Carbon-Aware Computing: How to Get Power Systems and Data Centers to Talk to Each Other

Vladimir Dvorkin

Department of Electrical Engineering and Computer Science

University of Michigan – Ann Arbor

IEEE CSS Day 2024

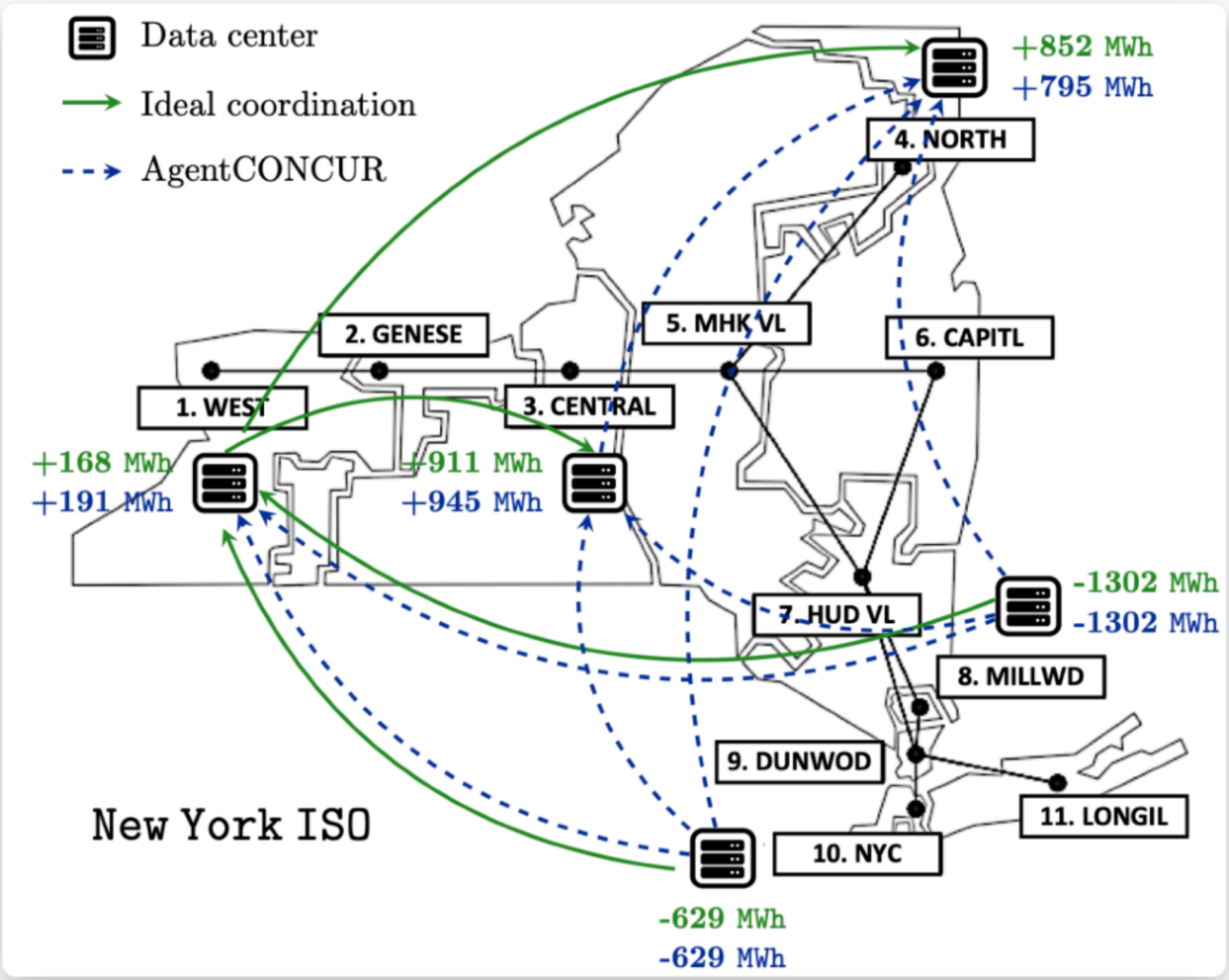
October 21, 2024

OptiML

https://web.eecs.umich.edu/~dvorkin/

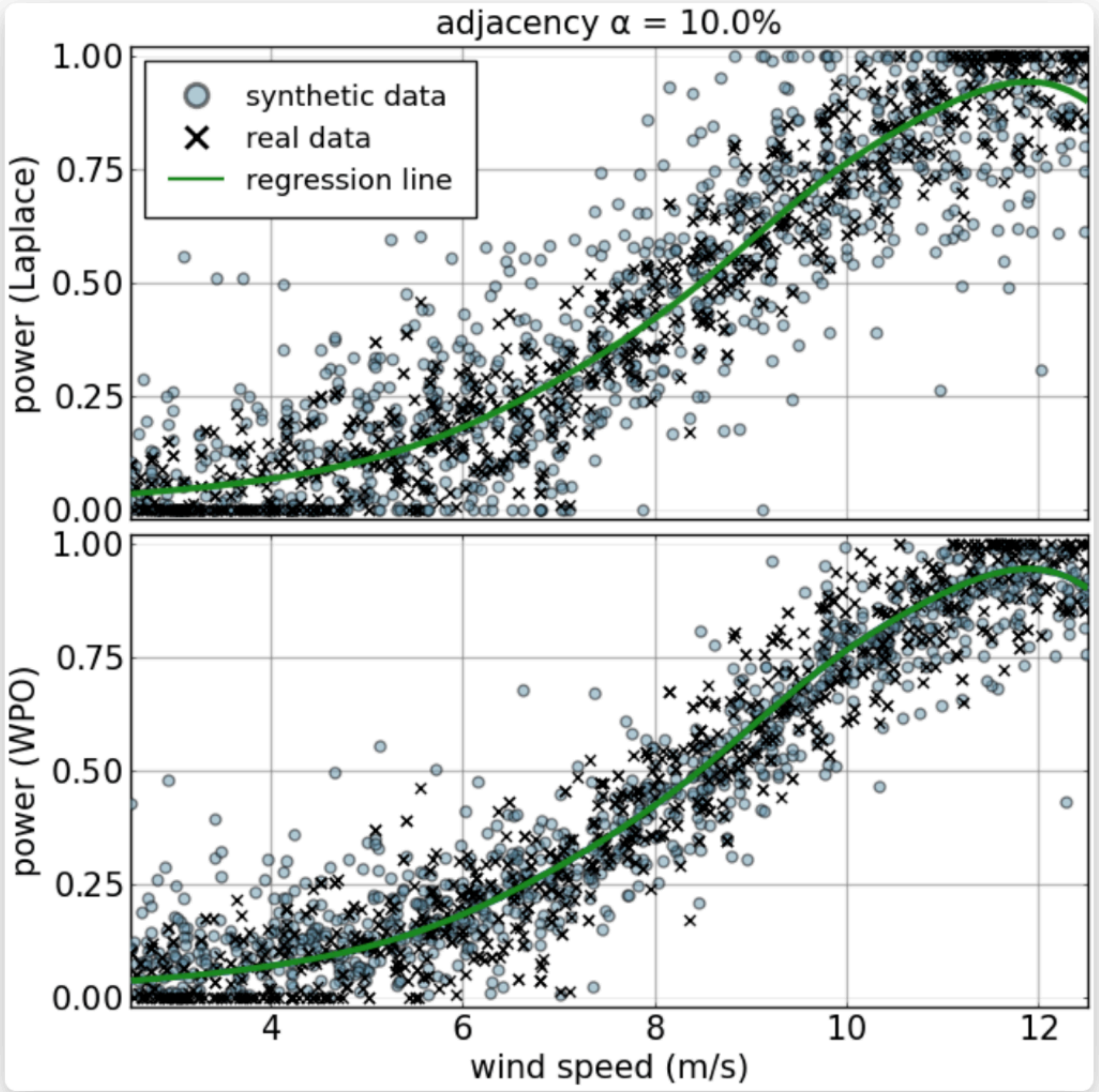
/ˈɑp·tə·məl/

Optimizing energy. Empowering society.



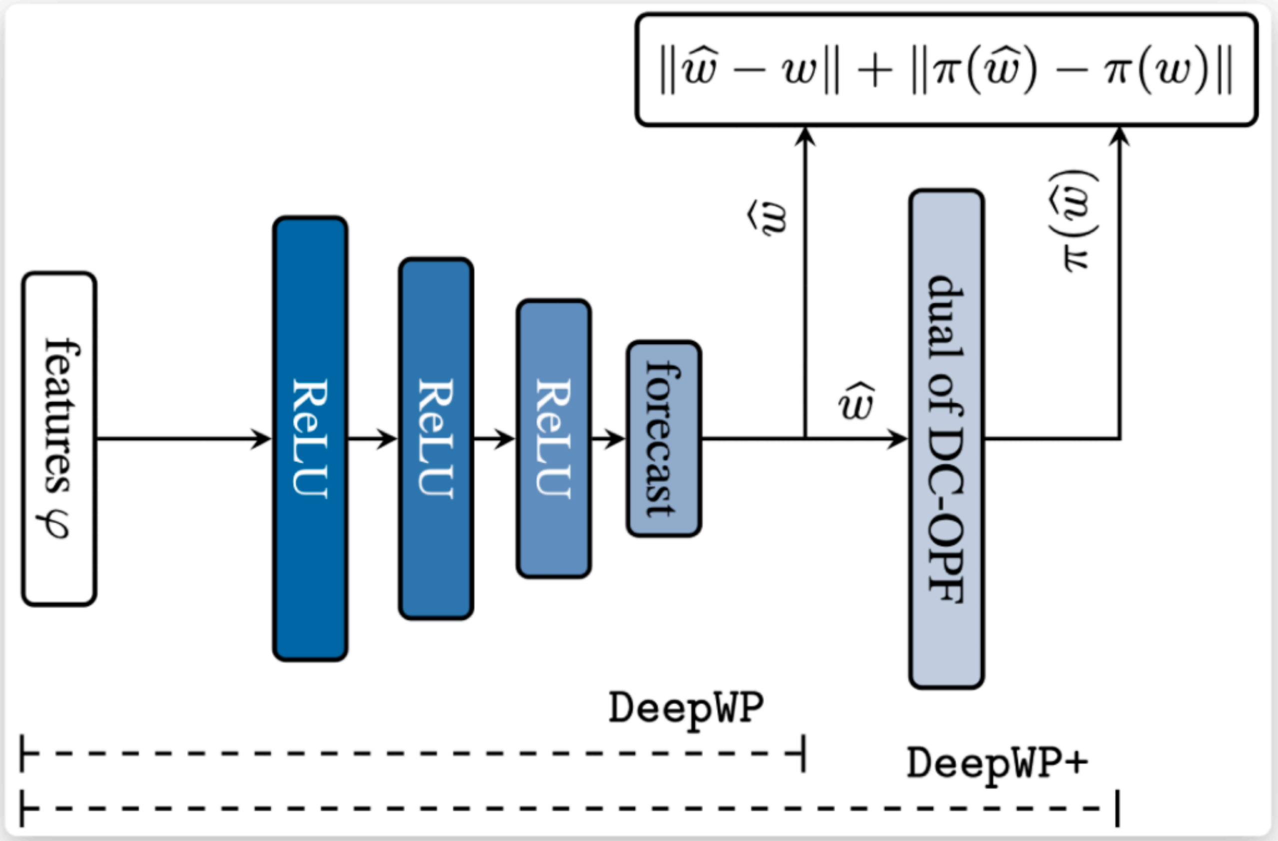
Agent coordination via contextual regression (AgentCONCUR) is a cost-effective protocol, both in terms of data and computation, designed to coordinate spatial shifts of data center electricity consumption to support power grid operations during peak hours.

[paper] [code]



Wind Power Obfuscation (WPO) is a noise-additive algorithm to create synthetic data from real measurements while maintaining higher quality than standard Laplace-based algorithms.

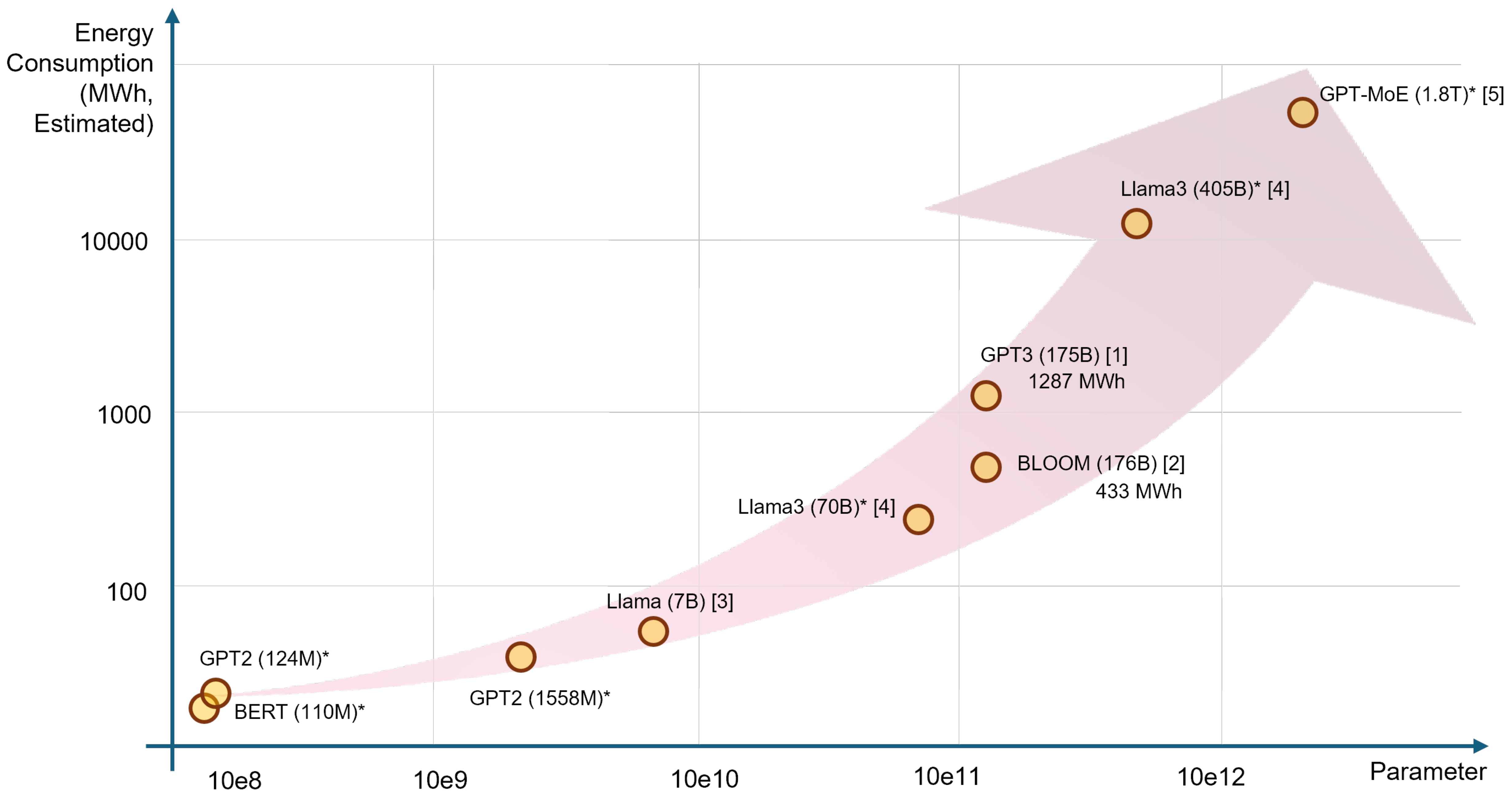
[paper] [code] [FERC talk]



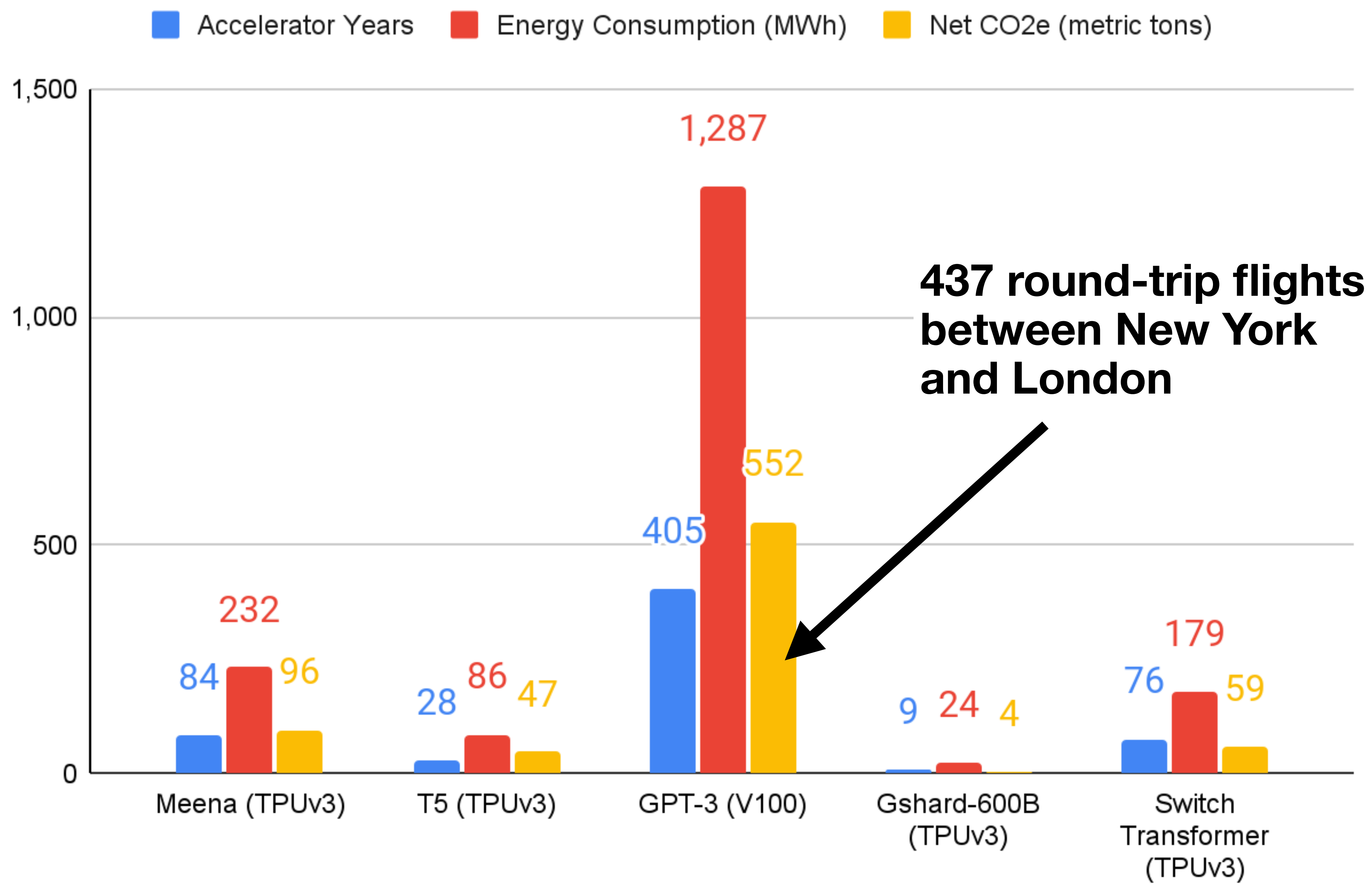
DeepWP+ is a new deep learning architecture designed to enhance wind power predictions by addressing errors and unfairness in electricity prices. Unlike traditional architectures, DeepWP+ embeds a market-clearing optimization problem that guides predictions towards more accurate and fair outcomes.

