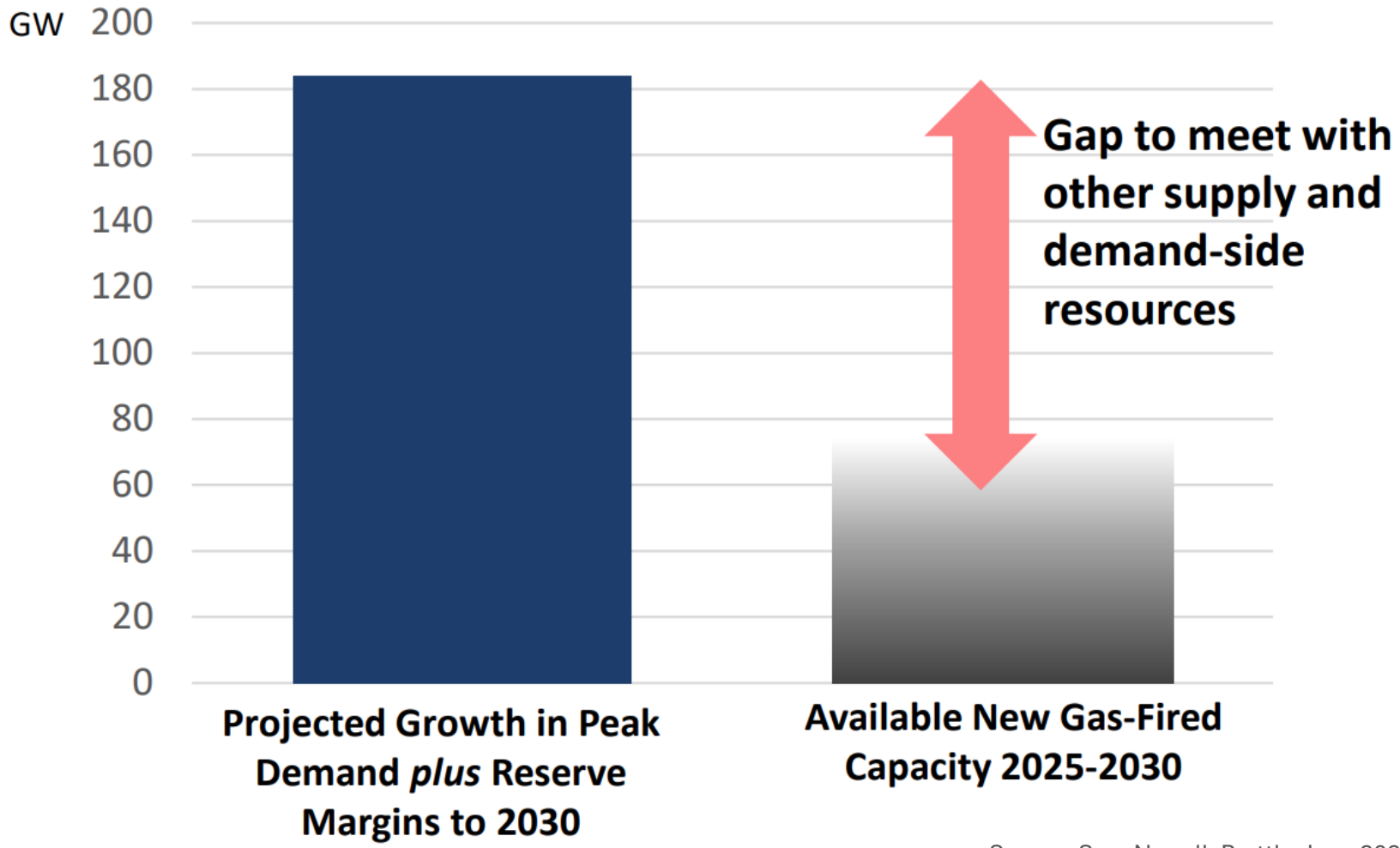
August 21, 2025

Duke NICHOLAS INSTITUTE for ENERGY, ENVIRONMENT & SUSTAINABILITY

GRACE Lab
A Grid that is Risk-Aware for Clean Electricity

NICHOLAS SCHOOL of the ENVIRONMENT

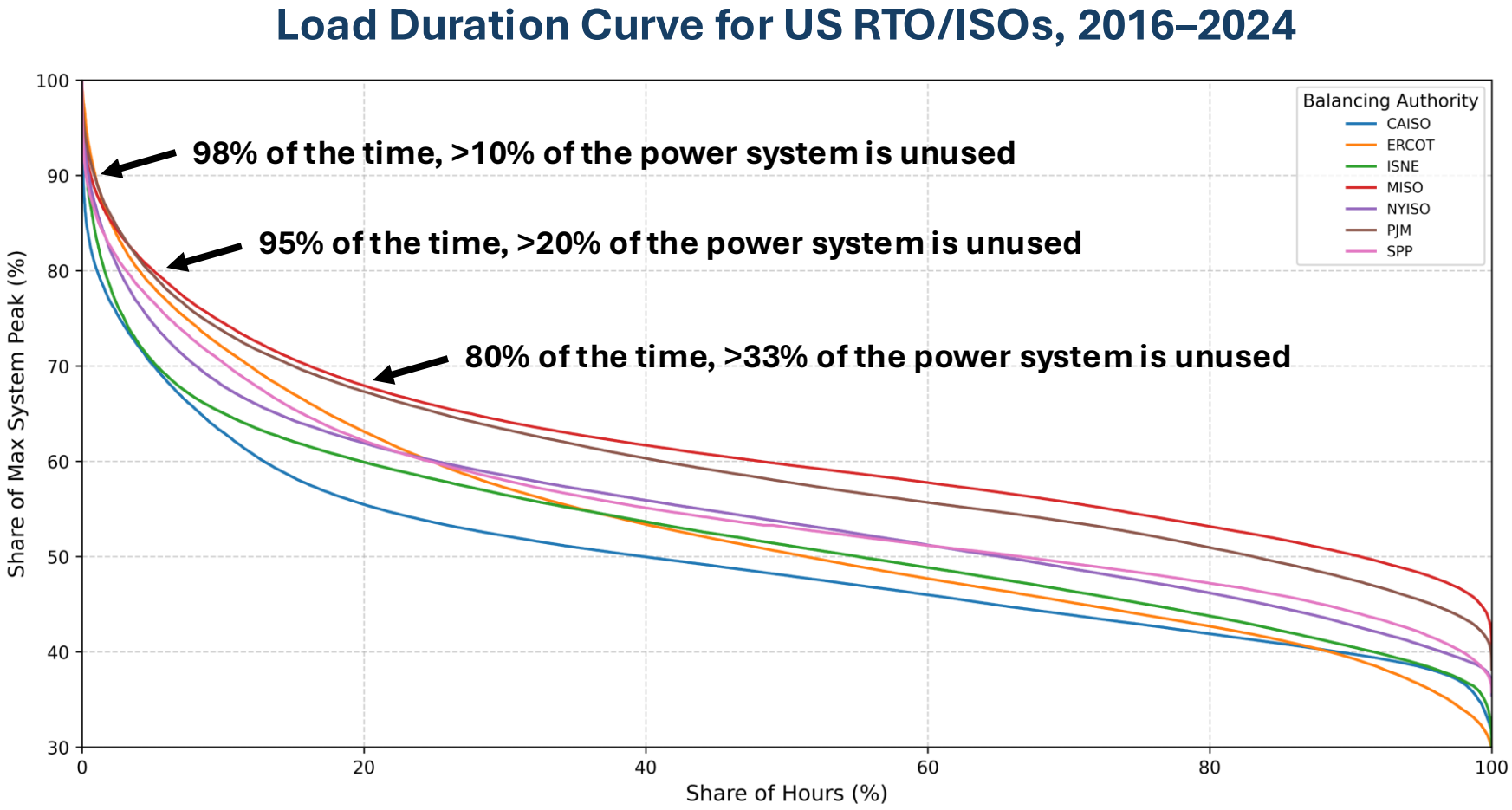
Bridging the supply-side gap



Source: Sam Newell, Brattle, June 2025

