Carbon- and Cost-Aware Workload Scheduling for Data Centers on the CAISO Grid with Battery Energy Storage System (BESS) Extension

Stefanie Aipassa, Peijie Dai, Saskia Kusnadi, Gregory Saragih, Parker Tikson

1. Abstract

This study evaluates carbon- and cost-aware workload scheduling for a large data center operating on the CAISO grid, with an extension for battery energy storage systems (BESS). We model a 500 MW data center over a 168-hour weekly horizon with four workload classes: inflexible interactive serving, ETL pipelines, batch ML training, and cold backup. Flexible workloads are shifted within operational delay windows using hourly CAISO electricity price and marginal emissions signals for summer, winter, and shoulder seasons. We then compare a battery-off case against four BESS sizes to evaluate cost savings, carbon savings, battery throughput, and net incremental BESS value. Results show that workload shifting alone provides meaningful operating savings, especially in summer for cost and in shoulder season for carbon. Larger BESS systems increase gross operational savings, but all tested battery cases have negative net incremental value after ownership and degradation costs. Overall, the results suggest that software-based workload flexibility should be used first, while BESS deployment requires broader value stacking beyond energy arbitrage alone.

2. Introduction

2.1 Motivation and Background

Global data center electricity consumption reached approximately 460 TWh in 2022 and is projected to grow substantially as AI workloads scale [8]. U.S. data centers consumed an estimated 176 TWh in 2023—about 4.4% of national electricity—with projections of 325–580 TWh by 2028 [23]. At the same time, a large fraction of data center compute is temporally deferrable: batch ML training, ETL pipelines, and nightly backup tasks carry completion deadlines measured in hours, not milliseconds. Google workload data shows approximately 70% of total compute is flexible [13], while Alibaba production traces suggest up to 90% is delay-tolerant [24]. This flexibility enables shifting energy consumption toward periods of low cost and low carbon intensity without major changes to computing hardware.

Battery energy storage adds another important layer to this problem. While workload shifting changes when flexible computation is performed, a battery can shift when grid electricity is drawn by charging during lower-cost or lower-carbon periods and discharging during expensive or carbon-intensive periods. However, batteries also introduce operational and economic constraints, including limited energy capacity, charge/discharge power limits, round-trip efficiency losses, degradation costs, and ownership costs. Therefore, this study does not only ask whether flexible workload scheduling is useful, but also whether adding a battery provides enough additional value to justify its cost.

The CAISO grid provides a rich optimization environment. California's high solar penetration produces pronounced daily cycles in both price and carbon intensity—the duck curve [22]. Crucially, the structure of these signals varies by season in ways that affect what scheduling and battery operation can achieve. In summer, strong demand and high solar generation coexist: the marginal generator remains a gas peaker even at midday, keeping carbon intensity elevated, while a steep evening price ramp as solar drops off creates a strong cost signal. In the shoulder season, solar supply exceeds demand at midday, causing curtailment and driving marginal carbon intensity near zero—a dramatic carbon signal the optimizer can exploit. In winter, both signals are relatively flat and gas-heavy, limiting the benefit of temporal shifting and storage. Characterizing these seasonal differences, and evaluating when workload flexibility and BESS deployment are most valuable, is a central contribution of this work.

2.2 Relevant Literature

Radovanovic et al. at Google demonstrated that temporal workload shifting can reduce carbon-attributed consumption by 1–2% during peak intensity hours at scale [13]. Wierman et al. showed the interactive and background nature of tasks [16]. Zheng et al. showed that geographic load migration between fossil-heavy and renewable-heavy grid regions can reduce emissions by 113–239 ktCO_2e/year, and identified temporal flexibility as a complementary lever [20]. On the market side, Liu et al. developed a convex framework for price-responsive geographic load balancing [10], and Ghatikar et al. validated demand response strategies in California field tests [5].

Most directly relevant is Liu, Wierman et al. [11], [16] who formulated a convex optimization model for renewable- and cooling-aware workload management with time-varying prices and deadline constraints. Our work extends this with 30-day averaged CAISO LMP and WattTime MOER data over a one-week multi-class horizon, producing a season-specific Pareto characterization and real-world KPIs. He et al. [6] and Cheng et al. [3] provide battery degradation cost frameworks that define the natural next extension of this project.

2.3 Focus of this Study

This study focuses on optimizing data center electricity use by coordinating flexible workload scheduling with battery energy storage under hourly electricity price and carbon-intensity signals. Within the broader challenge of reducing the cost and emissions of growing computing demand, we evaluate how different workload flexibility levels and BESS sizes affect cost savings, carbon reductions, and overall system value.

3. Technical Description

3.1 Data Sources

CAISO Day-Ahead LMP. Hourly local marginal prices are fetched from the CAISO OASIS API via the gridstatus library at node TH_NP15_GEN-APND (NP15 hub), with an example of data seen in Figure 1. We fetch all available days within each seasonal window and compute an hourly mean by hour-of-day, producing a 24-hour seasonal average robust to anomalous days: Summer (July 1–31, 2024, 31 days), Winter (January 1–31, 2024, 31 days), Shoulder (April 1–30, 2024, 30 days). All 92 days were fetched successfully.

Grid Carbon Intensity with Marginal Operating Emissions Rates (MOER) (lb CO_2/MWh) represents the emissions intensity of the marginal generator—the correct signal for load-shifting, as it captures what emissions are actually avoided when demand shifts rather than the average grid mix. Values are sourced from WattTime for the CAISO_NORTH balancing authority for the same seasonal periods. A 30-day seasonal average will replace current values once WattTime credentials are refreshed before final submission.