[paper] [slides]

AI is currently the fastest growing electricity demand



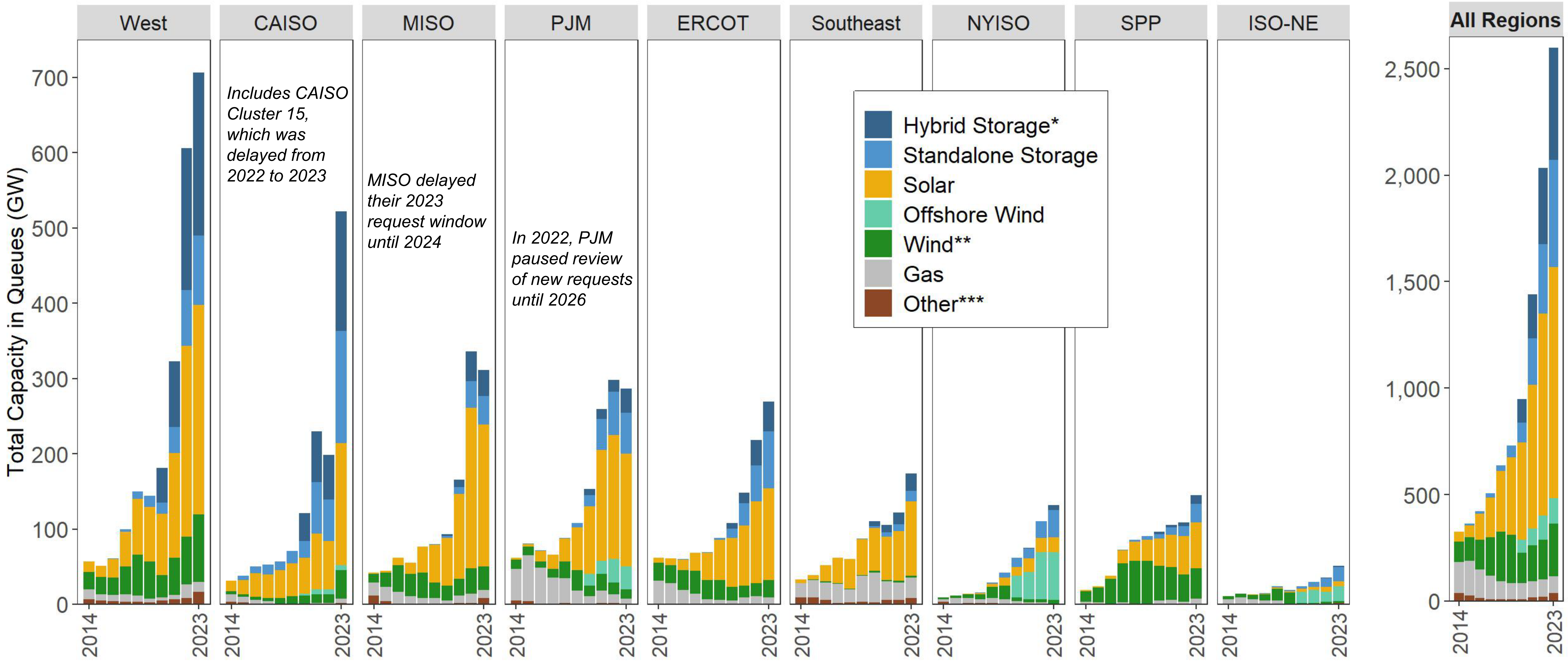
Li *et al.* The Unseen AI Disruptions for Power Grids: LLM-Induced Transients. 2024



Patterson *et al.* Carbon Emissions and Large Neural Network Training. 2021

- Data centers in the U.S. could consume as much electricity by 2030 as some entire industrialized economies
- Data centers have a significant environmental impact

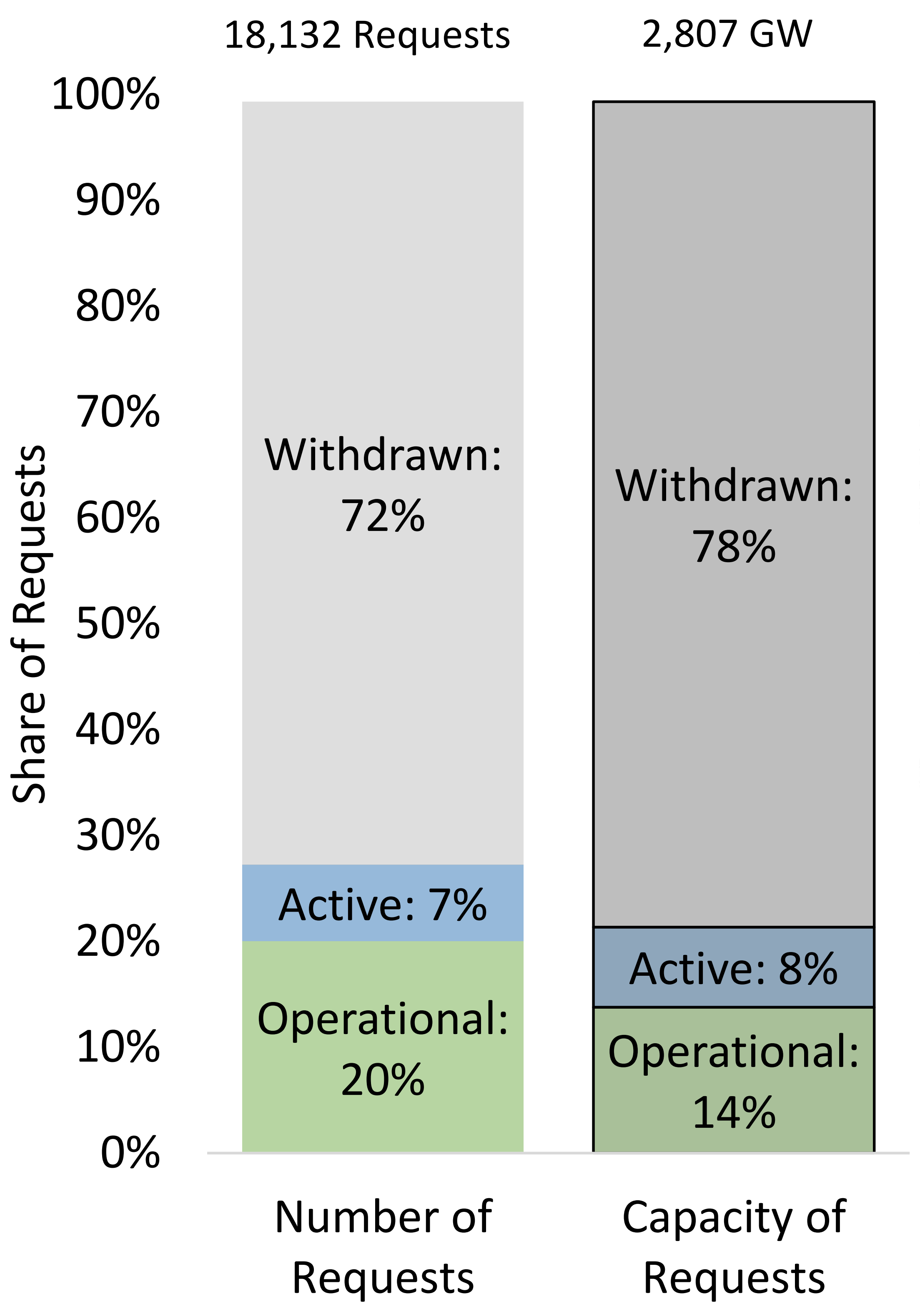
AI electricity demand growth outpaces power grid development

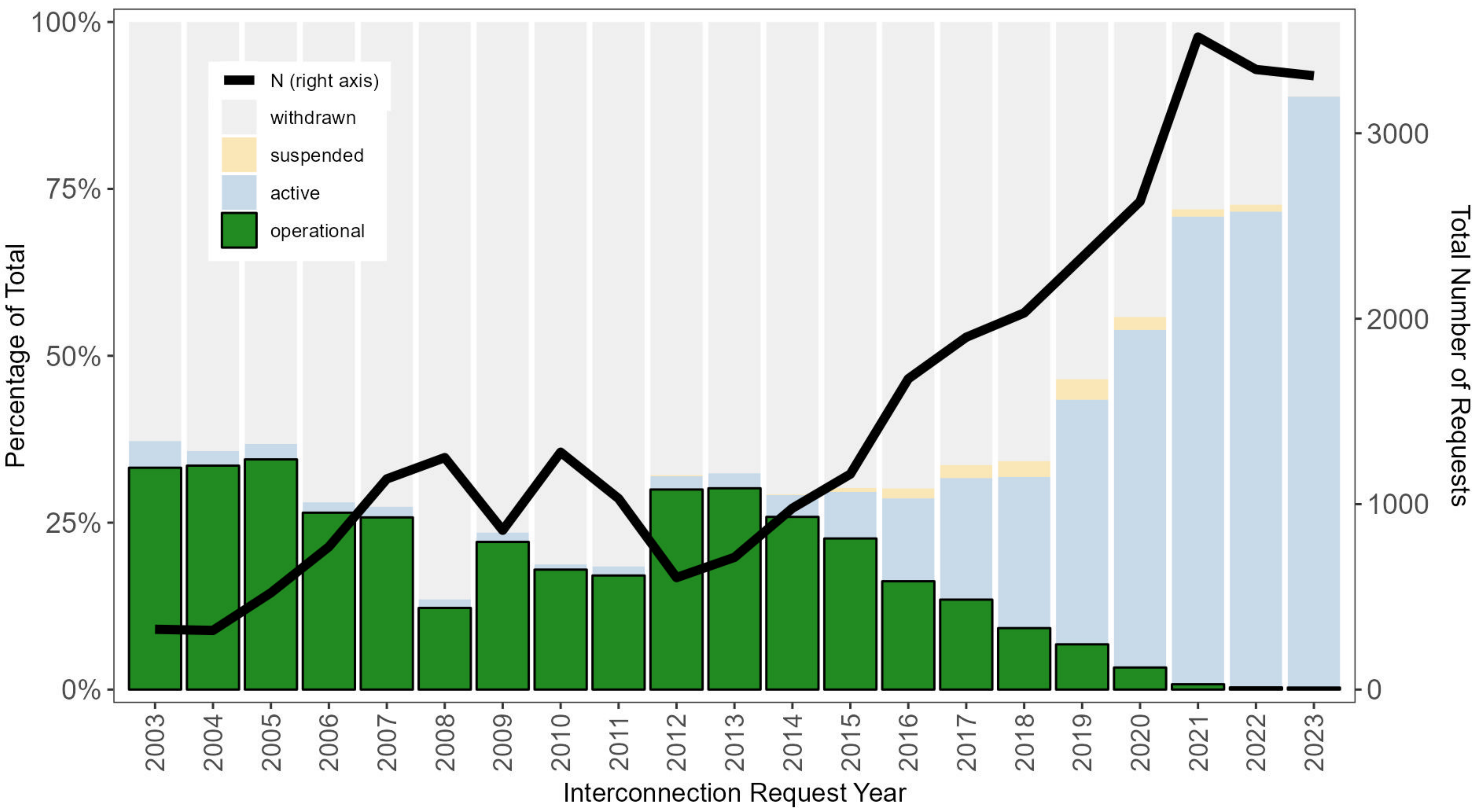


Lawrence Berkeley NL. 2024 Edition Characteristics of Power Plants Seeking Transmission Interconnection As of the End of 2023.

- Grid operators require generation projects to undergo a series of impact studies before they can be built
- The total capacity in the queues is growing year-over-year, but the impact studies do not keep up

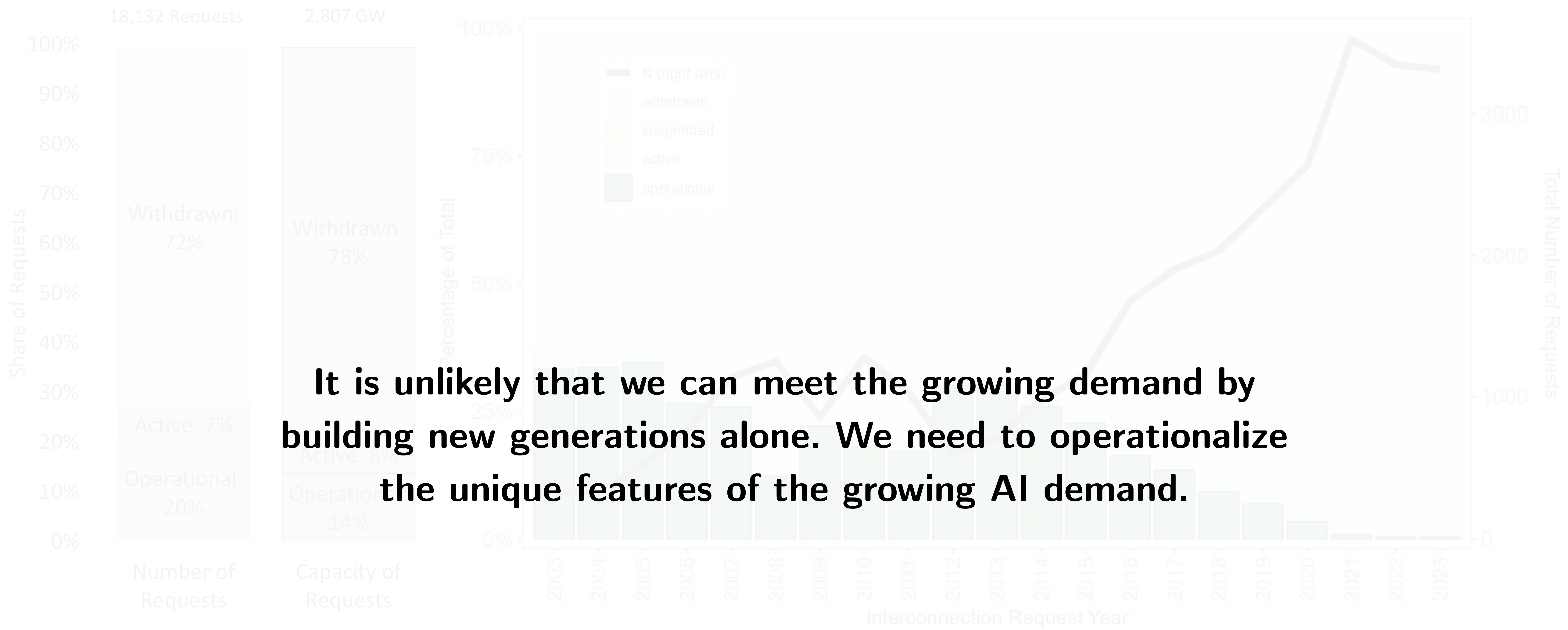
AI electricity demand growth outpaces power grid development





Lawrence Berkeley NL. 2024 Edition Characteristics of Power Plants Seeking Transmission Interconnection As of the End of 2023.

- The majority ($> 70\%$) of interconnection requests are withdrawn
- Just 20% of requests (14% of capacity) submitted from 2000-2018 had been built as of the end of 2023
- Power grids expansion does not keep with the pace of AI electricity growth



It is unlikely that we can meet the growing demand by building new generations alone. We need to operationalize the unique features of the growing AI demand.