Headroom exists outside extreme peak hours

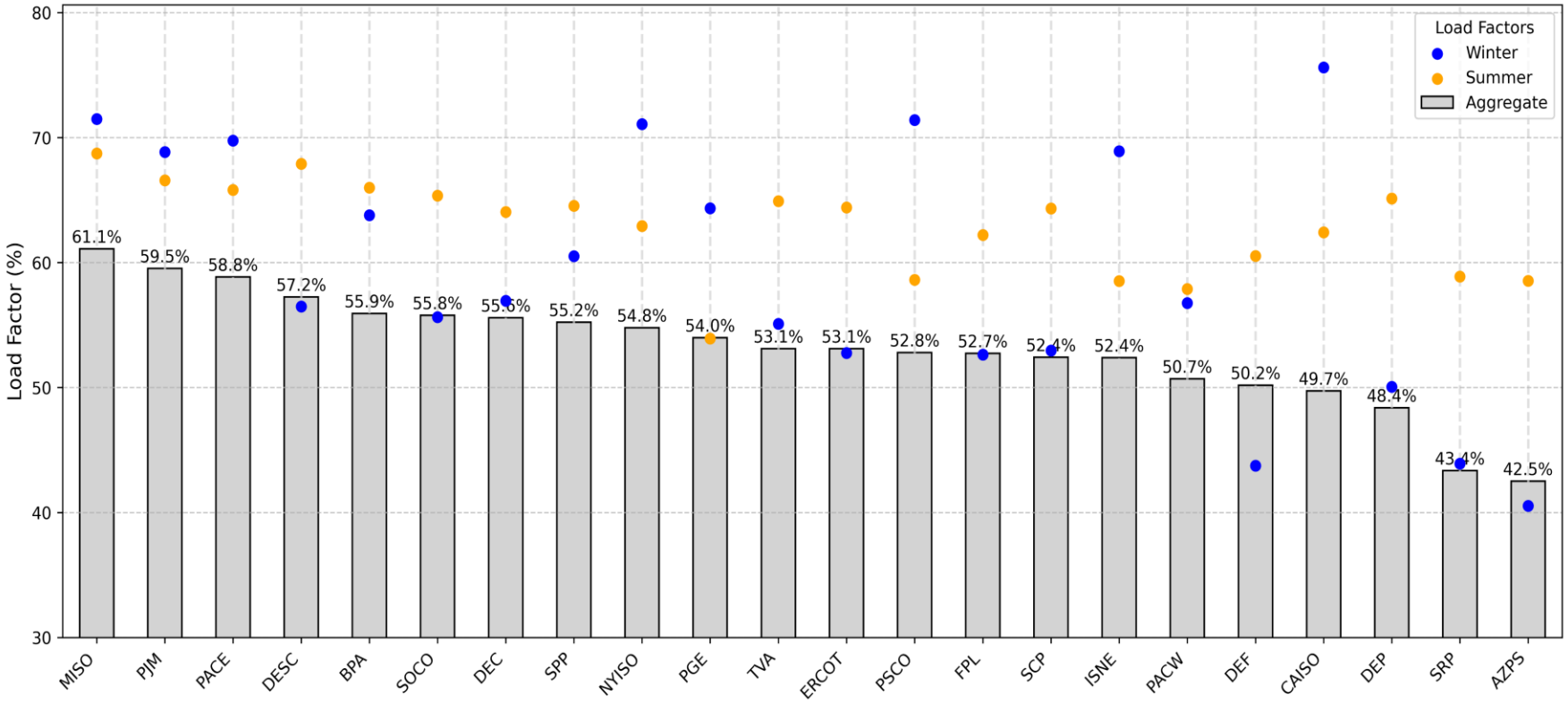
- The load duration curve illustrates system utilization by ranking demand from highest to lowest over a given period
- More than 10% of the system is built to serve 35 hours/year of extreme peak load (avg.)



US power systems operate at 53% avg. load factor

- Load factor is the ratio of average demand to peak demand and is an indication of system utilization
- Aggregate load factors range between 43% to 61%, with an average and median value of 53%
- Winter load factors were notably lower than summer (59% vs. 63% average)