Table 1. CAISO LMP (30-day average) and WattTime MOER summary statistics by season.

Signal	Season	Mean	Min	Max	Key Feature
LMP ($/MWh)	Summer	$40	$27	$141	Evening ramp peak at hour 20
LMP ($/MWh)	Winter	$78	$65	$91	Elevated, relatively flat all day
LMP ($/MWh)	Shoulder	$22	−$1	$45	Near-zero midday from solar surplus
MOER (lb/MWh)	Summer	708	400	980	Flat; gas peaker dominates all day
MOER (lb/MWh)	Winter	882	790	1020	Flat; minimal solar displacement
MOER (lb/MWh)	Shoulder	718	0	1069	Near-zero midday solar curtailment

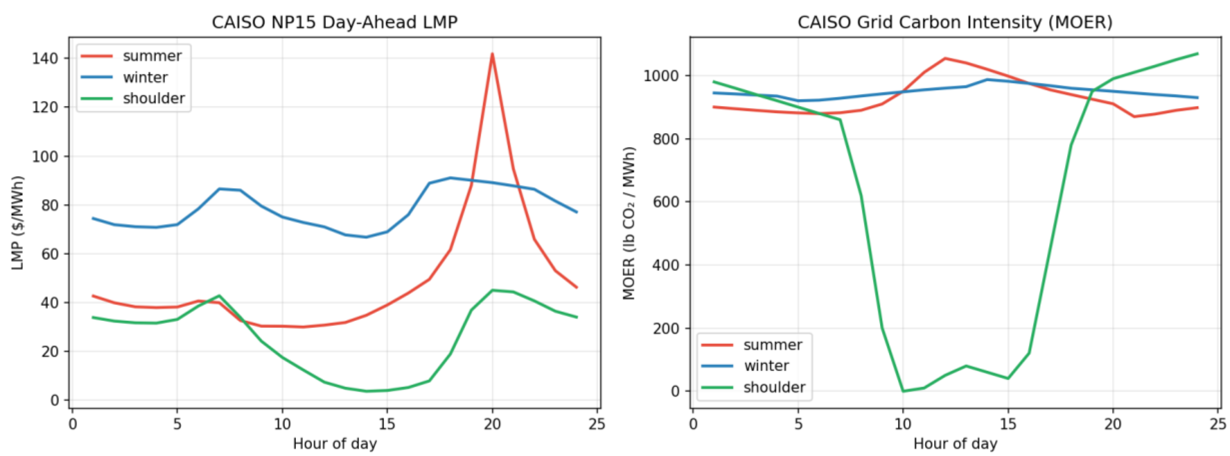

Figure 1. CAISO NP15 LMP (left) and WattTime MOER carbon intensity (right) by season. Summer LMP shows the duck curve evening ramp. Shoulder MOER drops near zero at midday due to solar curtailment. Summer MOER stays elevated because strong demand absorbs solar generation without reaching the marginal gas peaker threshold.

3.2 Demand and Workload Model

The workload model was formed into a four-class demand model that separates an inflexible user-facing load floor from flexible computation. In the current configuration, interactive serving is the inflexible portion (30% of total demand, W = 0 h), while ETL pipelines, batch ML training, and cold backup form the flexible portion (70% of total demand). This 30%/70% split is a configurable engineering assumption for a mixed-use data center, but it is consistent with carbon-aware computing and cluster-management literature [13], [14], [15]. Each class has a distinct baseline profile, demand share, flexibility window, and urgency penalty. Interactive serving is fixed; ETL/data pipelines are short-window and dependency-constrained; batch ML training is delay-tolerant but conservatively bounded to 6 hours in the current code; and cold backup is the most flexible background workload. This structure follows the scheduling-relevance logic in the workload literature: workloads should be separated when they require different operational constraints, not only when their magnitudes differ [9], [17], [19].

Each workload profile is first generated as a 24-hour shape, then repeated across the 168-hour week with class-specific day-of-week scaling and normalized to a weekly mean of one before being multiplied by the task share and 500 MW total capacity. Thus, the equations below describe temporal shape, as seen in Table 2.

Table 2. Profile equations, behavioral differences, and literature justification.