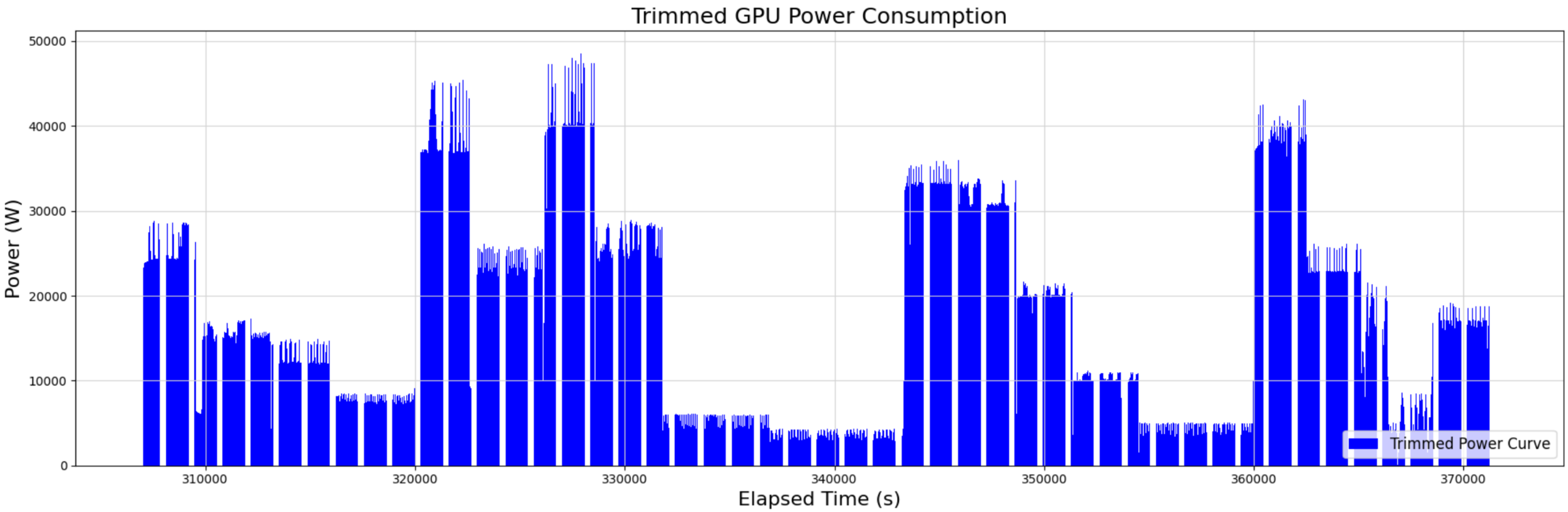
Lawrence Berkeley NL. 2024 Edition Characteristics of Power Plants Seeking Transmission Interconnection As of the End of 2023.

- The majority (> 70%) of interconnection requests are withdrawn
- Just 20% of requests (14% of capacity) submitted from 2000-2018 had been built as of the end of 2023
- Power grids expansion does not keep with the pace of AI electricity growth

Disruptive electricity consumption patterns

Li, Yuzhuo, et al. "The Unseen AI Disruptions for Power Grids: LLM-Induced Transients." *arXiv preprint arXiv:2409.11416* (2024).

Disruptive electricity consumption patterns

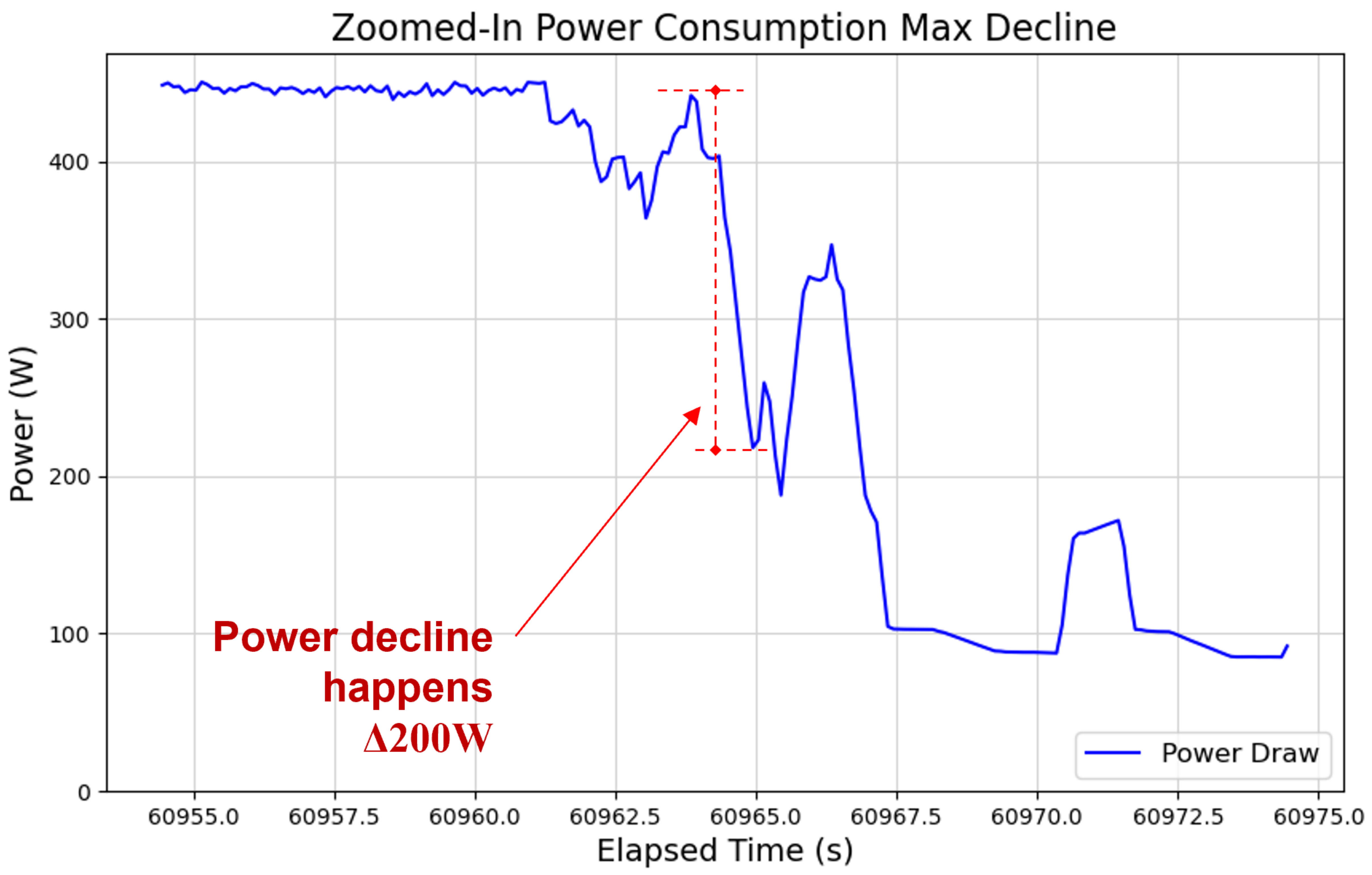


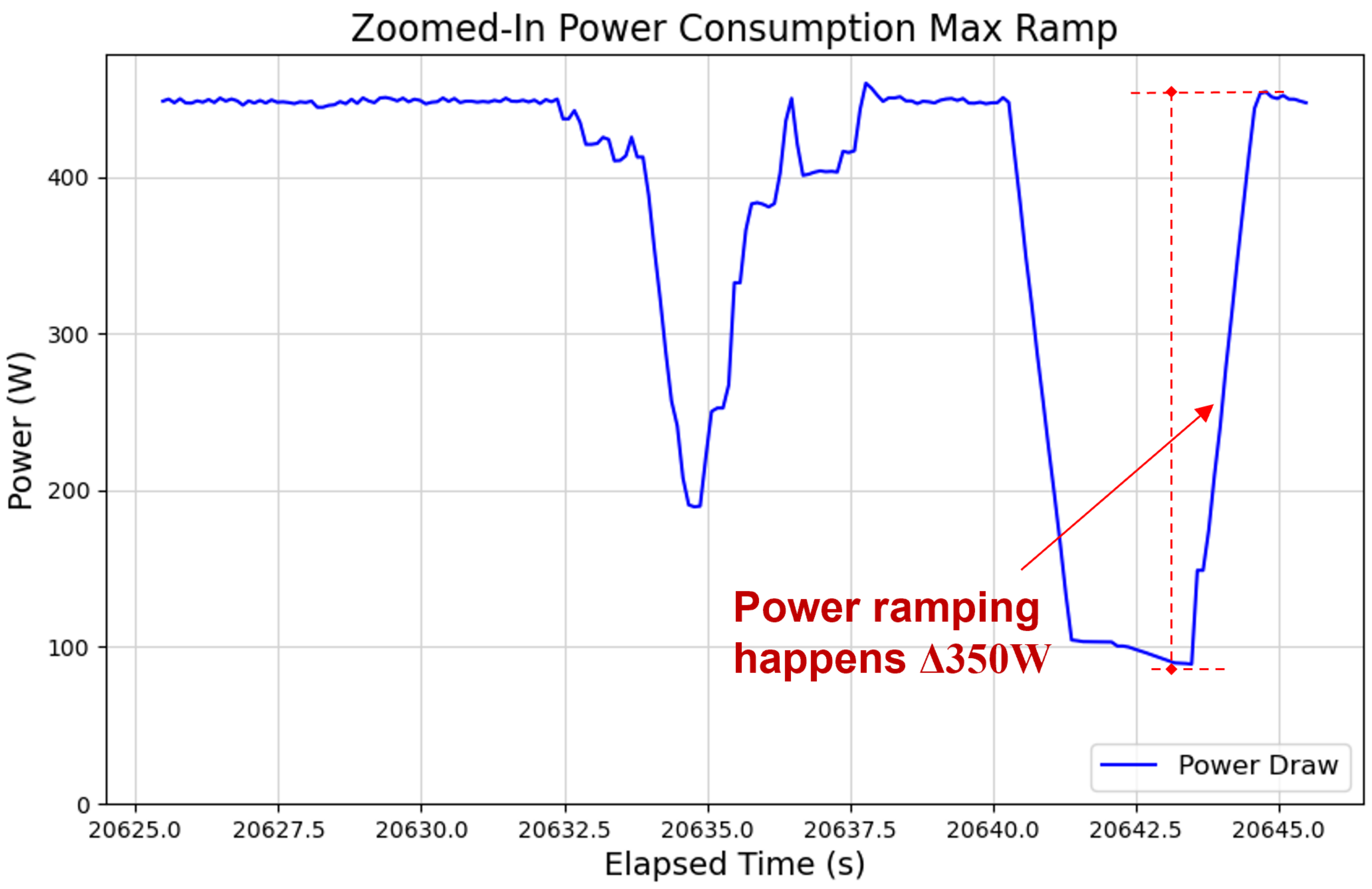
Power Consumption of BERT in MIT Supercloud Dataset

Li, Yuzhuo, et al. "The Unseen AI Disruptions for Power Grids: LLM-Induced Transients." *arXiv preprint arXiv:2409.11416* (2024).

What makes AI such a special load?

Disruptive electricity consumption patterns

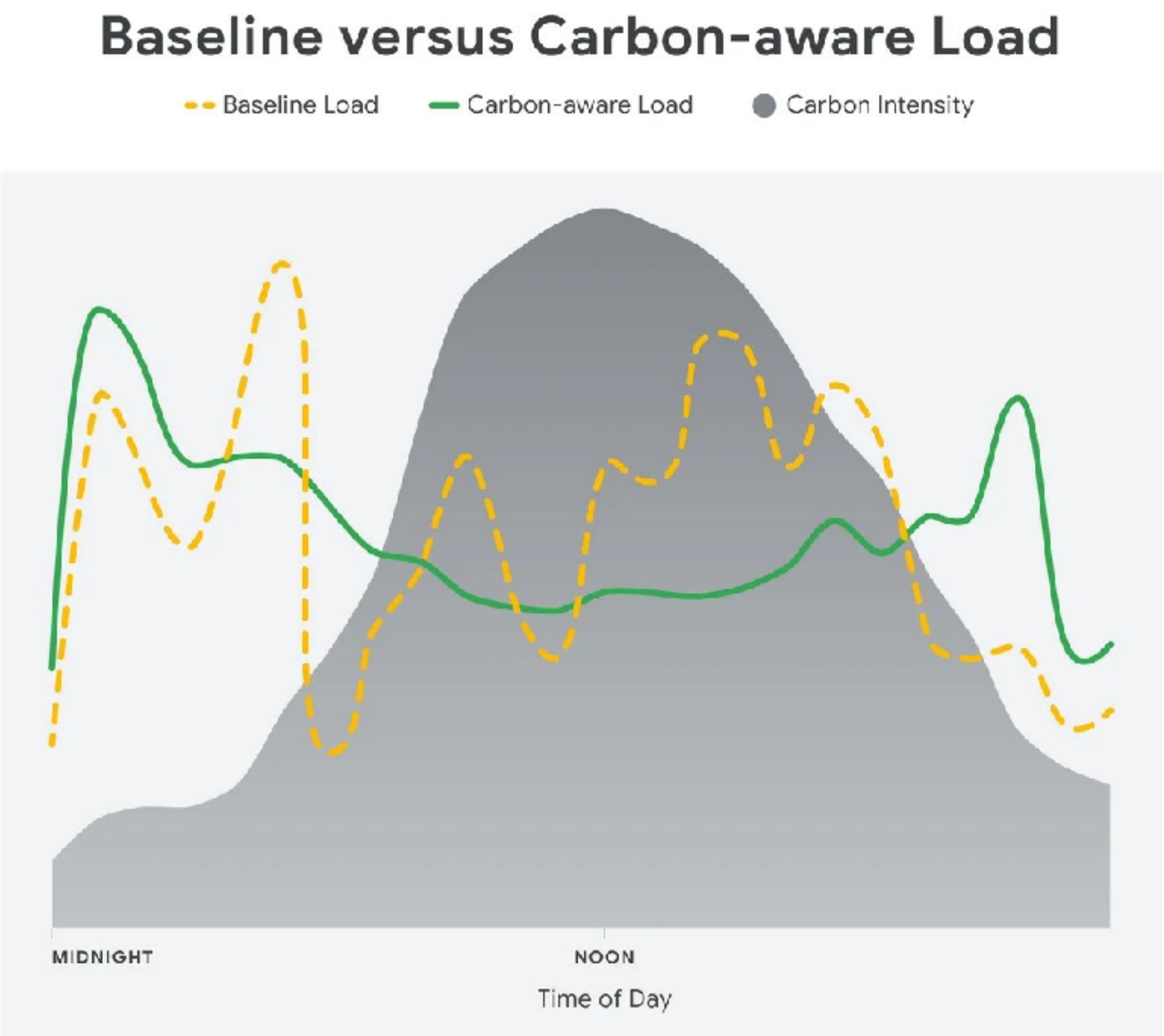




Power Transients of GPT-2 124M

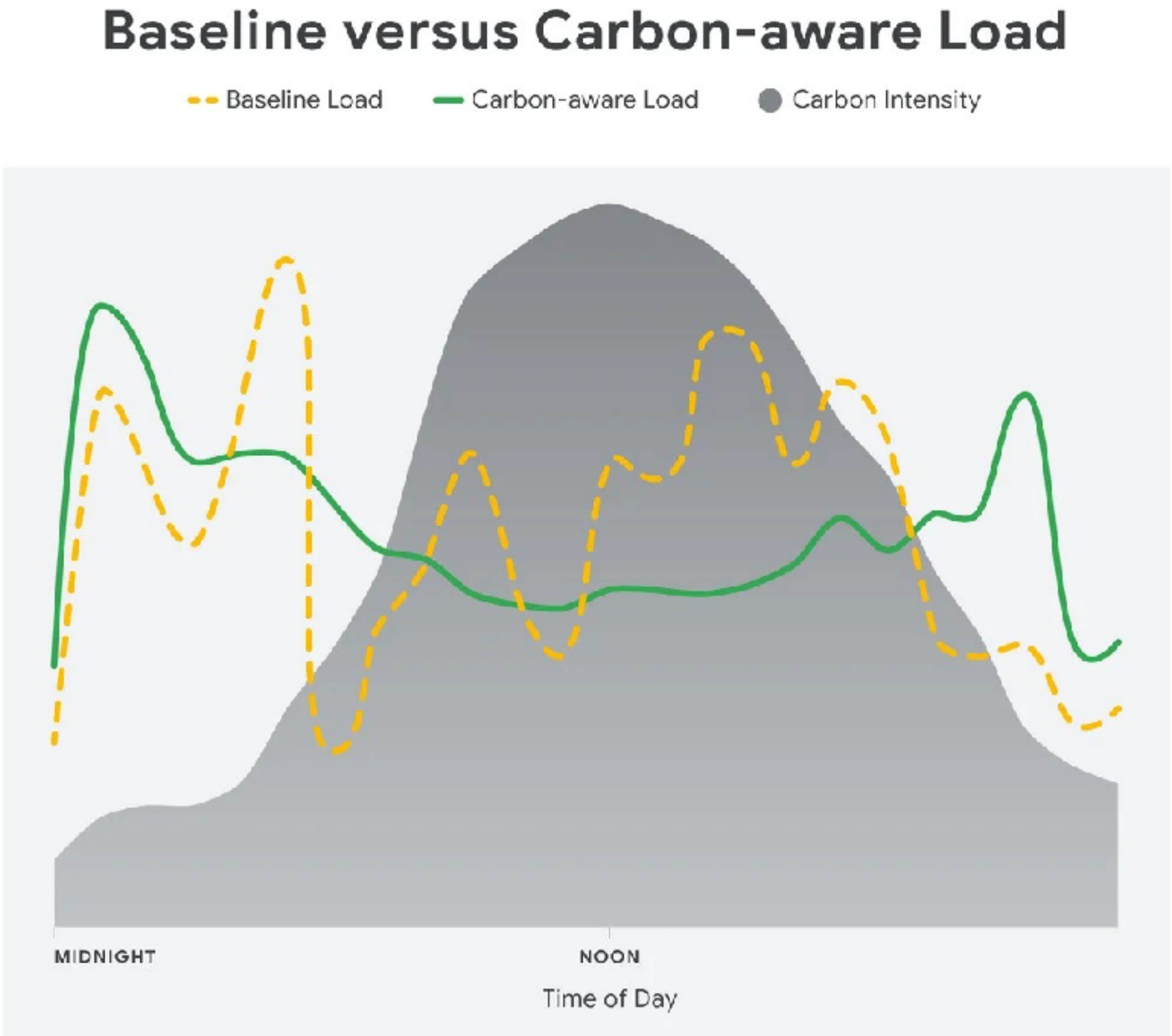
Li, Yuzhuo, et al. "The Unseen AI Disruptions for Power Grids: LLM-Induced Transients." *arXiv preprint arXiv:2409.11416* (2024).

