AI Data Center Power Curtailment

Potential, Challenges, and Implementation

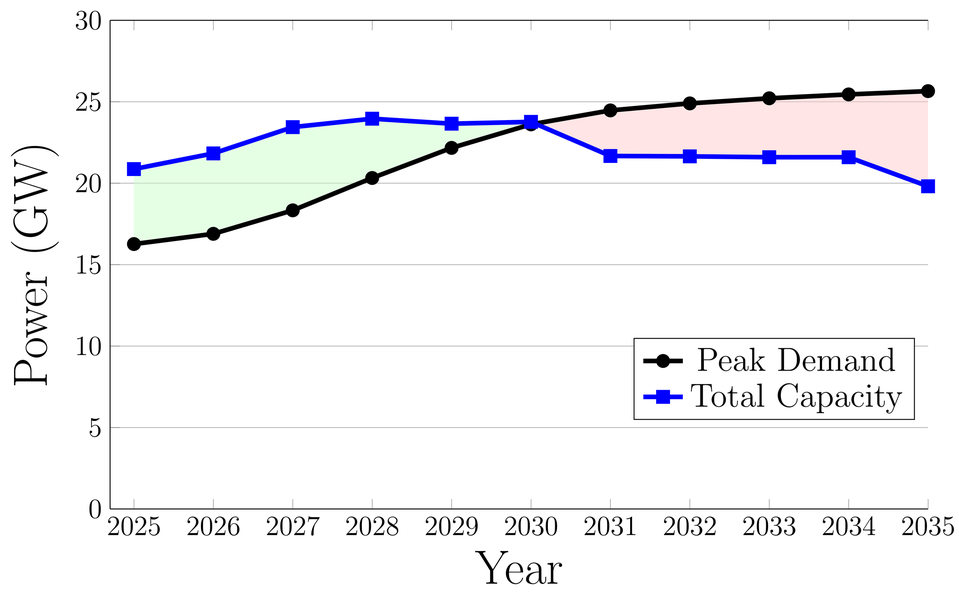
Max Hawkins and Richard Rex — In collaboration with Carbon Direct, Inc.

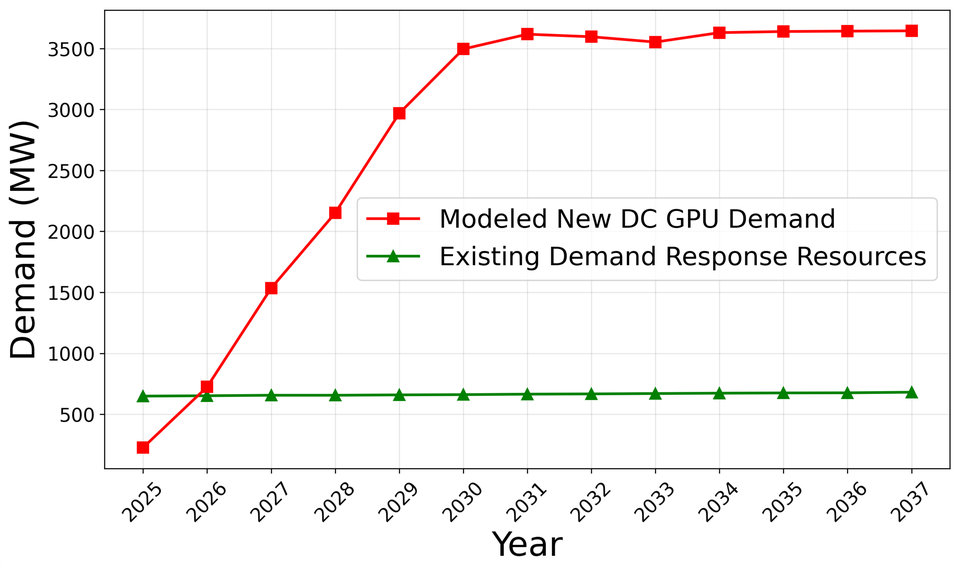
Motivation

Scale: AI adoption is causing immense data center demand - primarily using GPUs for AI training/inference.

Impact: Data centers now consume *significant* energy and water and cause substantial CO2 emissions.

Grid Stress: The resulting decreased power reserve margins pose major risks to grid stability (see below).





Opportunity: AI as Flexible Load

AI workloads are:

- ***Interruptible***: Can checkpoint/resume training jobs and slow/delay inference.
- ***Predictable***: Schedulers can align workloads with predicted grid conditions.
- ***Controllable***: AI power draw is dominated by GPUs, which can be throttled.