Workload	Profile equation / model	Shape and amplitude	Scheduling behavior (W & λ)	Source(s)
Interactive serving	$\ell\_int(t)=clip(0.5+0.5 \sin(2\pi(t-6)/24)+\varepsilon\_t, 0.30, 1.0)$; $\varepsilon\_t \sim N(0,0.01^2)$	Strongest sinusoid; amplitude 0.50; noon peak and midnight trough; clipped to preserve always-on service floor.	W=0, only s=0 feasible; this is the fixed 30% inflexible demand floor.	Radovanovic et al. [13] separate user-facing services such as Search, Maps, and YouTube as non-temporally flexible workloads, while Sukprasert et al. [14] describe interactive workloads as latency-constrained. Yang et al. [19] support user-driven diurnal behavior for latency-critical services.
ETL pipeline	$\ell\_ETL(t)$=1.55 for 06–09, 1.45 for 12–15, 1.15 for 19–21, 0.35 otherwise, plus ε_t	Step-function profile; discrete morning, midday, and evening bursts; not a smooth sinusoid.	W=4 h with highest urgency λ=3.0 because pipeline delays can block downstream tasks.	Lechowicz et al. [9] support the DAG/pipeline constraint logic: data-processing jobs can have precedence constraints, so delaying upstream stages affects downstream completion. Acun et al. [1] can support the idea of limited-SLO workload classes if you are using their 4-hour tier.
Batch ML training	$\ell\_batch(t)=0.75+0.22 \sin(2\pi(t-5)/24)+\varepsilon\_t$; $\varepsilon\_t \sim N(0,0.01^2)$	Mild sinusoid; amplitude 0.22; 11:00 peak; lower amplitude reflects queue smoothing.	W=6 h with medium urgency λ=1.5; conservative relative to 12–24 h in literature.	Radovanovic et al. [13] support that flexible workloads such as machine learning, simulations, and data-processing pipelines can tolerate delays if completed within 24 hours. Hu et al. [7] can support ML/batch workload behavior and power relevance. However, the current W=6W=6 h is our conservative modeling choice.
Cold backup	$\ell\_cold(t)=0.85-0.30 \sin(2\pi(t-6)/24)+\varepsilon\_t$; $\varepsilon\_t \sim N(0,0.025^2)$	Inverse of interactive serving; overnight peak and noon trough; amplitude 0.30.	W=24 h with lowest urgency λ=0.5; most flexible scheduling headroom.	Radovanovic et al. [13] show that flexible data center workloads can be delayed away from high-carbon hours using hourly limits on flexible tasks. The inverse overnight-oriented cold-backup profile is our simplifying modeling assumption, not a measured cold-backup trace.

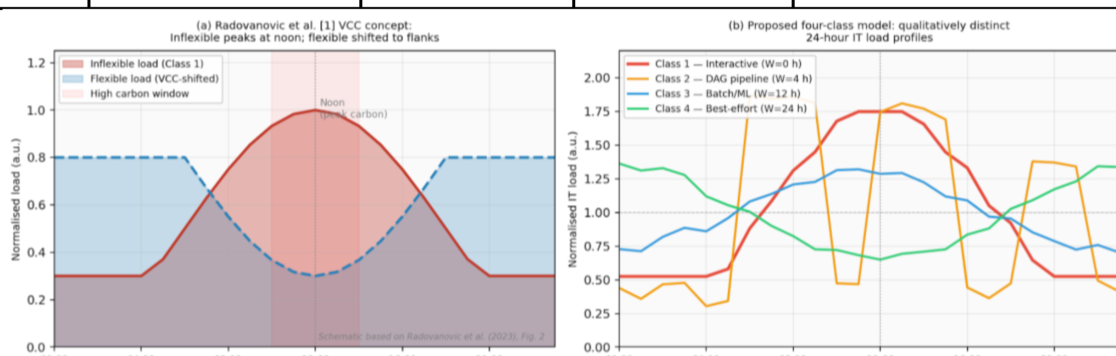

Figure 2. Comparison of the literature load model and the proposed four-class workload model. Panel (a) on the left shows the Radovanovic et al. [13] concept of inflexible interactive load and flexible compute shifting, with the interactive component peaking around noon and flexible load shifted toward overnight and early-morning hours. (b) on the right shows the proposed normalized four-class IT load profiles, which preserve this qualitative structure while adding more operational detail.

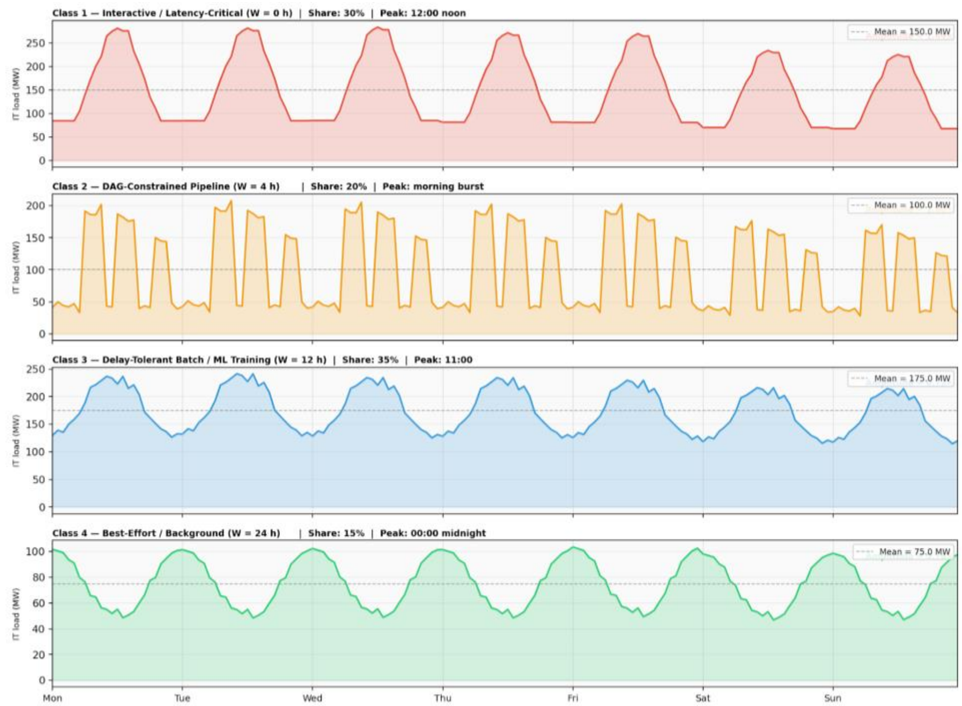

Figure 3. Baseline 168-hour IT load demand profiles for all four workload classes. 500 MW IT capacity; CAISO zone; Dashed horizontal lines indicate weekly mean for each class. Vertical dotted lines mark day boundaries (Monday through Sunday).