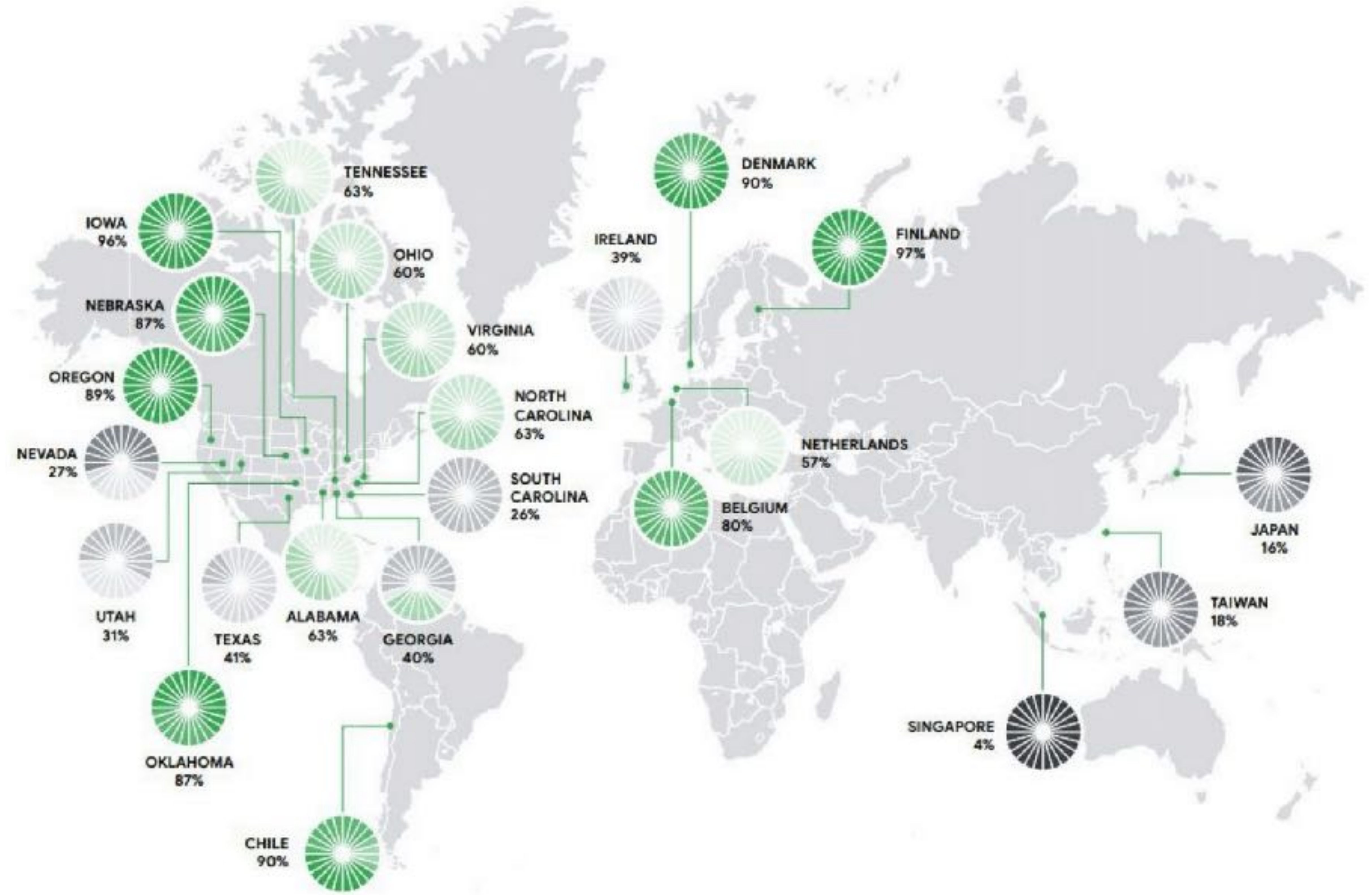
Temporal and geospatial load flexibility



Visualizations are from Google carbon-aware computing reports

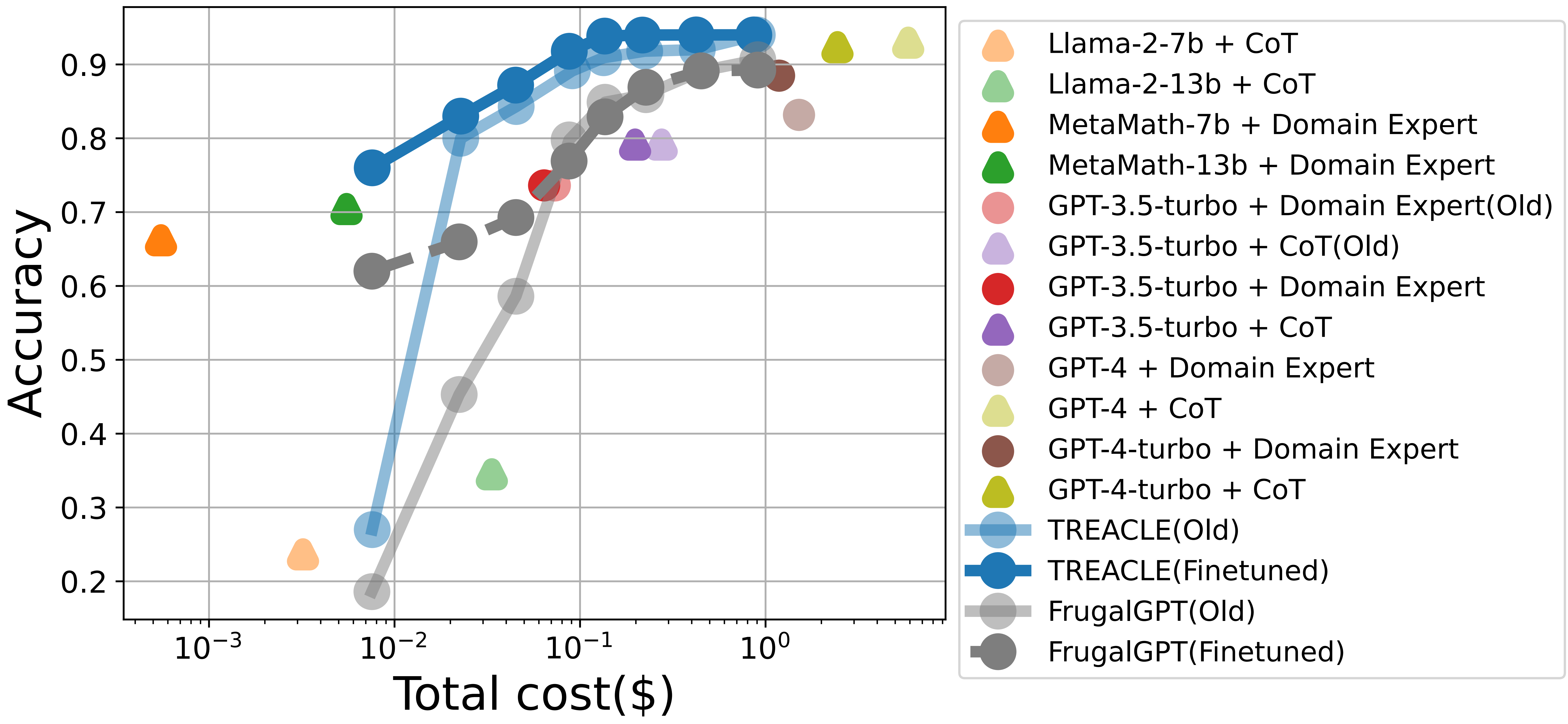
Temporal and geospatial load flexibility





Visualizations are from Google carbon-aware computing reports

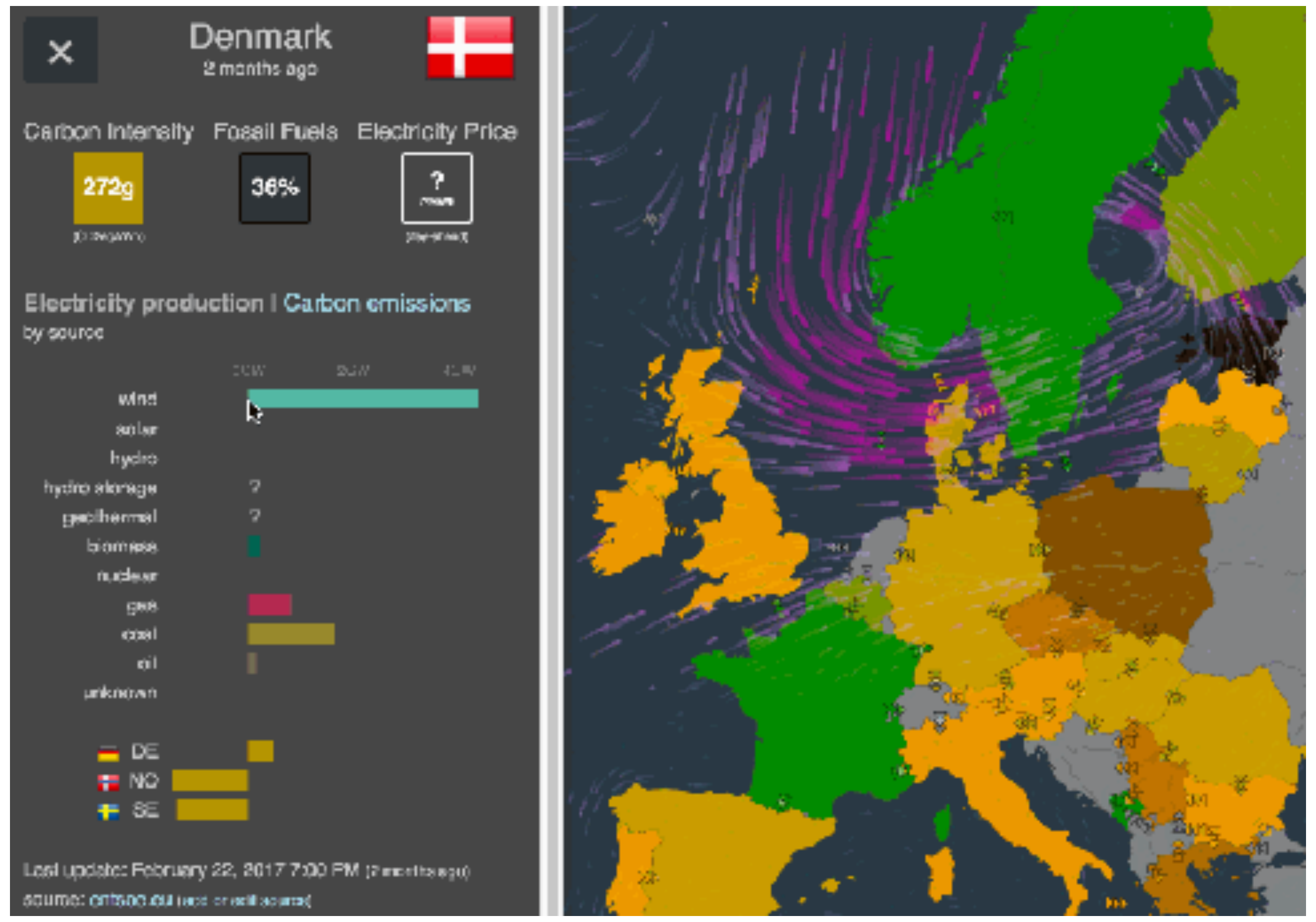
LLM service can be instantaneously delivered by multiple vendors: random job allocation in the grid, depending on user preferences.



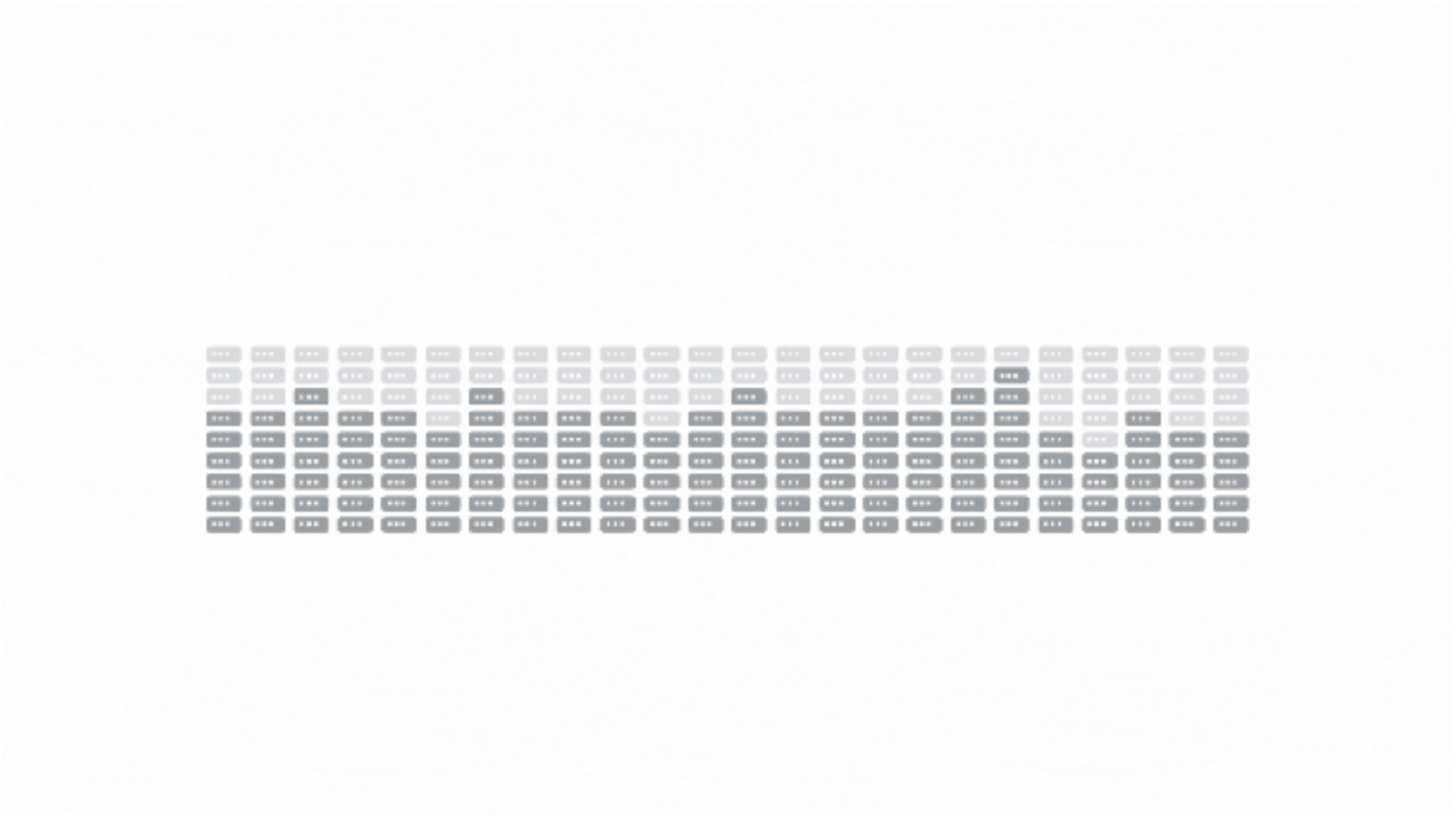
Zhang, Xuechen, et al. TREACLE: Thrifty Reasoning via Context-Aware LLM and Prompt Selection. 2024

How to coordinate power grids and data centers, considering such a disruptive nature of AI loads?