*Profiled on a GH200 system running vLLM

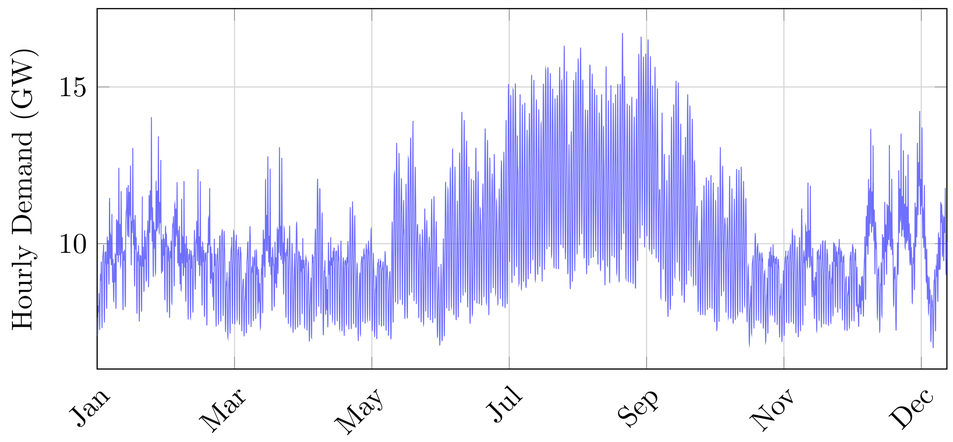
†Simplified, hypothetical model

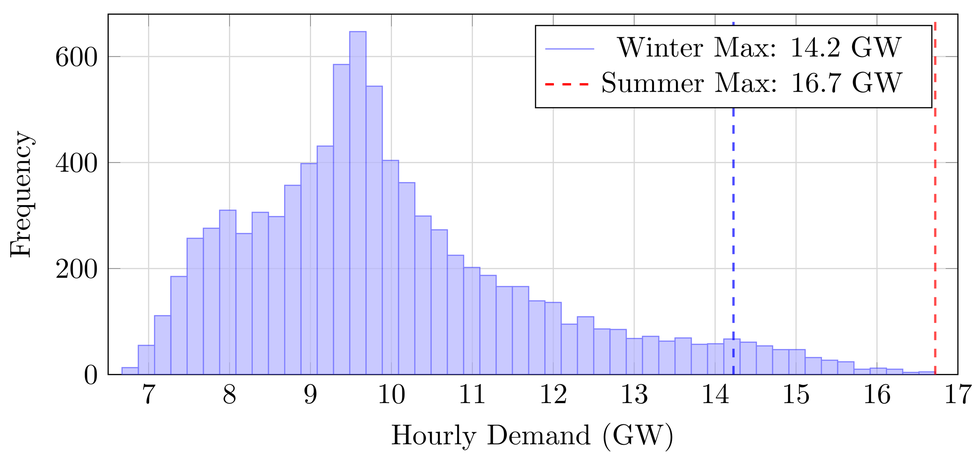
Curtailment is:

- ***Short***: Typically hours per year, coinciding with grid peaks (for now).
- ***Granular***: GPU-level or even subsystem (comp/mem/net) power control.
- **Beneficial**: Reducing power draw can also reduce water use or CCS risks.

Understanding Demand Variability

- Grid demand is highly variable: Most hours of the year sit well below peak.
- Seasonal and hourly swings necessitate coordination between the grid and flexible loads.





Financial Incentives?

When is AI power curtailment financially incentivized?

Lower-bound estimate: Relate hardware depreciation to peak power draw

Real-time market price breakeven value → $P_{\text{breakeven}}$

$$P_{\text{breakeven}} = \frac{\text{Compute CapEx depreciation}}{\text{Thermal Design Power}}$$

AI System	Thousands of $	TDP (kW)	Breakeven Price ($/MWh)
GB200 NVL72	4,100	120	775
DGX B200	380 - 800	12.3	706 - 1,463
GH200	42.5	1	970

Challenges

Incentives and Predictions: Aligning market price signals with AI scheduling and managing shared data centers.

Financial: Breakeven prices for AI systems are ~700–1,500 $/MWh, and most data centers have fixed-price power contracts → no incentive.

Regulation: Recent shifts. e.g. Texas Senate Bill 6 (2025) mandates curtailment or use of backup generation for large loads.

Implementation: For HPC/education, SLURM plugins (#SBATCH --gpu-freq <freq> #SBATCH --gpu-power <cap>). Corporations: Unique, in-house?