3.2.1 Mathematical Representation in The Optimizer

Total IT demand is represented as the sum of four normalized class profiles scaled by their demand shares. The base data center capacity is 500 MW and the 1.30 peak multiplier gives a 650 MW operating ceiling.With t as time/hour, K as set of workload classes, k one workload class, $P_{IT}(t)$ as total IT demand at hour t, $P_k(t)$ as power demand of workload k at hour t, s_k as share of total demand for workload k, P_{total} as base data center capacity (500 MW), and $l_k(t)$ as normalized hourly load shape for workload k.

$$P_{IT}(t) = \sum_{k\epsilon K} P_k(t) \quad \text{(Eq. 1)}$$

$$P_{max} = 1.30 \times 500\,MW = 650\,MW \quad \text{(Eq. 2)}$$

$$P_k(t) = s_k \times P_{total} \times l_k(t) \quad \text{(Eq. 3)}$$

For scheduling, $x_{k,t,s}$ is the fraction of workload class k released at hour t that executes after delay s. The optimizer must assign every released workload to exactly one feasible delay. Since W=0 for interactive serving, it has only one feasible assignment, s=0, and cannot shift. W_k as the maximum flexibility window for workload class k and τ as the actual execution hour. $P_{k,opt}(\tau)$ represents the optimized power of the workload class k at execution hour τ, while $P_{k,base}(t)$ represents the original baseline power released at hour t. The indicator term $1_{\{\tau = t + s\}}$ equals 1 only when workload released at hour t with delay s is executed at hour τ, and equals 0 otherwise.

Therefore, the optimized workload profile is built by accumulating all workload fractions that are shifted into each execution hour.

$$P_{k,opt}(\tau) = \sum_t \sum_{s=0}^{W_k} P_{k,base}(t) x_{k,t,s} 1_{\{\tau = t + s\}} \text{ (Eq. 4)}$$

The model includes a task-specific urgency penalty to discourage unnecessary delay. ETL has the highest urgency because it represents downstream-dependent pipeline work, batch ML has moderate urgency, and cold backup has the lowest urgency because it provides the largest scheduling headroom. $D_{k,s}$ is the urgency penalty assigned to workload class k when its execution is delayed by s hours, where λ_k controls how steeply the penalty grows (higher = more urgent), W_k is the maximum allowed delay window, and s is the actual delay applied.

$$D_{k,s} = exp(\lambda_k \frac{s}{W_k}) - 1, D_{k,0} = 0 \text{ (Eq. 5)}$$

3.3 Battery Sizing & Economics

Battery CAPEX is estimated using an NREL-style duration-based structure that separates energy-related and power-related costs. This is important because some storage components scale with energy capacity, such as battery cabinets, while others scale with power capacity, such as inverters and balance-of-system equipment [4], [12].

Table 3. Battery size scenarios

Scenario	Power	Energy	Duration
BESS 50 MW / 100 MWh	50 MW	100 MWh	2.0 h
BESS 100 MW / 200 MWh	100 MW	200 MWh	2.0 h
BESS 200 MW / 500 MWh	200 MW	500 MWh	2.5 h
BESS 200 MW / 800 MWh	200 MW	800 MWh	4.0 h

Fixed O&M is modeled as 2.5% of battery CAPEX per year, and CAPEX is annualized with a 7% discount rate and 15-year lifetime with the energy cost at $241/kWh and power cost at $372/kW according to NREL ATB. It also represents lithium-ion utility-scale battery storage with fixed O&M rather than variable O&M in the base case and states that fixed O&M includes augmentation needed to maintain rated capacity over the lifetime [12].

$$CAPEX = E_{kWh} \times \$241/kWh + P_{kW} \times \$372/kW \text{ (Eq. 6)}$$

$$CRF = \frac{r(1+r)^n}{(1+r)^n - 1} \text{ (Eq. 7)}$$

$$Weekly\ ownership\ cost = \frac{CAPEX \times CRF + 2.5\% \times CAPEX}{52} \text{ (Eq. 8)}$$

The economic decision is based on incremental BESS value rather than gross savings. Incremental BESS savings compare each BESS case to the battery-off case, which already includes workload shifting. Net incremental BESS value subtracts weekly ownership and degradation costs. This structure follows storage valuation logic: a battery should be judged by the additional value it creates relative to its additional cost [2], [8], [18]. The degradation cost rate of $11/MWh is taken as the midpoint of the $5–$17/MWh range established by He et al. [21] for lithium-ion battery storage under typical dispatch conditions.

$$Incremental\ BESS\ Savings = Savings\ with\ BESS - Savings\ with\ no\ BESS\ (Battery\ OFF) \quad \text{(Eq. 9)}$$

$$Battery\ throughput\ (\Theta) = \sum_{t=1}^{168} \quad (P_{charge}(t) + P_{discharge}(t)) \quad \text{(Eq. 10)}$$

$$Degradation\ cost = \Theta \times \$11/MWh \quad \text{(Eq. 11)}$$

$$Net\ incremental\ BESS\ Value = Incremental\ BESS\ Savings - Weekly\ Ownership\ Cost - Degradation\ Cost \quad \text{(Eq. 12)}$$

3.4 Optimization Model

3.4.1 Parameters and Notation

We model one week of data center operation using a 168-hour horizon, indexed by hour t = 1, ..., 168. Each workload task k has a baseline demand $d_{k,t}$, a flexibility window W_k, a share of total demand, and a maximum power limit. The model uses hourly electricity price LMP_t and marginal grid emissions $MOER_t$, tiled from 24-hour seasonal profiles for summer, winter, and shoulder seasons.