- Data center and grid co-design[1]:
  - Optimal sizing, siting and timing of data centers in the grid
  - Delivers benefits in the long run, but not in the short run
- New electricity market design[2]:
  - Definition of new market products is difficult (e.g., pricing geospatial shifts)
  - Mostly bilateral, out of market contracts. Microsoft & Three Mile Island
- Demand response participation[3]
  - Currently a working solution, yet with limited scalability
- Co-optimization of grid and data-center operations:
  - Ideal yet unattainable in practice solution
  - Significant privacy concerns and computational requirements



https://app.electricitymaps.com/map



[1]Y. Abdennadher et al. Carbon Efficient Placement of Data Center Locations. 2022
[2]W. Zhang et al. Flexibility from networks of data centers: A market clearing formulation with virtual links. 2020
[3]https://cloud.google.com/blog/products/infrastructure/using-demand-response-to-reduce-data-center-power-consumption

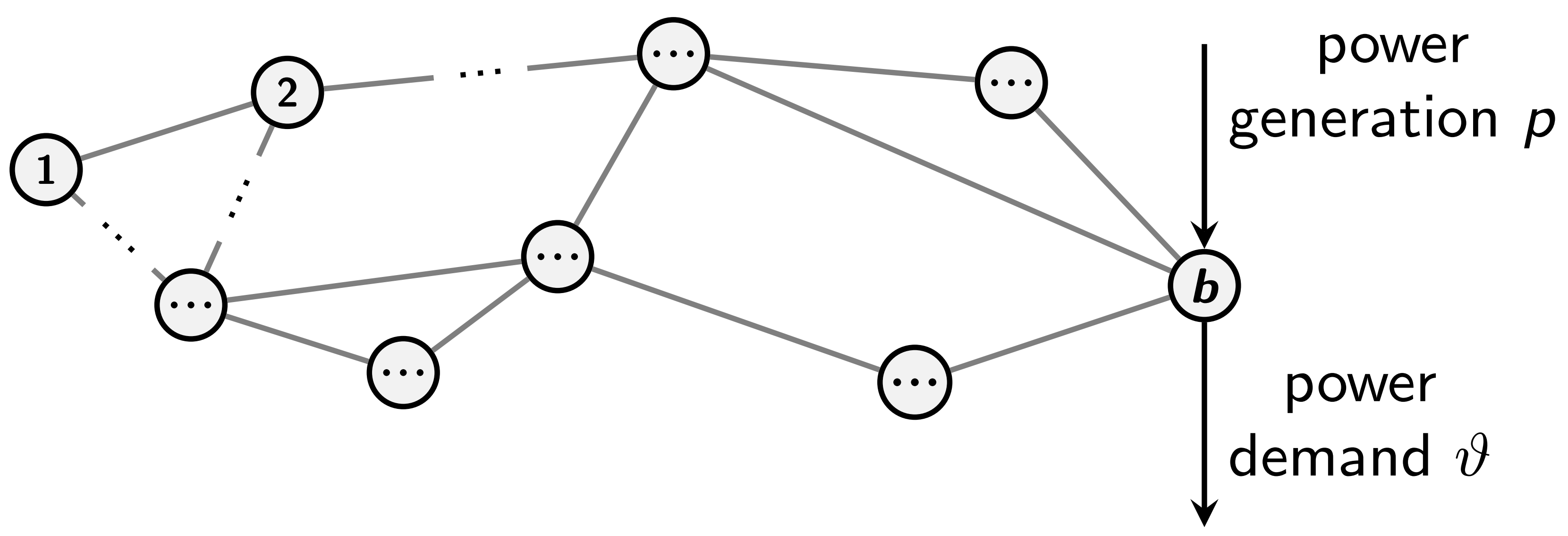
- Data center and grid co-design[1]:
  - Optimal sizing, siting and timing of data centers in the grid
  - Delivers benefits in the long run, but not in the short run
- New electricity market design[2]:
  - Definition of new market products is difficult (e.g., pricing geospatial shifts)
  - Mostly bilateral, out of market contracts. Microsoft & Three Mile Island
- Demand response participation[3]
  - Currently a working solution, yet with limited scalability
- Co-optimization of grid and data-center operations:
  - Ideal yet unattainable in practice solution
  - Significant privacy concerns and computational requirements

https://app.electricitymaps.com/map

[1] Y. Abdennadher et al. Carbon Efficient Placement of Data Center Locations. 2022
[2] W. Zhang et al. Flexibility from networks of data centers: A market clearing formulation with virtual links. 2020
[3] https://cloud.google.com/blog/products/infrastructure/using-demand-response-to-reduce-data-center-power-consumption

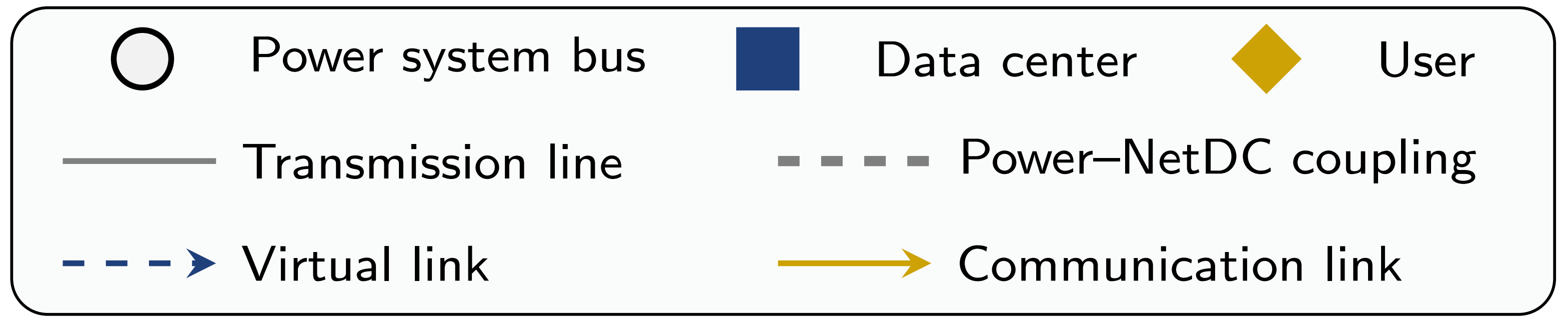
Co-optimization of grid and data-center operations

Co-optimization of networked data centers (NetDC) and power grid operations

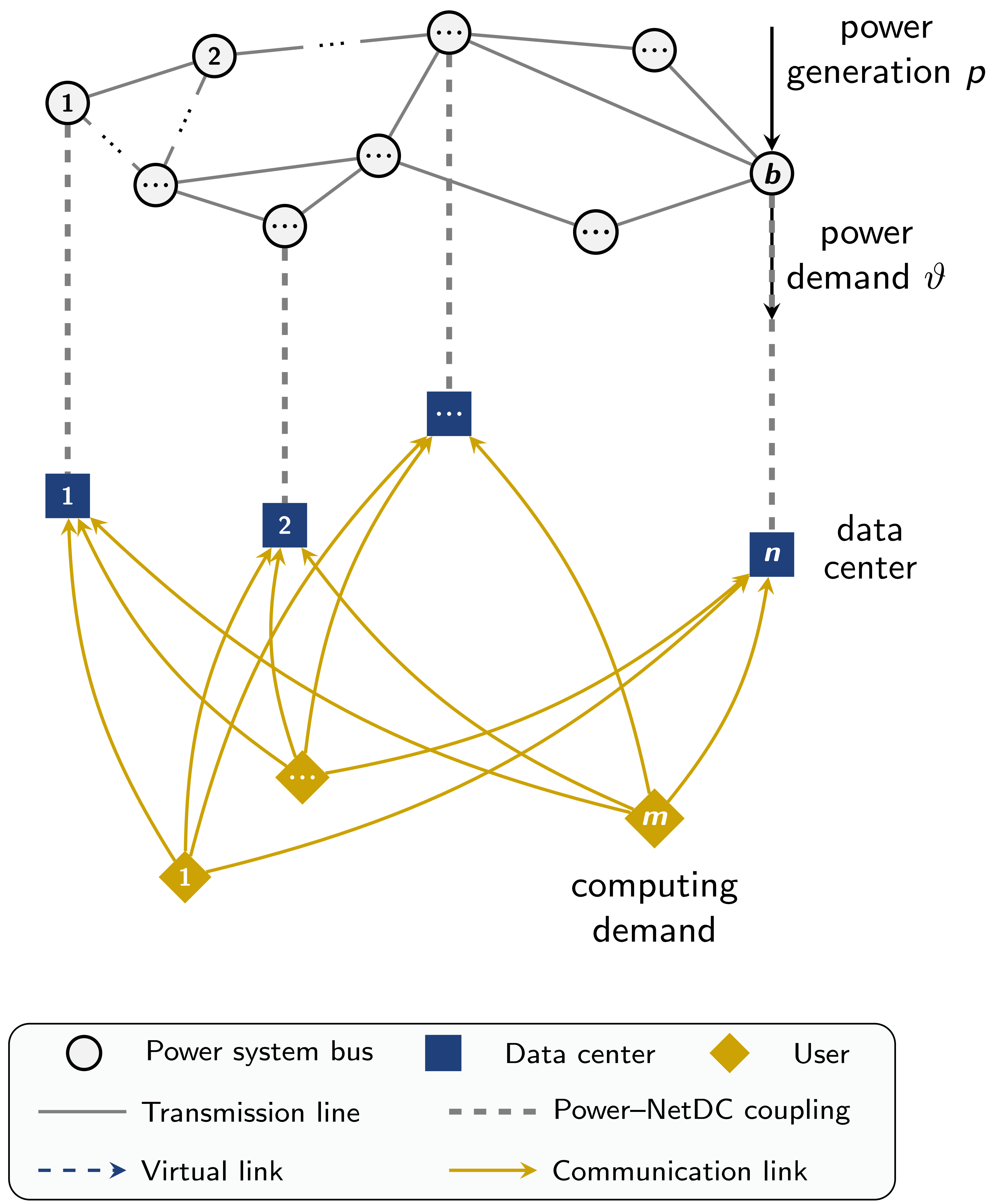


Power grid optimization problem:

$$\begin{aligned} \underset{p}{\text{minimize}} \quad & c_{\text{pwr}}(p) && \triangleright \texttt{ Dispatch cost} \\ \text{subject to} \quad & p \in \mathcal{P}_{\text{pwr}}(\vartheta) && \triangleright \texttt{ Grid feasibility} \end{aligned}$$



Co-optimization of networked data centers (NetDC) and power grid operations



Power grid optimization problem:

$$\underset{p}{\text{minimize}} \quad c_{\text{pwr}}(p) \qquad \triangleright \text{ Dispatch cost}$$

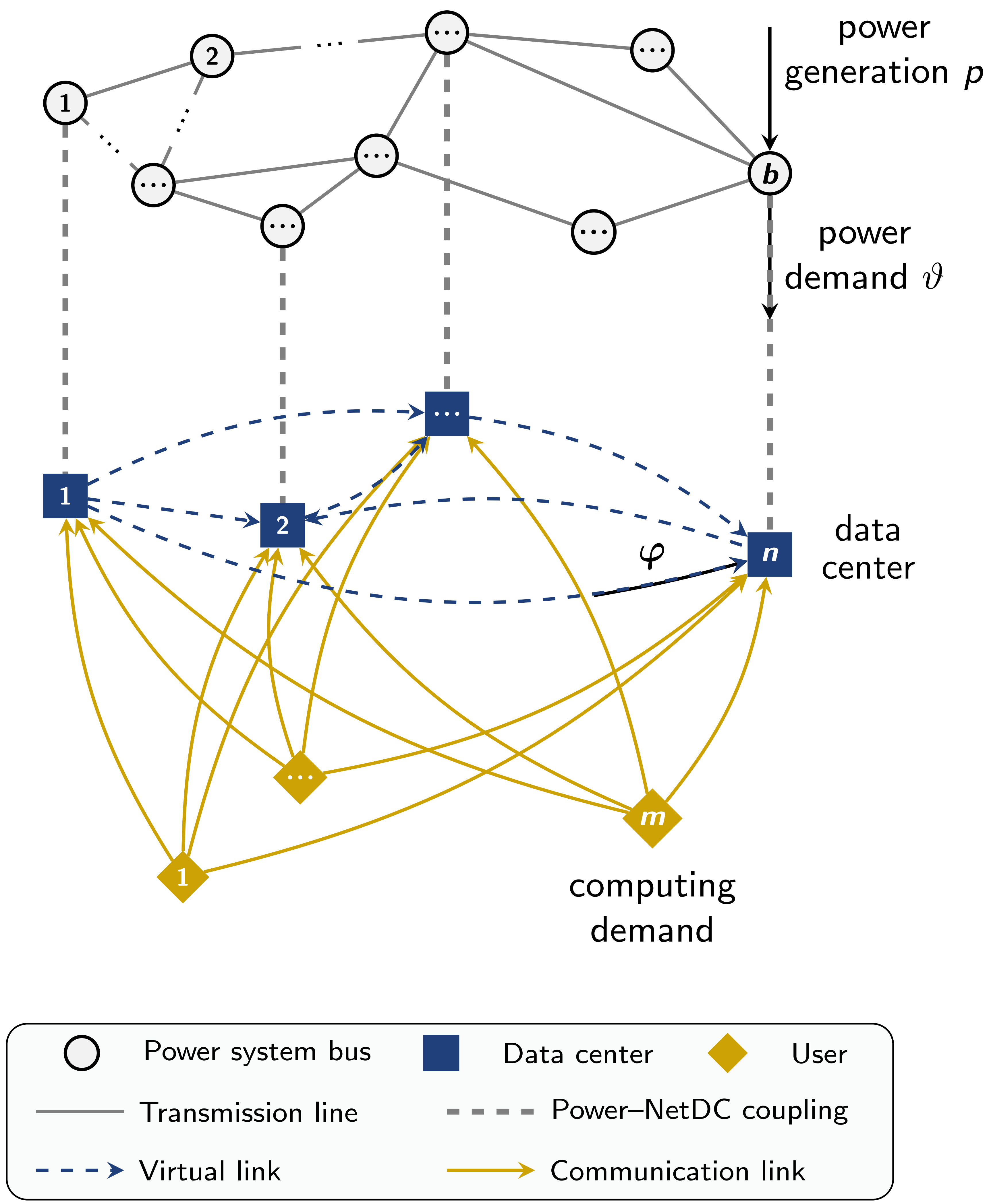
$$\text{subject to} \quad p \in \mathcal{P}_{\text{pwr}}(\vartheta) \qquad \triangleright \text{ Grid feasibility}$$

Data centers optimization problem:

$$\underset{\vartheta}{\text{minimize}} \quad c_{\text{net-dc}}(\vartheta) \qquad \triangleright \text{ Latency loss}$$

$$\text{subject to} \quad \vartheta \in \mathcal{W}_{\text{net-dc}}(\varphi) \qquad \triangleright \text{ NetDC feasibility}$$

Co-optimization of networked data centers (NetDC) and power grid operations

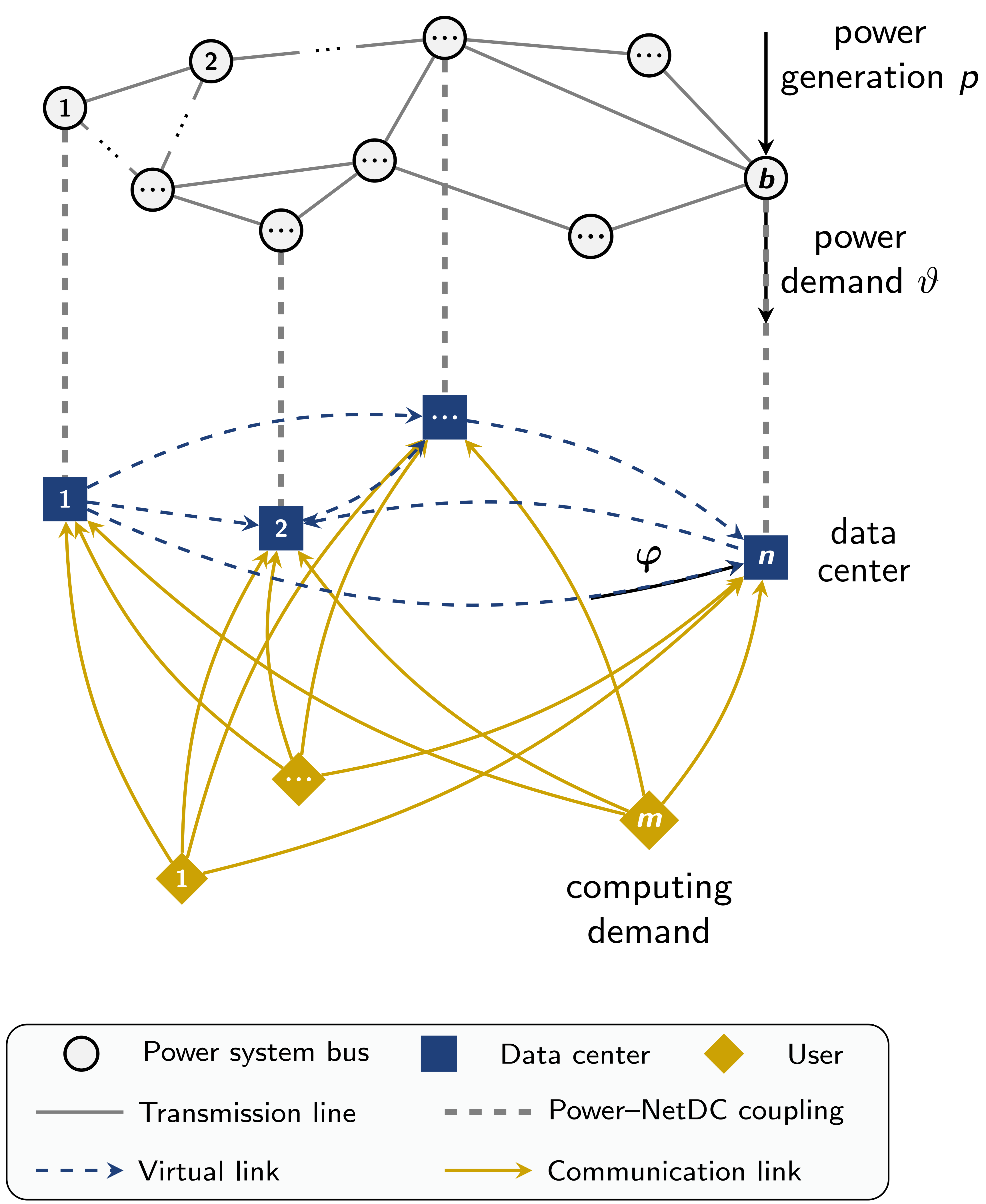


Power grid optimization problem:

$$\begin{aligned} \underset{p}{\text{minimize}} \quad & c_{\text{pwr}}(p) & \triangleright \texttt{ Dispatch cost} \\ \text{subject to} \quad & p \in \mathcal{P}_{\text{pwr}}(\vartheta) & \triangleright \texttt{ Grid feasibility} \end{aligned}$$

Data centers optimization problem:

$$\begin{aligned} \underset{\vartheta}{\text{minimize}} \quad & c_{\text{net-dc}}(\vartheta) & \triangleright \texttt{ Latency loss} \\ \text{subject to} \quad & \vartheta \in \mathcal{W}_{\text{net-dc}}(\varphi) & \triangleright \texttt{ NetDC feasibility} \end{aligned}$$



Power Grid–NetDC coordination problem:

$$\begin{aligned} \underset{\varphi, p}{\text{minimize}} \quad & c_{\text{pwr}}(p) && \triangleright \text{ Dispatch cost} \\ \text{subject to} \quad & p \in \mathcal{P}_{\text{pwr}}(\vartheta) && \triangleright \text{ Grid feasibility} \\ & \underset{\vartheta}{\text{minimize}} \quad c_{\text{net-dc}}(\vartheta) && \triangleright \text{ Latency loss} \\ & \text{subject to} \quad \vartheta \in \mathcal{W}_{\text{net-dc}}(\varphi) && \triangleright \text{ NetDC feasibility} \end{aligned}$$

el. demand

computing task shift

Bilevel structure: the grid operator minimizes dispatch costs by optimizing task shifts that reshape data-center electricity demand.