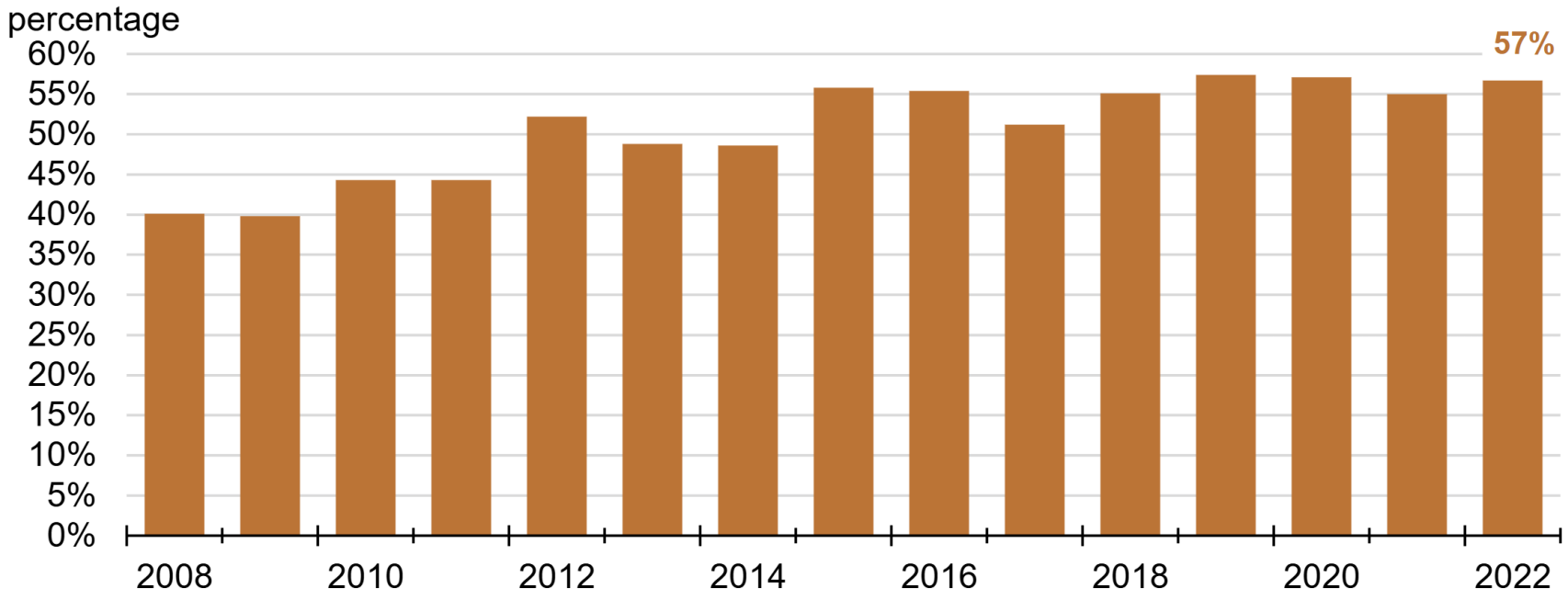
Load Factor by Balancing Authority and Season, 2016–2024



US NGCC fleet operates at <60% capacity factor

- NGCC plants run well below full potential, with a fleetwide capacity factor of 57% in 2022 – despite being among the most efficient thermal resources
- Even the most efficient NGCC units are underutilized – new plants average 64% utilization despite cutting-edge turbines and low marginal costs

Capacity Factor of US NGCC Power Plant Fleet (Annualized)



Source: EIA Form EIA-860M, Monthly Electric Power Industry Report

Types of load flexibility

On-site power and storage

Utilizing co-located storage, renewables, or other generators

Temporal flexibility

Scheduling computational loads before or after periods of high system stress

Spatial flexibility

Distributing workloads across one or multiple data centers in different geographic locations

Reduced operations

Planning for reduced workload during defined periods

Implications for system planning and interconnection

Inflexible Loads

- More burden on system peaks
- More likely to trigger need for additional generation and transmission capacity
- More likely to require longer interconnection

Flexible Loads

- Less burden on system peaks
- Less likely to trigger need for additional generation and transmission capacity
- Potential for accelerated grid interconnection