Key parameters include:

$d_{k,t}$ = released demand of task k at hour t

W_k= maximum allowed delay for task k

α = cost-carbon tradeoff weight (between 0 and 1)

P_{max} = facility peak power limit

LMP_t = electricity price signal at hour t

$MOER_t$ = grid carbon intensity signal at hour t

For battery scenarios, the model also includes battery energy capacity, charge and discharge power limits, round-trip efficiency, minimum and maximum state of charge, initial state of charge, and degradation cost.

3.4.2 Decision Variables

The main workload decision variable is $x_{k,t,s}$.

$x_{k,t,s}$ = the fraction of workload k released at hour t that is shifted by s hours. For each task, s = 0, ..., W_k. If s = 0, the job runs immediately. If s > 0, the job is delayed.

For battery-enabled cases, the model also decides:

$u_{ch,t}$ = battery charging power at hour t

$u_{dis,t}$ = battery discharging power at hour t

SOC_t = battery state of charge at hour t

The optimized facility load is the shifted computing load plus battery charging minus battery discharging.

3.4.3 Objective Function

We solve one LP per (season, α) combination over a padded one-week horizon. The padded horizon is 168 + $2W_{na}{}^{x}$ = 216 hours; only the inner 168 hours are reported. Grid signals are tiled from the 24-hour seasonal average over the padded horizon. The optimization minimizes a weighted combination of normalized electricity cost, normalized carbon emissions, delay penalty, and battery degradation cost:

$$min\left[\alpha \sum_t \widehat{LMP}_t P_t + (1-\alpha)\sum_t \widehat{MOER}_t P_t + \gamma \sum_t D_{k,s} d_{k,t} x_{k,t,s} + \kappa \sum_t (u_t^{ch} + u_t^{dis})\right] \text{ (Eq. 13)}$$

Where P_t is the optimized net facility power, $\widehat{LMP}_t$ and $MOER_t$ are normalized price and carbon signals, γ controls the weight of delay penalties, and κ represents normalized battery degradation cost. The parameter α controls the tradeoff: α = 1 prioritizes electricity cost, while α = 0 prioritizes carbon reduction.

The delay penalty increases with the shift length:

$$D_{k,s} = e^{\lambda_k s/W_k} - 1 \text{ (Eq. 14)}$$

This makes long delays more costly than short delays, especially for more urgent workloads.

3.4.4 Constraints

Each released workload must be fully scheduled:

$$\sum_{s=0}^{W_k} x_{k,t,s} = 1 \text{ (Eq. 15)}$$

$$x_{k,t,s} \geq 0 \text{ (Eq. 16)}$$

Workloads cannot be shifted outside the optimization horizon:

$$x_{k,t,s} = 0 \text{ if t + s > 168 (Eq. 17)}$$

Each task must obey its hourly power cap:

$$P_{k,t} \leq P_{max} \text{ (Eq. 18)}$$

The total net facility power must stay within physical limits:

$$0 \leq P_t \leq P_{max} \text{ (Eq. 19)}$$

For battery cases, the battery state of charge evolves according to charging and discharging decisions:

$$SOC_{t+1} = SOC_t + \eta u_t^{ch} - \frac{1}{\eta} u_t^{dis} \text{ (Eq. 20)}$$

with limits:

$$SOC_{min} \leq SOC_t \leq SOC_{max} \text{ (Eq. 21)}$$

$$0 \leq u_t^{ch} \leq u^{ch,max} \text{ (Eq. 22)}$$

$$0 \leq u_t^{dis} \leq u^{dis,max} \quad \text{(Eq. 23)}$$

The model also prevents excessive simultaneous charge and discharge by limiting total battery power:

$$u_t^{ch} + u_t^{dis} \leq max\,(u^{ch,max}, u^{dis,max}) \quad \text{(Eq. 24)}$$

Finally, the battery begins and ends the week at the same state of charge, so the optimizer cannot gain artificial value by fully draining the battery at the end:

$$SOC_0 = SOC_{168} \quad \text{(Eq. 25)}$$

The resulting problem is a linear program solved with CVXPY using the CLARABEL solver. We evaluate the model across seasons, workload flexibility assumptions, and multiple battery sizes to compare cost savings, carbon savings, battery throughput, and net incremental BESS value.

4. Discussion

4.1 Demand Shift

Figure 4 shows the stacked workload composition across baseline, workload shifting only, and BESS sizing scenarios. The baseline row preserves the original task profiles, while optimized rows show temporal reshaping of flexible classes within their feasible windows. Interactive serving remains anchored at the bottom of the stack and does not move, which is necessary because user-facing services are excluded from temporal flexibility in carbon-aware computing systems [13], [14]. This reflects that data center sustainability is not only about adding hardware but also about coordinating flexible computation with grid conditions. Similar conclusions appear in temporal workload-shifting studies, where achievable carbon benefits depend on workload deadlines and the temporal structure of the grid carbon signal [14], [17]. The benefits of the carbon and cost saved are extended in Appendix A and further explained in subsection 4.2.