Power-NetDC coordination: From optimization to regression

From optimization to regression for Power-NetDC coordination

- Solving the bilevel problem in real-time is extremely challenging:
  - Power and NetDC needs to be exchanged in real time (privacy barriers)
  - Large-scale bilevel optimization in real time (computationally intractable)

- Instead, we consider a contextual regression policy for real-time coordination

$$\phi(x) = \beta_0 + \beta_1 x$$

where $\beta = (\beta_0, \beta_1)$ are regression weights and x is the vector of real-time coordination features (e.g, locational marginal prices, zonal electricity demand, renewable power generation)

- The baseline training of the regression policy:
  1. Collect a dataset $\{(x_1, \varphi_1^\star), \dots, (x_q, \varphi_q^\star)\}$ of q coordination scenarios.

     Each scenario i includes contextual features x_i and the optimal solution $\varphi_i^\star$ to the coordination problem
  2. Train a contextual regression to map coordination features into the optimal task shifts in real-time

$$\underset{\|\beta\|_1 \leqslant \varepsilon}{\text{minimize}} \quad \frac{1}{q}\sum_{i=1}^{q} \|\beta_0 + \beta_1 x_i - \varphi_i^\star\|_2^2$$

← *Minimum prediction loss, but no guarantees on policy cost-optimality and feasibility*

feature selection ↗

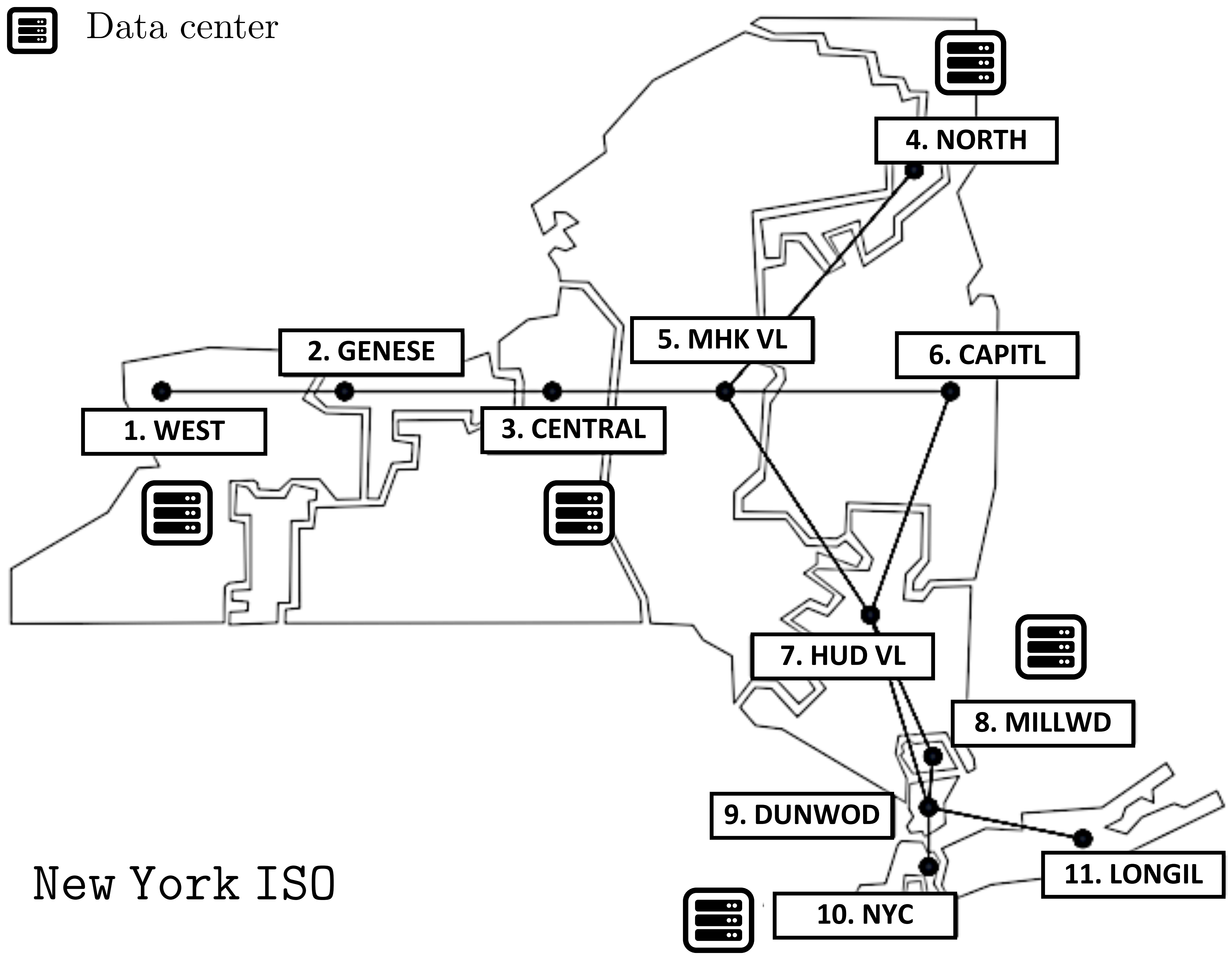
AgentCONCUR protocol with performance guarantees

AgentCONCUR: Ensuring feasibility of the contextual regression

$$\begin{aligned}
\underset{\beta,\varphi_i,\dots,\varphi_q,p_i,\dots,p_q}{\text{minimize}} \quad & \frac{1}{q}\sum_{i=1}^{q} c_{\text{pwr}}(p_i) && \triangleright \texttt{ Average dispatch cost} \\
\text{subject to} \quad & p_i \in \mathcal{P}_{\text{pwr}}(\vartheta_i^{\star}), \quad \forall i = 1,\dots,q && \triangleright \texttt{ Grid equations for each scenario} \\
& \varphi_i = \beta_0 + \beta_1 x_i,\ \|\beta\|_1 \leqslant \varepsilon, \quad \forall i = 1,\dots,q && \triangleright \texttt{ Coupling contextual regression} \\
& \vartheta_i^{\star} \in \underset{\vartheta_i}{\text{minimize}} \quad c_{\text{net-dc}}(\vartheta_i) && \triangleright \texttt{ Latency loss} \\
& \qquad \text{subject to} \quad \vartheta_i \in \mathcal{W}_{\text{net-dc}}(\varphi_i), \quad \forall i = 1,\dots,q && \triangleright \texttt{ NetDC feasibility}
\end{aligned}$$

- The task shifts are restricted to the affine policy of contextual features.
- Optimization anticipates how the affine restriction affects the average OPF costs.
- Feasibility guarantees on the training set $\rightarrow$ also holds on the testing set

Numerical Experiments on the New York Independent System Operator's System

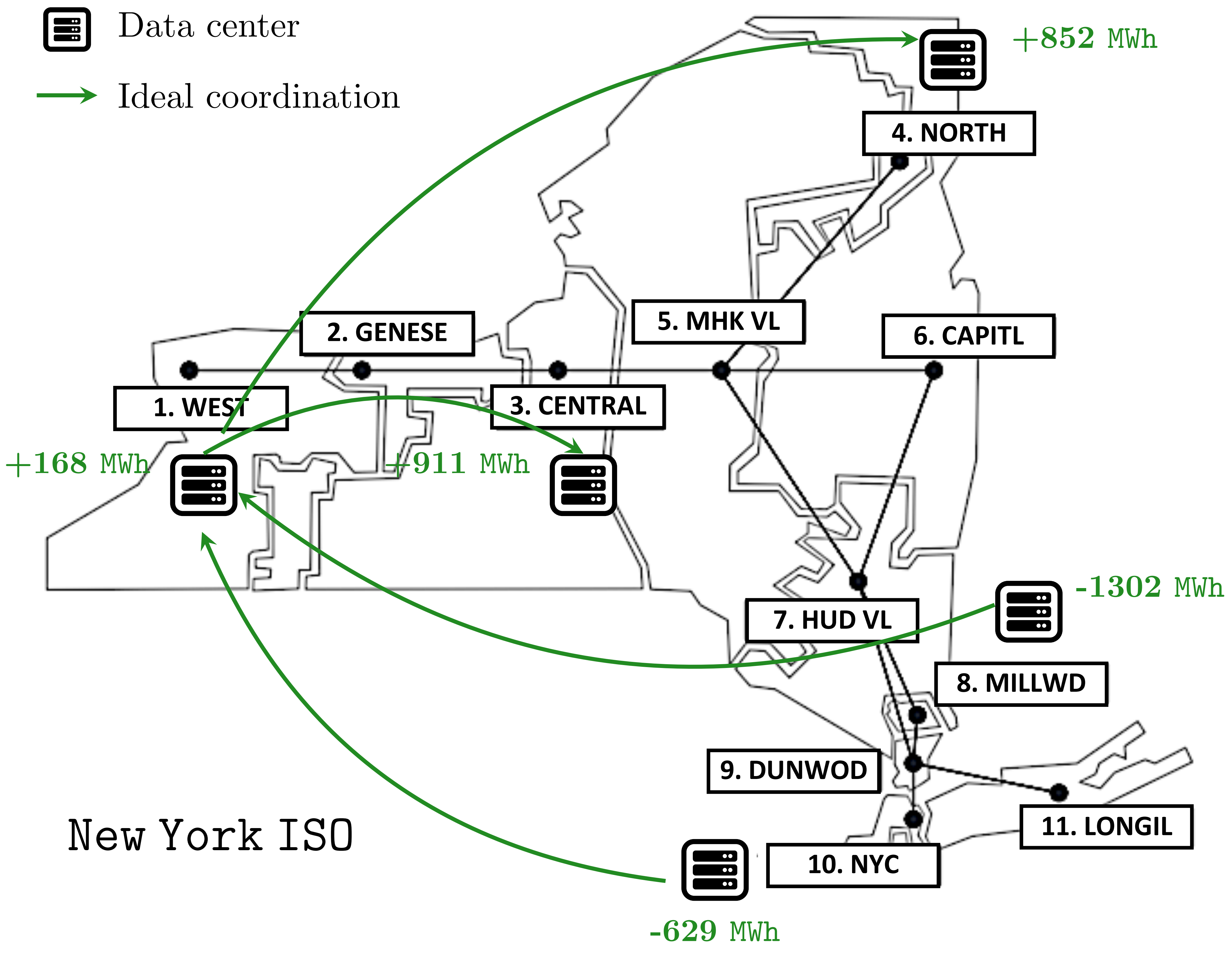


Data inputs:

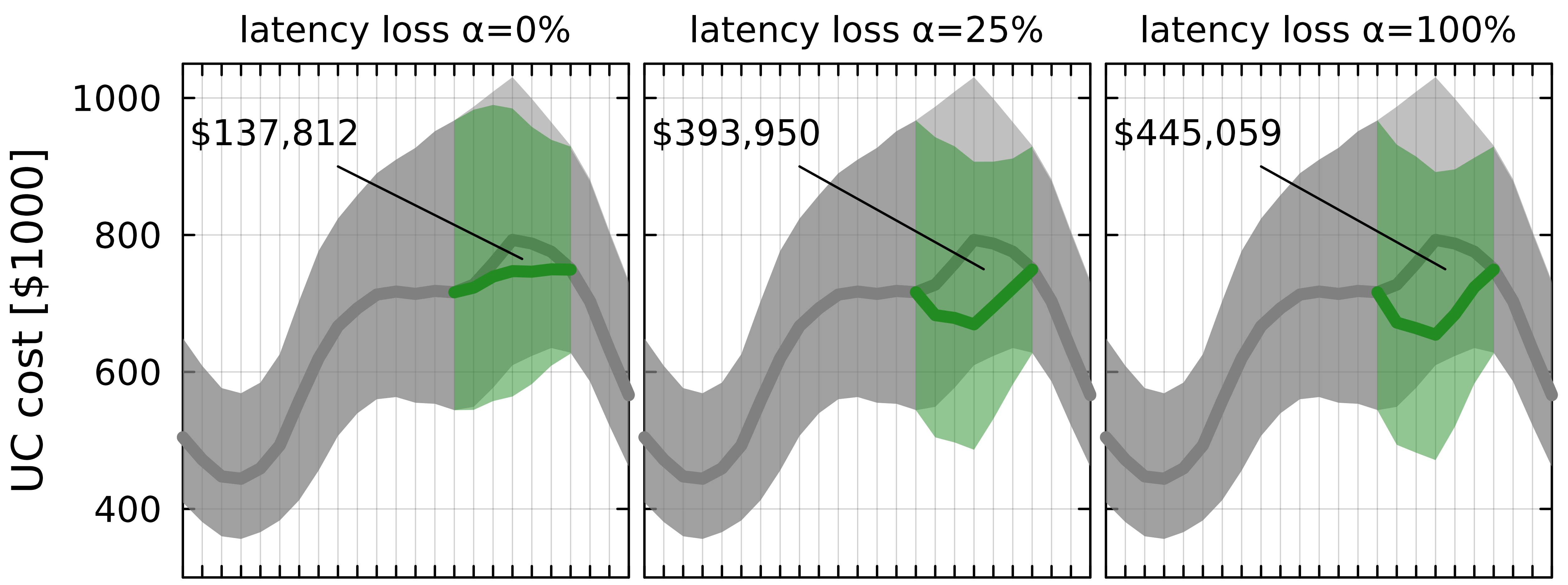
- 11-zone aggregation of the New York ISO
- Network of 5 data centers (10 virtual links)
- Varying demand from 5% to 30% of the peak load

We study two coordination settings:

- Ideal day-ahead coordination with optimization
- Real-time coordination with contextual regression