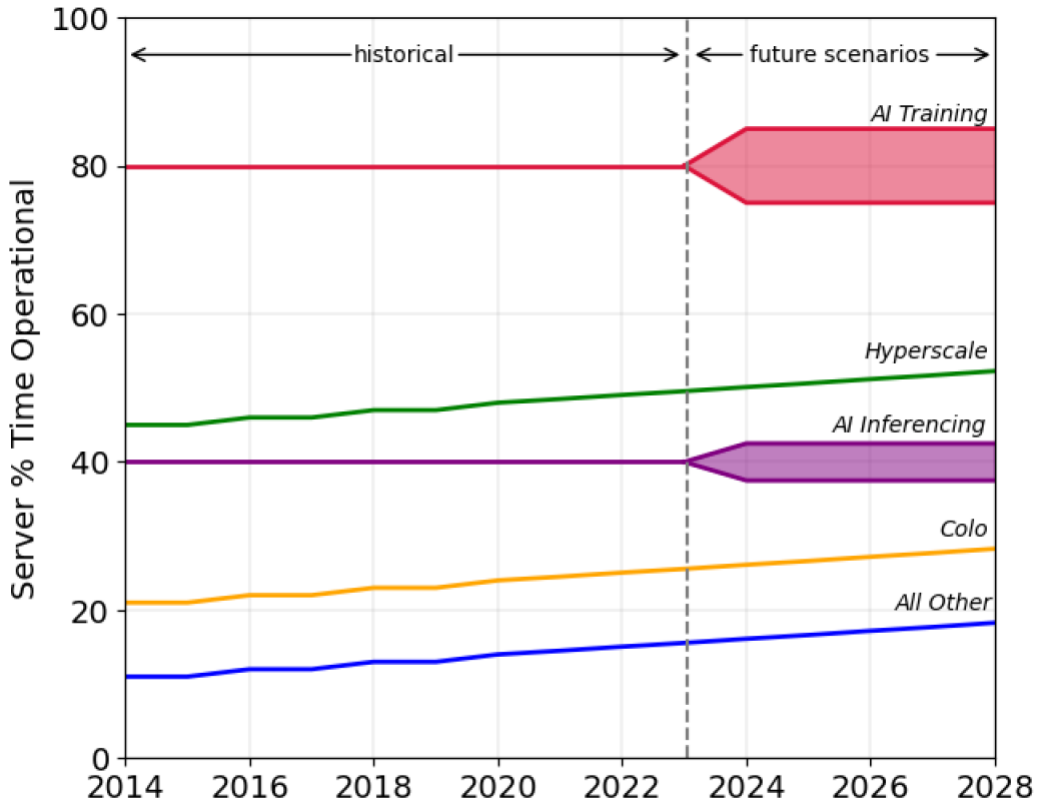
Data center utilization rates: perception vs. reality

Question: When an electric utility cites X gigawatts of AI data center load requests, how much utilization should regulators anticipate?

Source	Figure	Paper
Shehabi et al. 2025 (LBNL)	50%	*2024 United States Data Center Energy Usage Report*
Guidi et al. 2024	75%	*Environmental Burden of United States Data Centers in the Artificial Intelligence Era*
Patel et al. 2024 (SemiAnalysis)	80%	*AI Datacenter Energy Dilemma – Race for AI Datacenter Space*
Knittel et al. 2025 (MIT)	80%	*Flexible Data Centers and the Grid: Lower Costs, Higher Emissions?*
Netherlands Central Bureau of Statistics	"Unknown"	Cited in *How Data Centers Have Come to Matter*

Source: Norris, T. "The Puzzle of Low Data Center Utilization Rates." Aug. 2025. Power & Policy Substack. https://www.powerpolicy.net

Operational time of servers by data center type



Source: Shehabi, A., et al. 2024 United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory, Berkeley, California. LBNL-2001637

Data center utilization rates: perception vs. reality

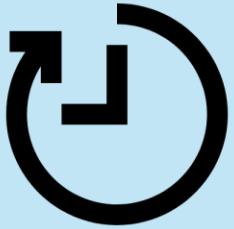
Clarification	Explanation
1. Load Factor ≠ Utilization Rate	Load factor is the ratio of average demand to realized peak demand. But if a facility's realized peak demand is only 80% of that facility's rated capacity, and its load factor is 90%, then its true *capacity utilization rate* is only 72%.
2. Load Factor ≠ Server Uptime	A facility with high electrical load factor may still have servers operating well below their compute capacity.
3. Server Uptime ≠ "Five Nines"	The industry standard uptime guarantee (99.999%) is about customer-facing availability, not actual server utilization.
4. Non-IT Load Variation Matters	Cooling and other infrastructure loads can fluctuate, especially during extreme weather events.
5. PUE ≠ IUE	PUE indicates nothing about a given facility's capacity utilization rate, which has been called "Infrastructure Usage Effectiveness (IUE)."