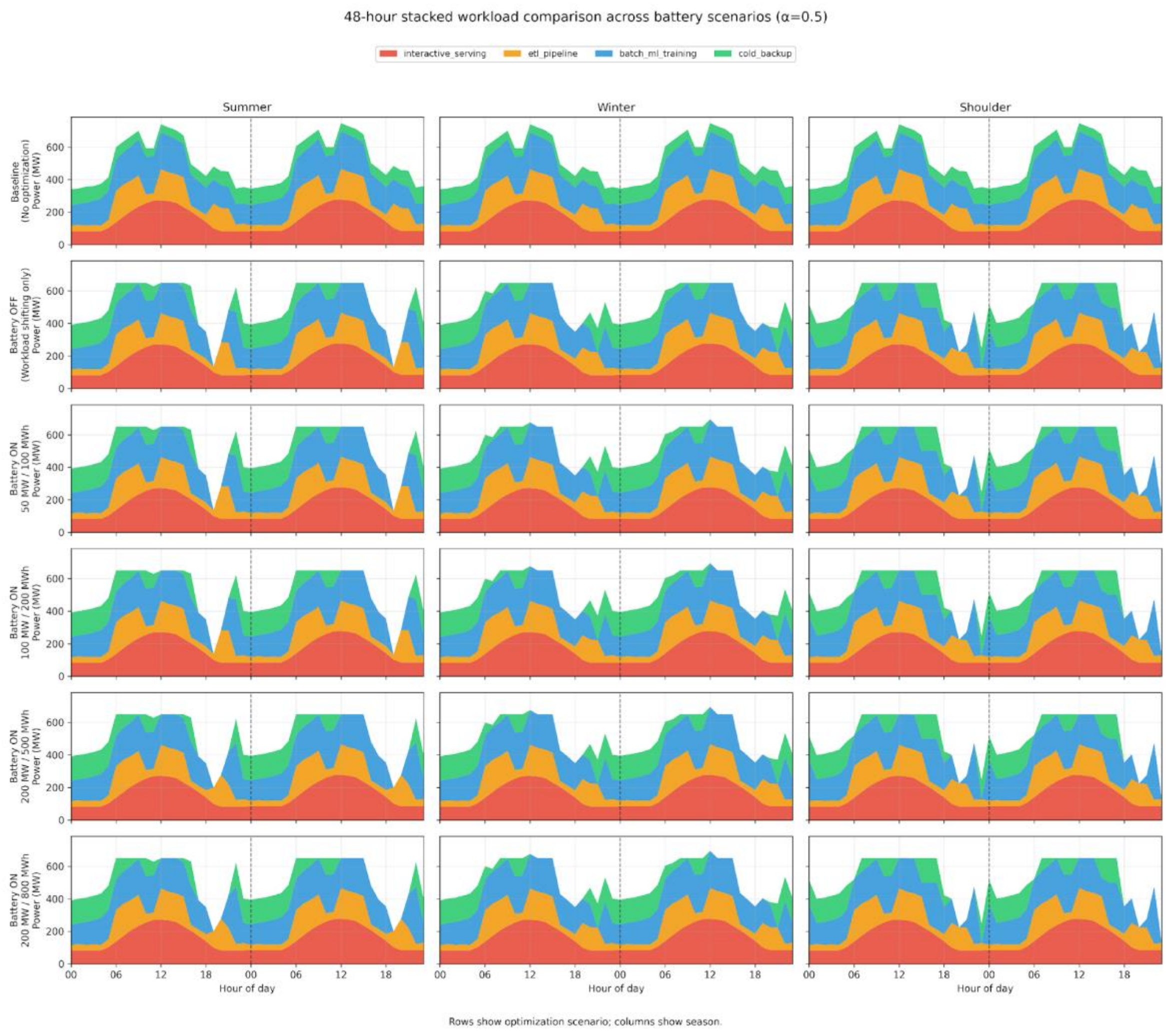

Figure 4. 48-hour stacked workload comparison across seasons and BESS sizing scenarios. Rows show baseline, workload shifting only, and BESS cases. Columns show season.

4.2 Cost and Carbon Reduction

4.2.1 Weekly Cost Savings by Scenario

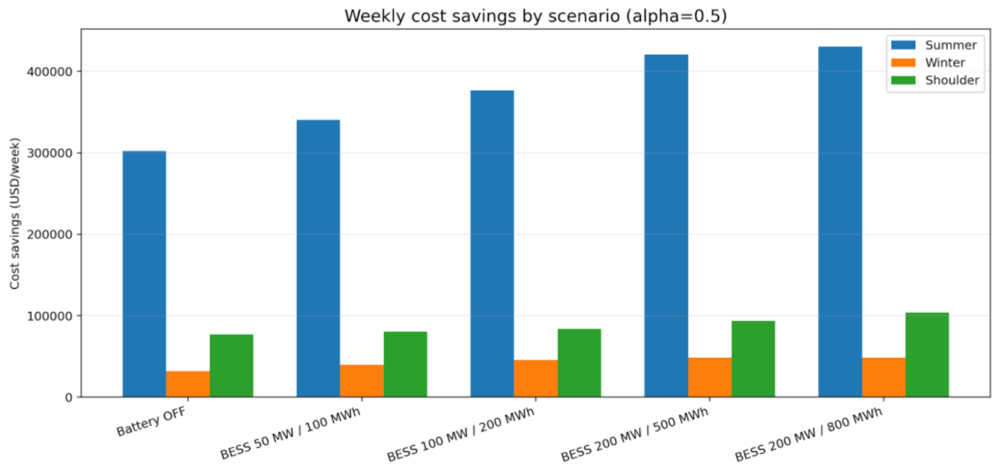

Figure 5. Weekly cost savings by scenario at α = 0.5. Values are weekly operating savings only and are not annualized.