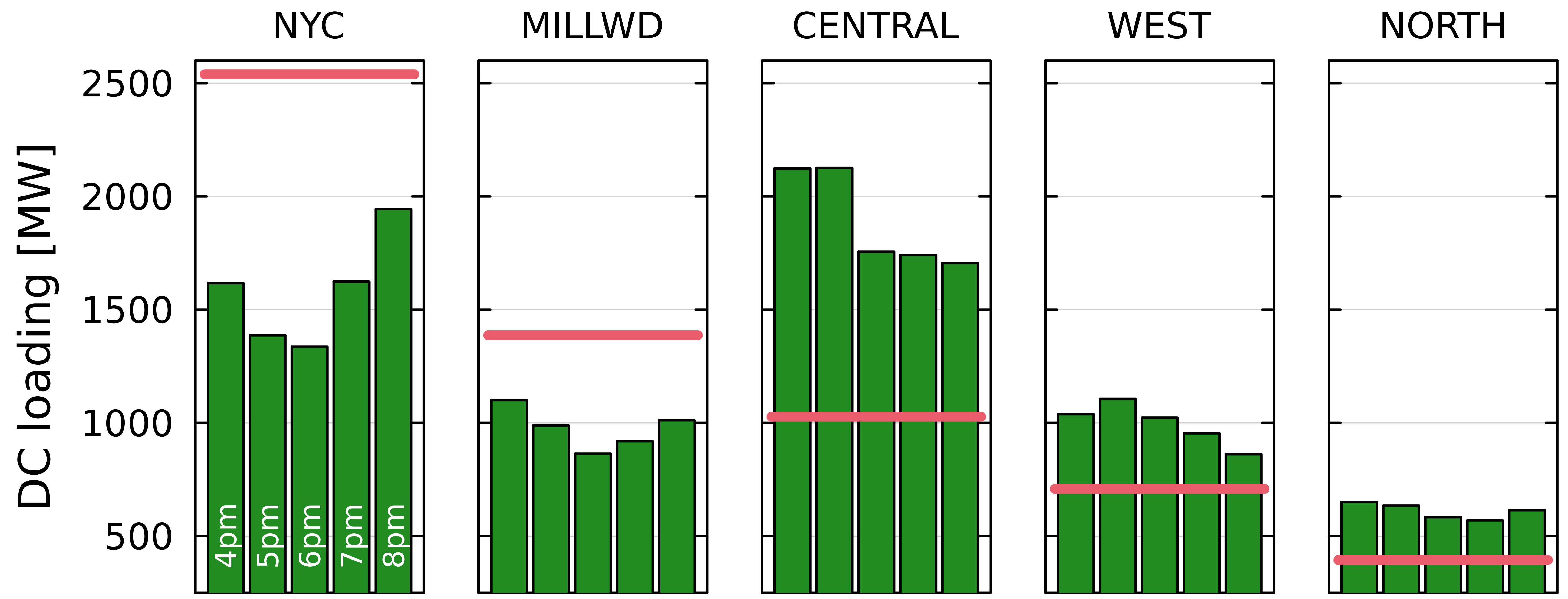


Example of task shifts across the network of data centers

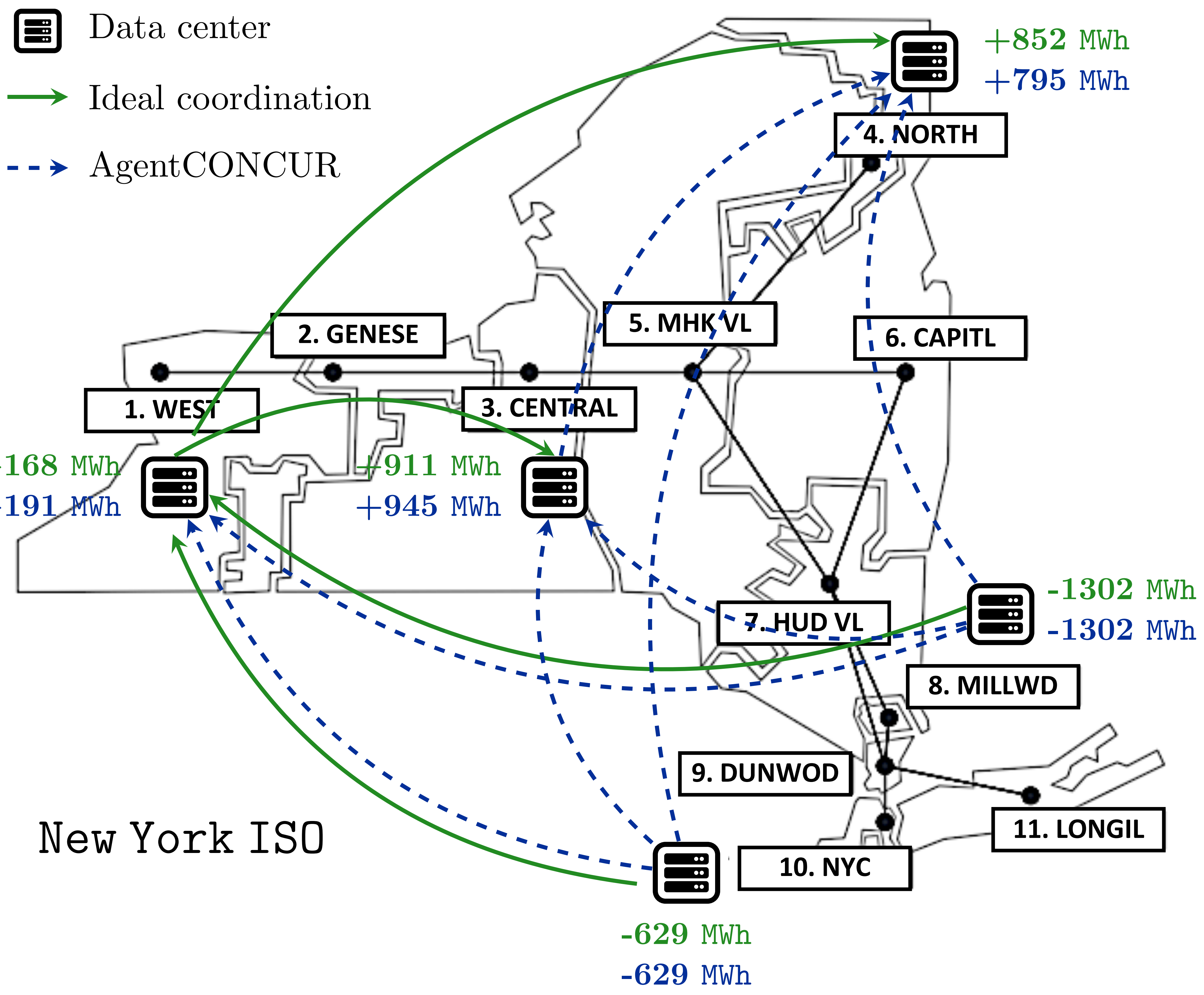


Relaxation of latency constraints ⇒ greater generation cost savings

Unit commitment constraints prevent unlocking the whole NetDC flexibility



The flat (in red) loading profile is re-distributed in space and time



Ideal coordination versus the AgentCONCUR solution

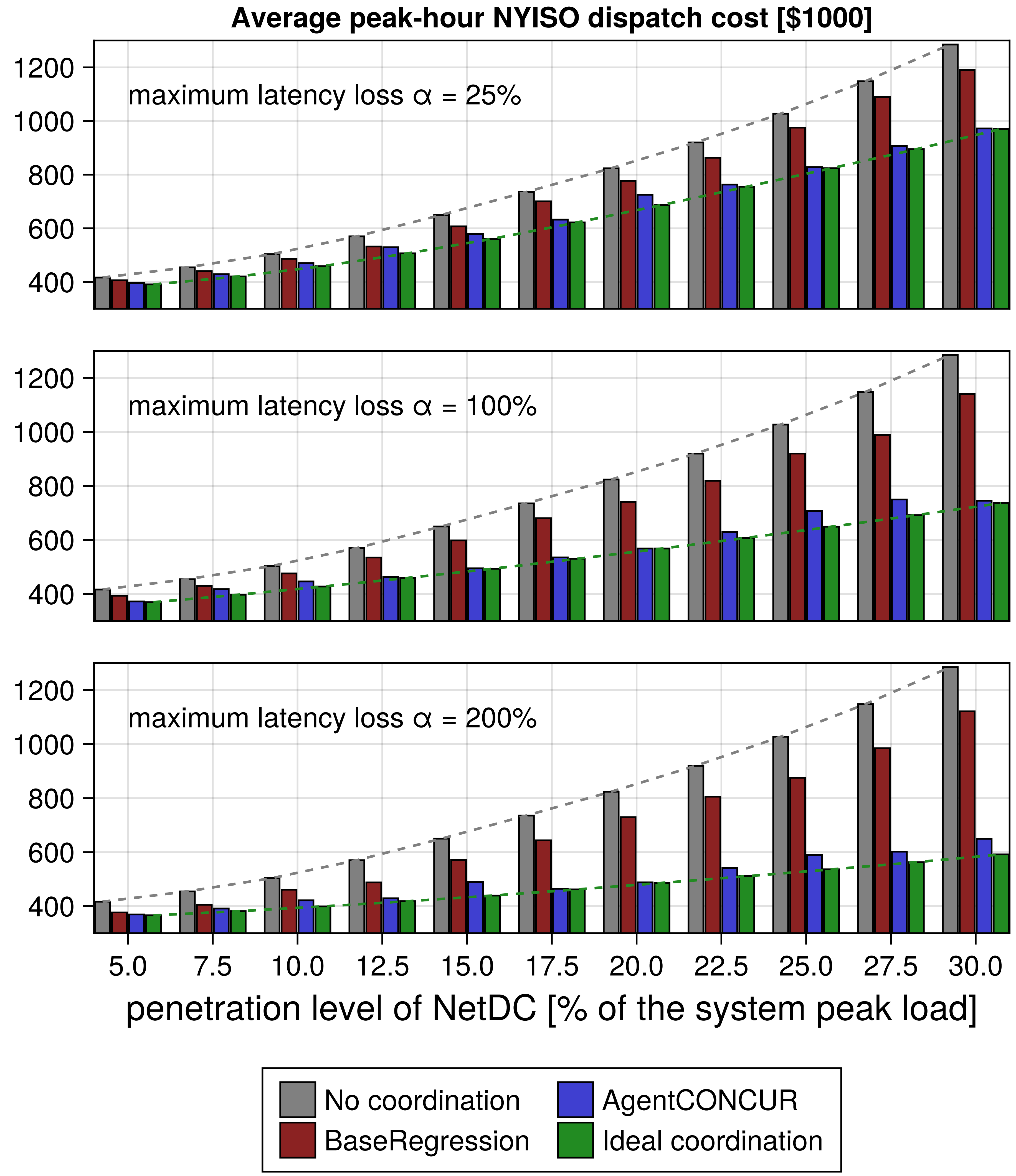
Contextual features from NYISO website

- Zonal real-time electricity demand (d);
- Zonal electricity prices (λ);
- Zonal renewable power generation (r);
- Power flows between aggregation zones (f).

Coordination policy to be optimized offline:

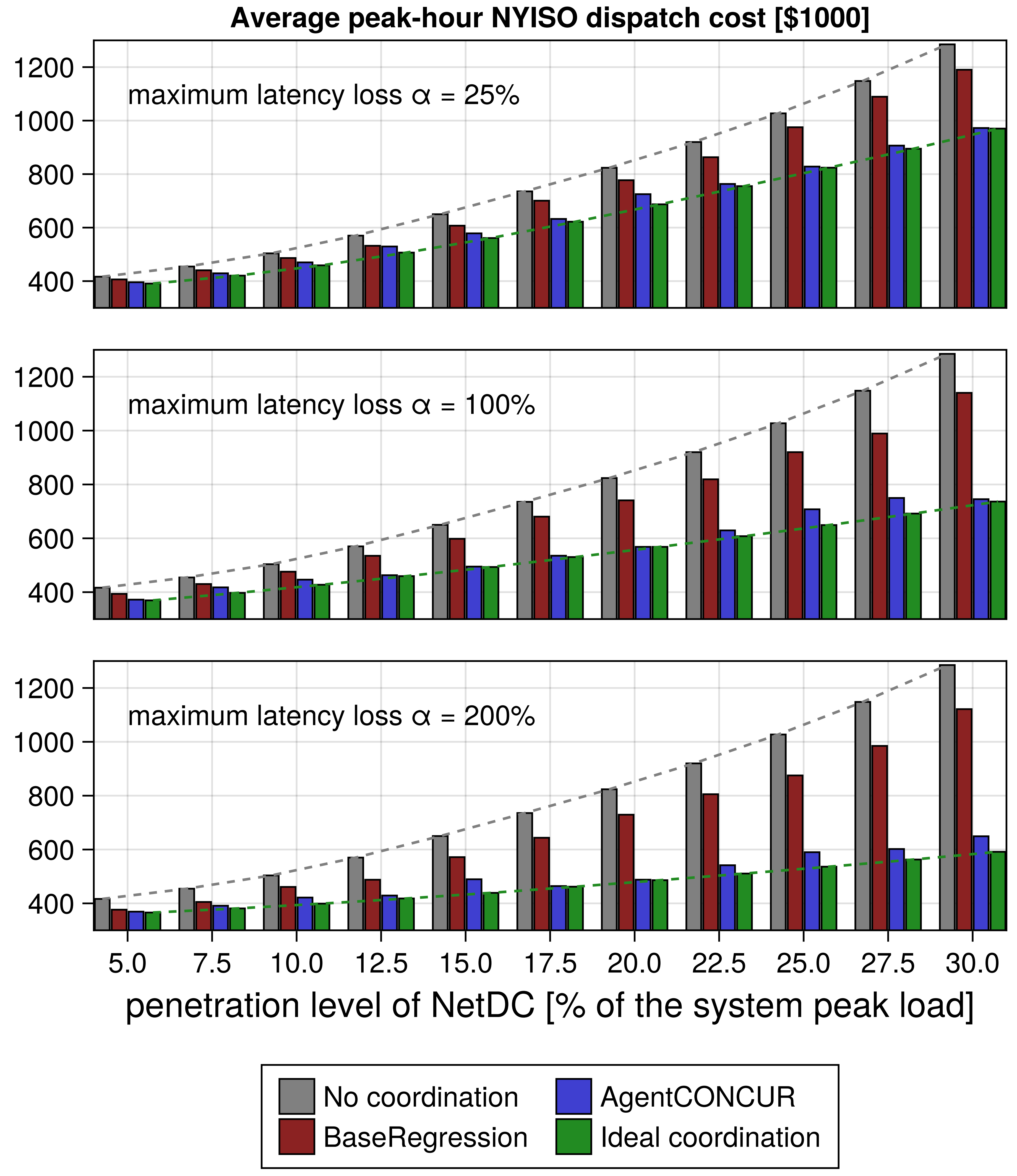
$$\phi \triangleq \beta_0 + \beta_1^d \begin{bmatrix} d_1 \\ \vdots \\ d_{11} \end{bmatrix} + \beta_1^\lambda \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_{11} \end{bmatrix} + \beta_1^r \begin{bmatrix} r_1 \\ \vdots \\ r_{11} \end{bmatrix} + \beta_1^f \begin{bmatrix} f_1 \\ \vdots \\ f_{12} \end{bmatrix}$$

NYISO: Cost-savings of AgentCONCUR

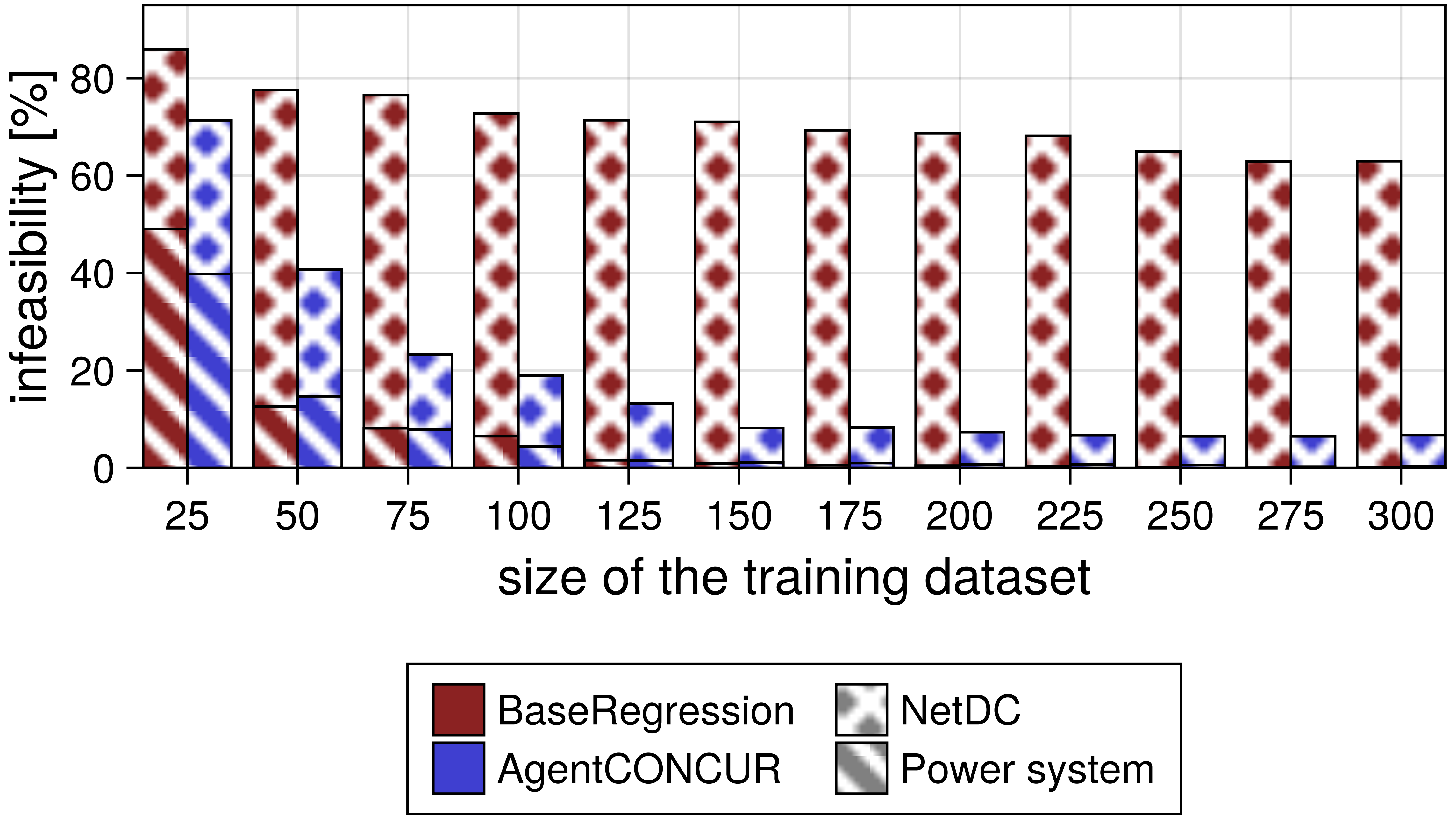


- **Non-coordinated solution** ⇒ quadratic cost growth
- **Ideal coordination** ⇒ more linear cost growth
- **Baseline regression** is often infeasible ⇒ small savings
- Feasibility guarantees of the **AgentCONCUR** protocol ⇒ efficient approximation of the ideal coordination