Source: Norris, T. "The Puzzle of Low Data Center Utilization Rates." Aug. 2025. Power & Policy Substack. https://www.powerpolicy.net

Potential values of flexibility

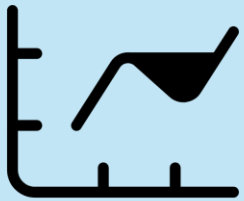
1 Faster Time-to-Power

- Dominion Energy: 7+ year wait time
- Centerpoint Energy: 700% increase in data center queue
- Several jurisdictions developing priority interconnection for flexible loads



2 Increased Interconnection Capacity

- Utilities and txm service providers limit load capacity based on worst-case load study results
- Flexible loads can be interconnected at higher capacities



3 Mandate Compliance

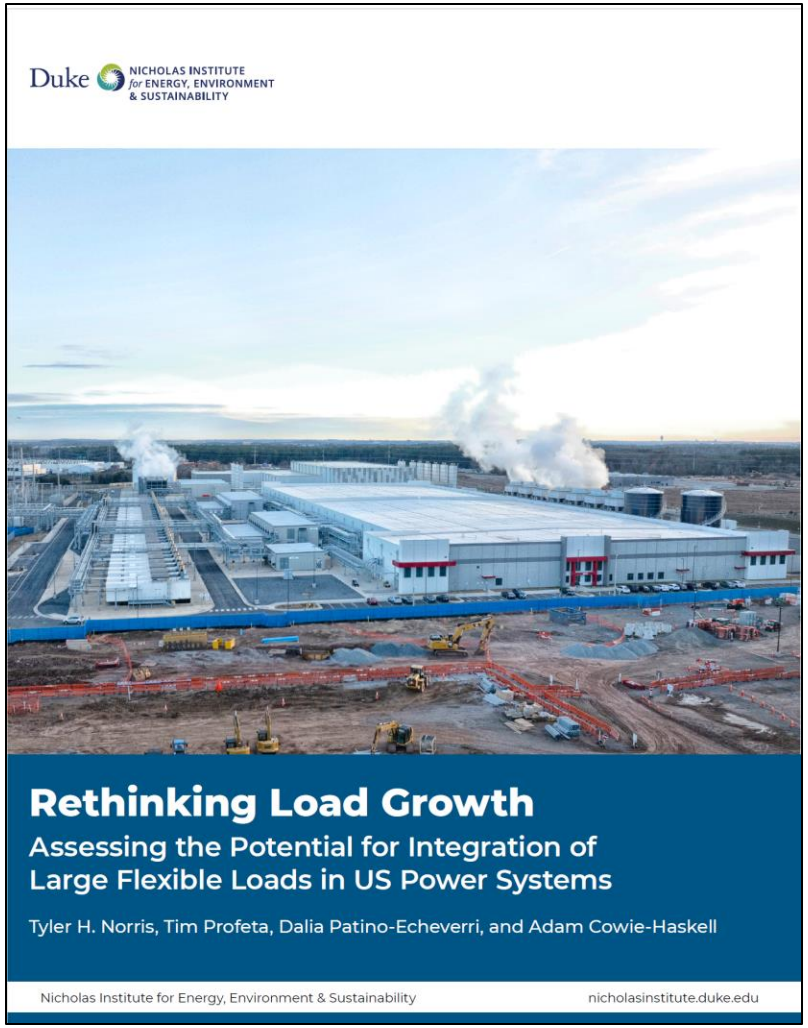
- Grid reliability and price affordability are suffering around the country
- Demand response mandates may be coming, e.g. TX SB6

4 Flexibility Revenues

- As peak demand soars, regional power systems are looking for shock absorbers
- Rising flexibility revenues could become material