Figure 5 shows that gross weekly cost savings increase as BESS size increases, particularly in summer. The battery-off case already saves $302,326/week through workload shifting alone, while the 200 MW / 800 MWh BESS reaches $430,468/week. This behavior is expected when the cost signal has strong intra-day variation: flexible workloads and batteries can avoid high-price hours and move consumption toward lower-price hours [8], [13]. Summer has the largest gross cost savings because the LMP profile has stronger price-spread opportunities than winter or shoulder. Winter gross savings are much smaller and saturate quickly at the larger BESS sizes, indicating that the storage and flexible workload options have fewer valuable hours to arbitrage. This aligns with storage-economics literature showing that arbitrage value depends on price volatility and spread width, not merely on the existence of a battery [2], [8].

4.2.2 Weekly Carbon Savings by Scenario

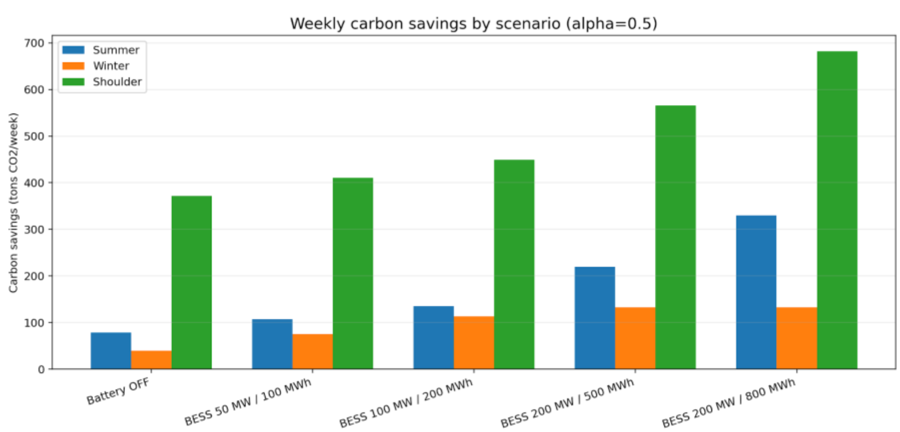

Figure 6. Weekly carbon savings by scenario at α = 0.5. Values are operational tons CO2 avoided per representative week.

Figure 6 shows that shoulder season produces the largest carbon savings, even though summer produces the largest cost savings. This indicates that the carbon signal and price signal are not identical. In shoulder season, the model finds more opportunities to move flexible demand and storage operations toward lower-MOER hours. The increase in carbon savings with BESS size is strongest in shoulder season, rising from 371.6 tons/week in the battery-off case to 681.7 tons/week for the 200 MW / 800 MWh BESS. This suggests that storage adds carbon value when it can absorb electricity in cleaner hours and reduce net grid demand in dirtier hours. However, this operational carbon benefit does not include embodied battery emissions or lifecycle impacts, which Acun et al. identifies as an extensive measure for a holistic 24/7 carbon-free data center design [1].

4.2.3 Incremental BESS savings

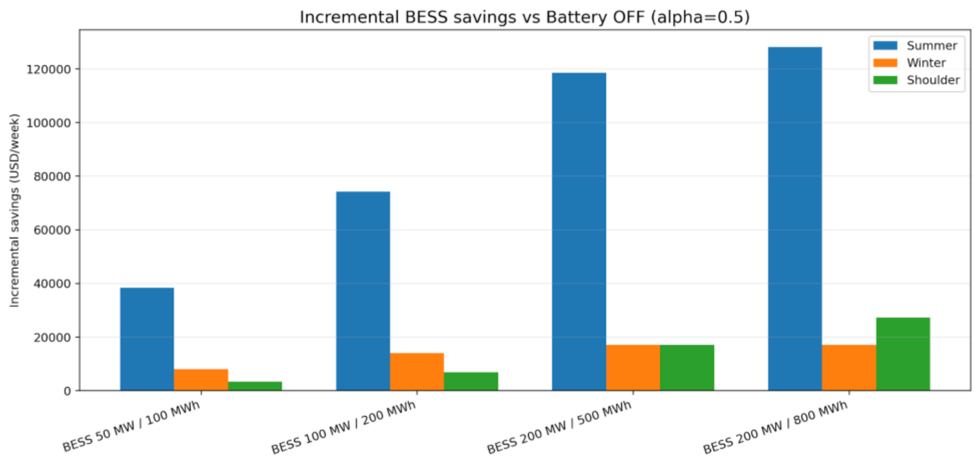

Figure 7. Incremental BESS savings versus the battery-off case. This isolates the marginal operating value of storage after workload shifting is already included.

Figure 7 isolates the marginal operating value of storage by subtracting the battery-off savings from each BESS case. The figure shows that BESS adds the most incremental cost savings in summer, increasing from $38,345/week for the 50 MW / 100 MWh system to $128,142/week for the 200 MW / 800 MWh system. The winter incremental savings are small and nearly flat for the larger systems. This suggests that energy capacity beyond a certain level is not valuable in winter under the modeled price and carbon profiles. Shoulder season has stronger carbon-saving potential, but the cost signal does not provide enough monetary value to justify the battery by itself. This is consistent with the idea that storage gains value when it can exploit wider price spreads [2], [8].