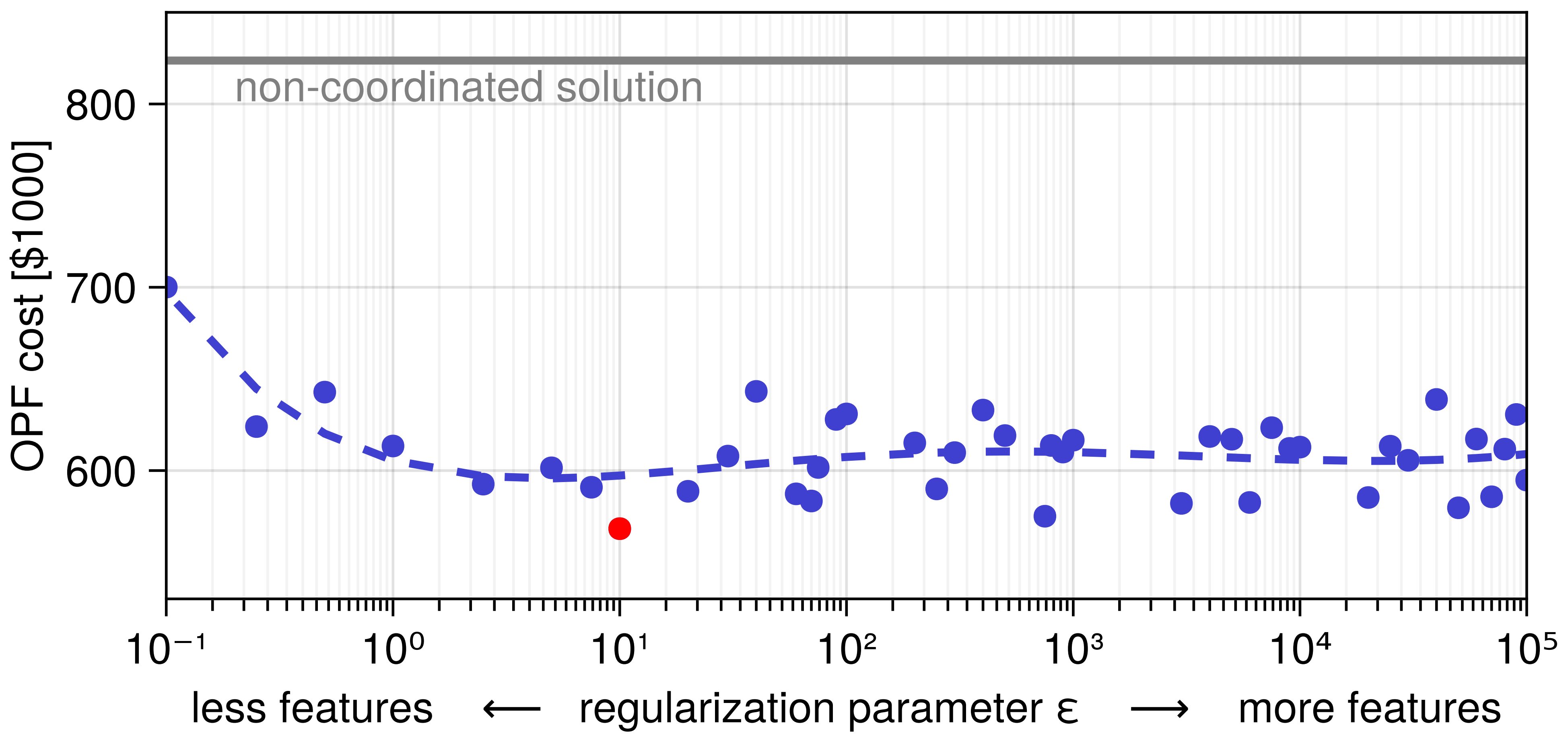
NYISO: Cost-savings of AgentCONCUR



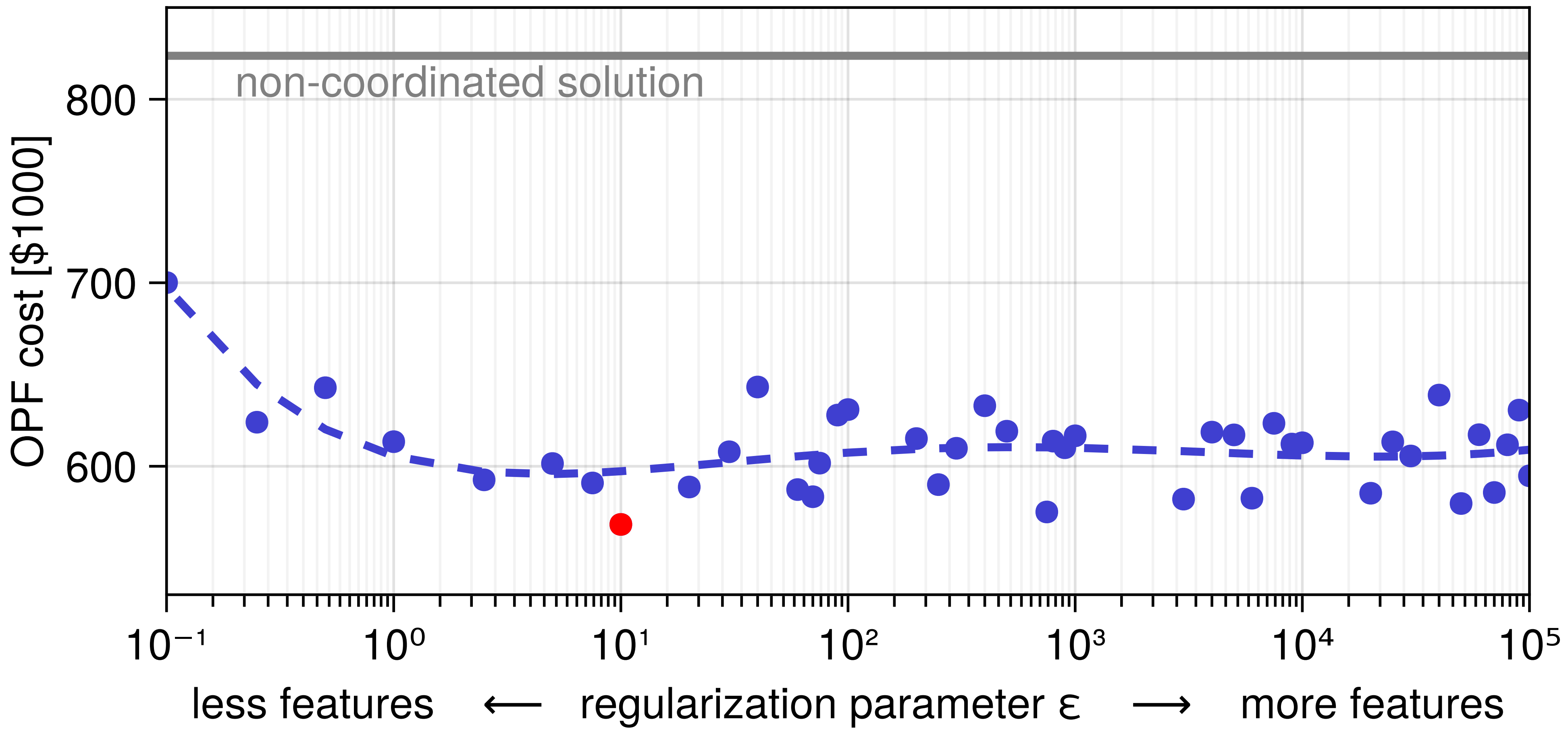
- Non-coordinated solution ⇒ quadratic cost growth
- Ideal coordination ⇒ more linear cost growth
- Baseline regression is often infeasible ⇒ small savings
- Feasibility guarantees of the AgentCONCUR protocol ⇒ efficient approximation of the ideal coordination



NYISO: Coordination feature selection for AgentCONCUR



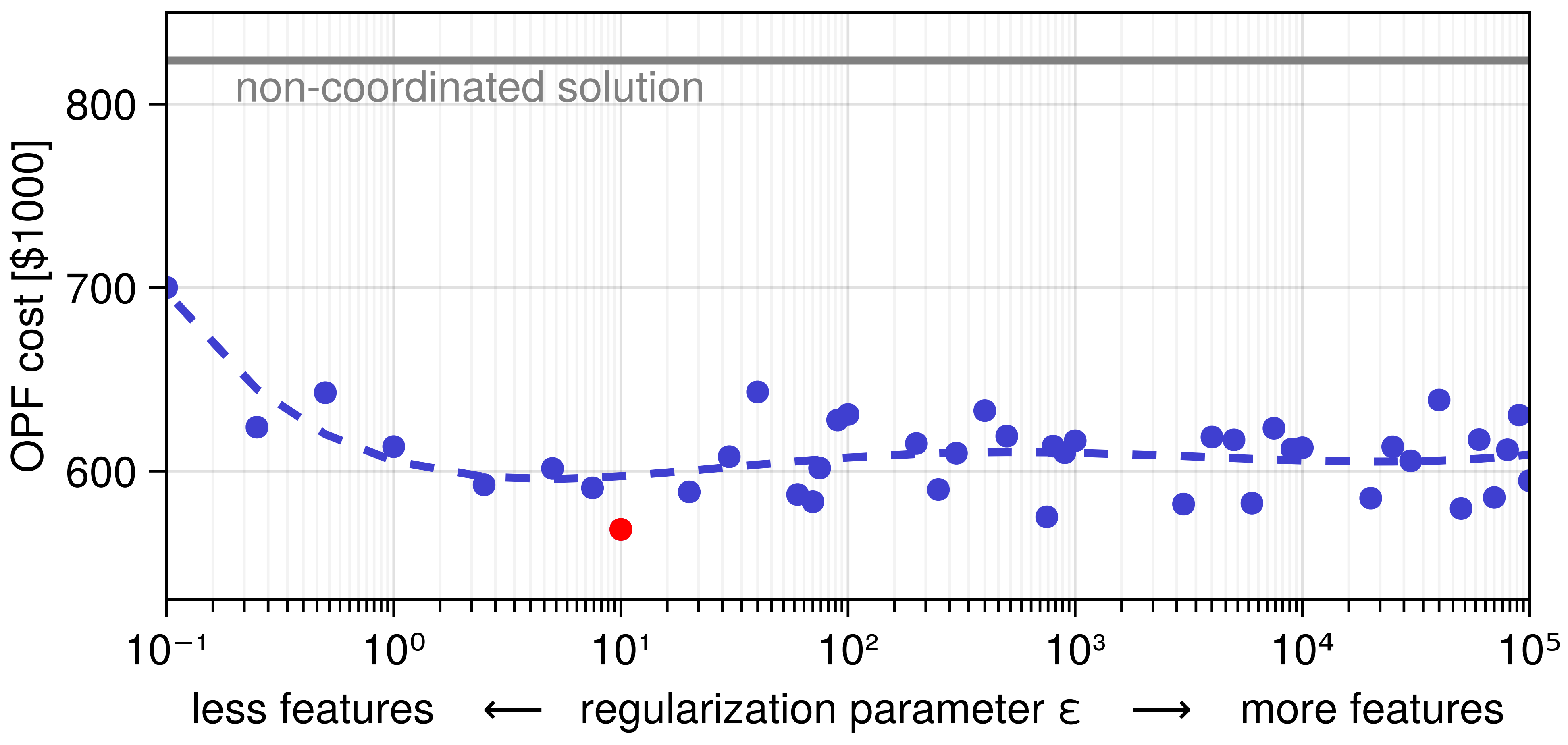
- Feature selection by means of ℓ_1-regularization
- ℓ_1-regularization also ensures coordination robustness
- Can we organize coordination using just one feature?



- Feature selection by means of ℓ_1-regularization
- ℓ_1-regularization also ensures coordination robustness
- Can we organize coordination using just one feature?

Selected contextual features for Grid-NetDC coordination for different ε

ε	# of features	zonal electricity demand											power flow												zonal electricity price											zonal renewable power output										
		d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	d_{11}	f_1	f_2	f_3	f_4	f_5	f_6	λ_7	f_8	f_9	f_{10}	f_{11}	f_{12}	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}	λ_{11}	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}	r_{11}
1000.0	29	○	●	●	●	○	●	○	●	●	○	○	●	●	●	●	●	●	●	●	●	●	●	○	●	●	●	●	●	●	●	○	●	●	○	○	●	●	○	○	○	○	○	○	○	●
100.0	28	●	●	●	●	●	●	○	●	●	●	○	●	●	●	●	●	●	●	●	○	●	○	○	○	●	●	●	●	○	○	●	●	●	●	○	●	○	○	○	○	○	○	○	○	●
10.0	24	●	●	●	●	○	●	●	●	●	●	●	●	●	●	●	○	●	●	●	●	●	○	○	○	○	○	○	○	○	○	●	●	●	○	○	●	○	○	○	○	○	○	○	○	●
5.0	20	●	●	○	●	●	●	●	●	●	●	●	●	●	●	○	●	○	●	●	●	○	○	○	○	○	○	○	○	○	○	○	●	●	○	○	●	○	○	○	○	○	○	○	○	○
2.5	13	●	○	●	○	○	●	●	●	●	●	○	●	●	○	○	○	●	○	○	●	○	○	○	○	○	○	○	○	○	○	○	●	●	○	○	○	○	○	○	○	○	○	○	○	○
1.0	6	●	○	●	○	○	○	○	○	○	●	○	●	●	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
0.5	3	●	○	○	○	○	○	○	○	○	●	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
0.1	1	○	○	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○



- Feature selection by means of ℓ_1-regularization
- ℓ_1-regularization also ensures coordination robustness
- Can we organize coordination using just one feature?

Selected contextual features for Grid-NetDC coordination for different ε

ε	# of features	zonal electricity demand											power flow												zonal electricity price											zonal renewable power output										
		d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	d_{11}	f_1	f_2	f_3	f_4	f_5	f_6	λ_7	f_8	f_9	f_{10}	f_{11}	f_{12}	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}	λ_{11}	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}	r_{11}
1000.0	29	○	●	●	●	○	●	○	●	●	○	○	●	●	●	●	●	●	●	●	●	●	●	○	●	●	●	●	●	●	●	○	●	●	○	○	●	●	○	○	○	○	○	○	○	●
100.0	28	●	●	●	●	●	●	○	●	●	●	○	●	●	●	●	●	●	●	●	○	●	○	○	○	●	●	●	●	○	○	●	●	●	●	○	●	○	○	○	○	○	○	○	○	●
10.0	24	●	●	●	●	○	●	●	●	●	●	●	●	●	●	●	○	●	●	●	●	●	○	○	○	○	○	○	○	○	○	●	●	●	○	○	●	○	○	○	○	○	○	○	○	●
5.0	20	●	●	○	●	●	●	●	●	●	●	●	●	●	●	○	●	○	●	●	●	○	○	○	○	○	○	○	○	○	○	○	●	●	○	○	●	○	○	○	○	○	○	○	○	○
2.5	13	●	○	●	○	○	●	●	●	●	●	○	●	●	○	○	○	●	○	○	●	○	○	○	○	○	○	○	○	○	○	○	●	●	○	○	○	○	○	○	○	○	○	○	○	○
1.0	6	●	○	●	○	○	○	○	○	○	●	○	●	●	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
0.5	3	●	○	○	○	○	○	○	○	○	●	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
0.1	1	○	○	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

peak electricity demand in New York City

Policy optimization still requires sensitive data

$$
\begin{aligned}
\underset{\beta,\varphi_i,\dots,\varphi_q,p_i,\dots,p_q}{\text{minimize}} \quad & \frac{1}{q}\sum_{i=1}^{q} c_{\text{pwr}}(p_i) \\
\text{subject to} \quad & p_i \in \mathcal{P}_{\text{pwr}}(\vartheta_i^\star), \quad \forall i = 1,\dots,q \\
& \varphi_i = \beta_0 + \beta_1 x_i, \ \|\beta\|_1 \leqslant \varepsilon, \quad \forall i = 1,\dots,q \\
& \vartheta_i^\star \in \underset{\vartheta_i}{\text{minimize}} \quad c_{\text{net-dc}}(\vartheta_i) \\
& \qquad \text{subject to} \quad \vartheta_i \in \mathcal{W}_{\text{net-dc}}(\varphi_i), \quad \forall i = 1,\dots,q
\end{aligned}
$$

cost function?

network/gen/load data?

- How to enable the exchange of sensitive training datasets?

Policy optimization still requires sensitive data

$$
\begin{aligned}
\underset{\beta,\varphi_i,\dots,\varphi_q,p_i,\dots,p_q}{\text{minimize}} \quad & \frac{1}{q}\sum_{i=1}^{q} c_{\text{pwr}}(p_i) \\
\text{subject to} \quad & p_i \in \mathcal{P}_{\text{pwr}}(\boldsymbol{\vartheta}_i^{\star}), \quad \forall i = 1,\dots,q \\
& \varphi_i = \beta_0 + \beta_1 x_i, \; \|\beta\|_1 \leqslant \varepsilon, \quad \forall i = 1,\dots,q \\
& \boldsymbol{\vartheta}_i^{\star} \in \underset{\boldsymbol{\vartheta}_i}{\text{minimize}} \quad c_{\text{net-dc}}(\boldsymbol{\vartheta}_i) \\
& \qquad \text{subject to} \quad \boldsymbol{\vartheta}_i \in \mathcal{W}_{\text{net-dc}}(\varphi_i), \quad \forall i = 1,\dots,q
\end{aligned}
$$

cost function?

network/gen/load data?

- How to enable the exchange of sensitive training datasets?
- Differential privacy provides an answer

Differentially Private Algorithms for Synthetic Power System Datasets

Vladimir Dvorkin, Jr., *Member, IEEE*, and Audun Botterud, *Member, IEEE*

Conclusions

- We build significantly less generation capacity than what we need to accommodate the growing AI demand
- We need to leverage the unique (geospatial) flexibility of datacenters to accommodate the loads
- Legacy optimization-based solutions to coordination are not the option
  - Lack of real-time communication interfaces
  - Privacy barriers for information exchange
  - Computationally intractable problem
- We developed a contextual regression mechanism (AgentCONCUR) to:
  - Minimize real-time communication requirements (contextual features)
  - Enable coordination at minimum data exchange (e.g., feature selection)
  - Computationally tractable real-time computations (millisecond)

IEEE TRANSACTIONS ON POWER SYSTEMS, AUGUST 2024 1

Agent Coordination via Contextual Regression (AgentCONCUR) for Data Center Flexibility

Vladimir Dvorkin, *Member, IEEE*