Goals for *Rethinking Load Growth*

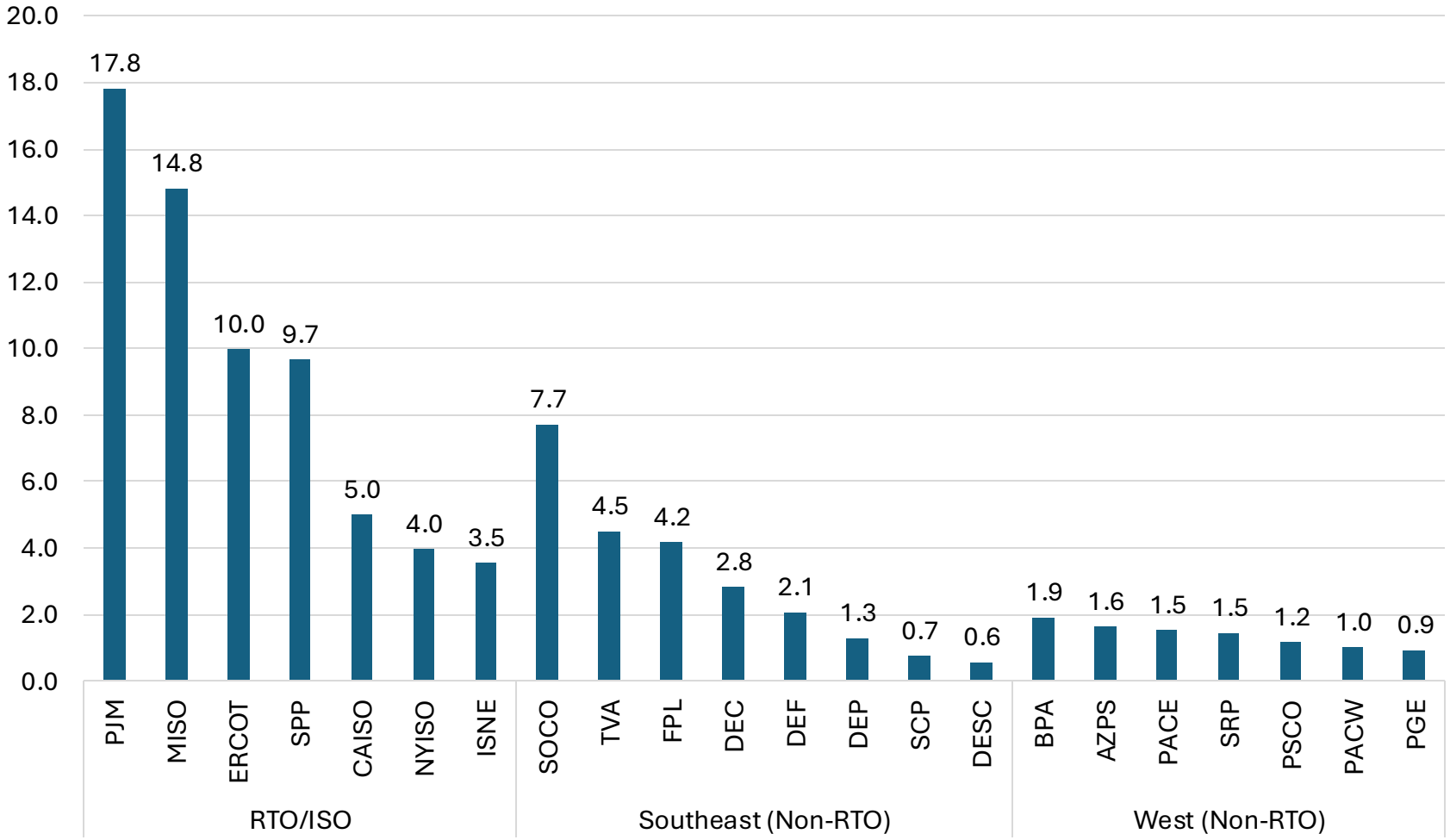
- Support regulators and stakeholders in identifying strategies to accommodate load growth without compromising reliability, affordability, or progress on decarbonization
- Provide informational resources and a first-order estimate of the potential for accommodating new loads while mitigating or deferring capacity expansion
- Motivate additional analysis to more precisely quantify headroom in each balancing authority



Results: Curtailment-Enabled Headroom

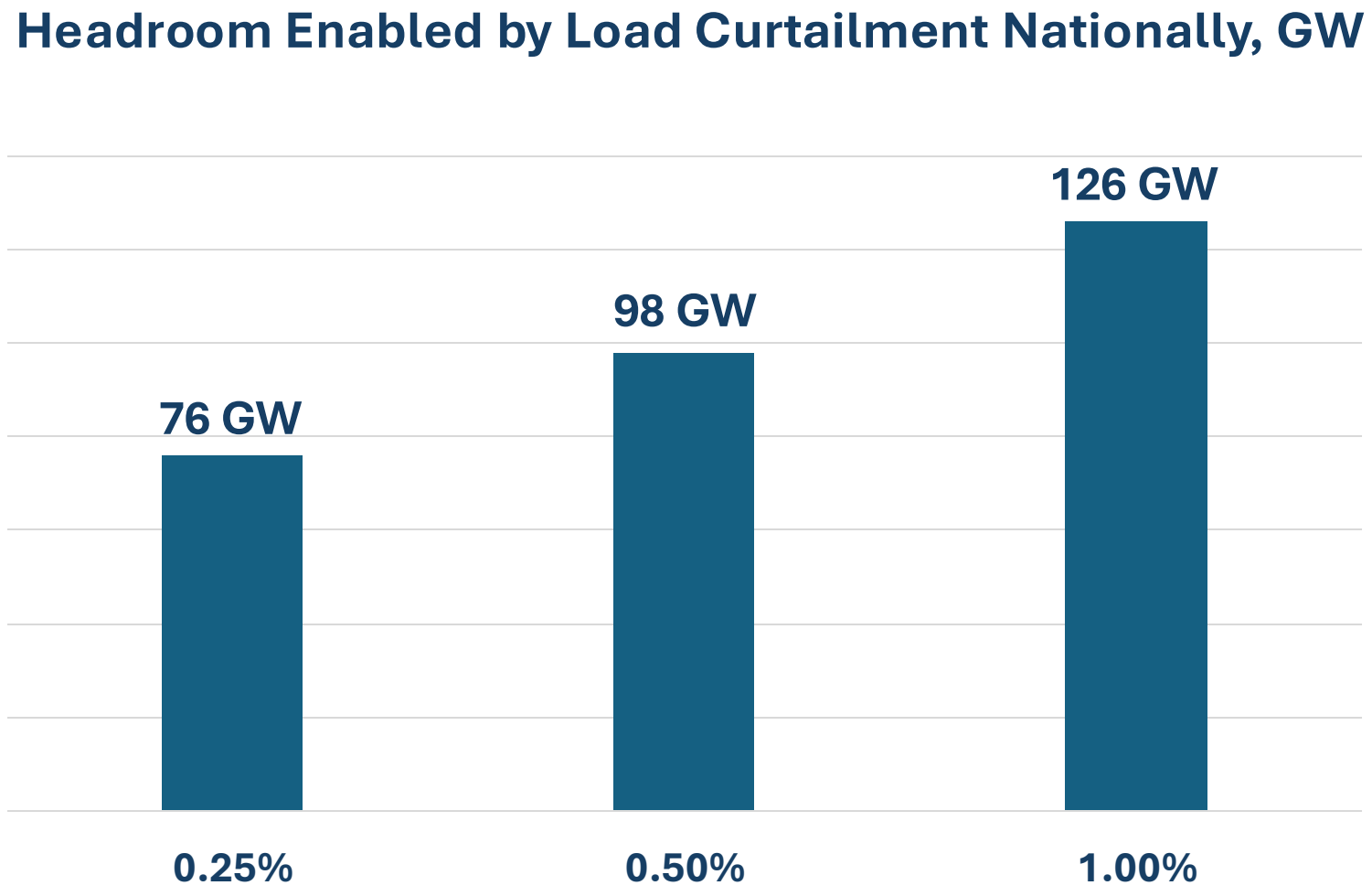
- Headroom across the 22 analyzed balancing authorities is between 76 to 215 GW, depending on the applicable load curtailment limit
- 76 GW of headroom is available at an expected load curtailment rate of 0.25%
- This headroom increases to 98 GW at 0.5% curtailment, 126 GW at 1.0%, and 215 GW at 5.0% curtailment

Headroom Enabled by 0.5% Load Curtailment by Market, GW



Results: Curtailment-Enabled Headroom

- Headroom across the 22 analyzed balancing authorities is between 76 to 215 GW, depending on the applicable load curtailment limit
- 76 GW of headroom is available at an expected load curtailment rate of 0.25%
- This headroom increases to 98 GW at 0.5% curtailment, and 126 GW at 1.0%



Recent developments

1

RTO/ISO and Utility Initiatives

- PG&E FlexConnect
- PJM colocation docket
- PJM Critical Issue Fast Path – Large Loads
- SPP non-firm service proposal
- NYISO flexible load forecast

2

Legislation

- Texas SB6
- New Jersey
- Massachusetts
- California
- US House Appropriations Committee

3

Market Participants

- EPRI DC Flex First Demonstrations
- Beacon AI commitment
- Verrus demo with NREL
- Google flexibility contracts: I&M and TVA

> *“We cannot do what we did in 1960s and just build. We have to build smart. And the need for speed will come from unlocking the headroom that you have in the grid through flexibility.*
>
> *… To keep electric rates for customers in check, flexibility will be the key. We have to not only build, but we have to use the underused capacity that we already have in the grid. And we've got to think differently.”*

– EPRI President & CEO Arshad Mansoor

EPRI Summer Seminar, August 2025

Thank you!

Contact:

tyler.norris@duke.edu

Duke NICHOLAS INSTITUTE *for* ENERGY, ENVIRONMENT & SUSTAINABILITY

GRACE Lab
A Grid that is Risk-Aware for Clean Electricity

NICHOLAS SCHOOL *of the* ENVIRONMENT