4.3 Battery Throughput and Economic Decisions

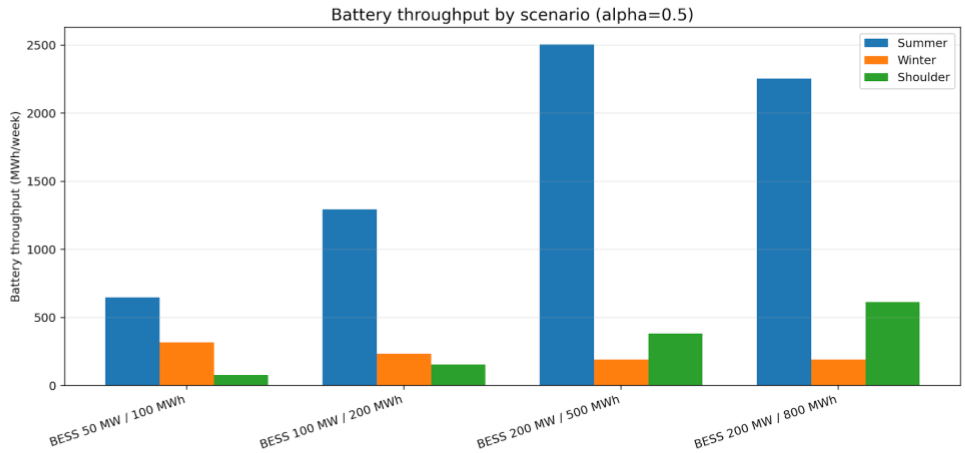

Figure 8. Battery throughput by scenario. Throughput equals weekly charge plus discharge energy and measures how actively the battery cycles.

Battery throughput varies by season and BESS size because it measures how much the battery cycles. Summer consistently produces the highest throughput across all scenarios due to the strong intraday LMP spread and aligned MOER signal creating clear daily charge-discharge windows. Meanwhile, winter throughput stays low and nearly flat because both price and carbon signals are relatively stable across hours. The more interesting result is that the 200 MW / 800 MWh case can deliver higher cost and carbon savings than the 200 MW / 500 MWh case despite lower throughput — the larger energy capacity lets the optimizer hold charge longer and dispatch selectively into the highest-value hours rather than cycling more frequently. This is consistent with what the storage literature describes: IRENA [8] notes that arbitrage value saturates with growing storage capacity. Bae et al. [2] reinforce this through their energy stacking argument, that profitability comes from combining services strategically from the same unit rather than running it at maximum throughput. NREL's 2025 report [4] supports this interpretation by showing that BESS economics depend on duration, lifetime, efficiency, O&M, and cycling assumptions, so lower throughput can be beneficial in the model when it still increases savings while reducing degradation exposure.

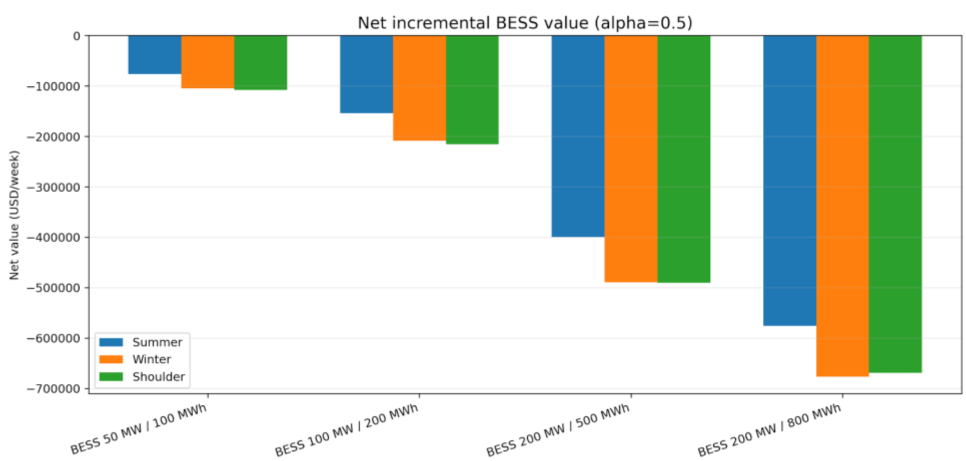

Figure 9. Net incremental BESS value by season and BESS size. All tested cases are negative after ownership and degradation costs.

Figures 9 show why the final decision differs from the gross savings figures. In summer, the 50 MW / 100 MWh system adds $38,345/week of incremental savings, but its weekly ownership cost is $110,687/week and degradation is $3,235/week. The result is a net value of -$75,577/week. Larger systems save more operationally, but their ownership costs rise faster than their incremental benefits [4], [12]. The negative net value across all BESS sizes does not mean batteries are technically ineffective. It means this project only credits BESS for cost-carbon arbitrage within the data center scheduling problem. Real storage projects often rely on value stacking, including frequency regulation, reserves, demand-charge reduction, capacity value, resilience, and renewable integration. The storage-economics literature emphasizes that excluding these services can understate the full economic case for batteries [2], [8], [18].

4.4 Pareto Frontier

The Pareto frontier is generated by sweeping alpha from 0 to 1. Alpha = 0 represents a carbon-focused solution, alpha = 1 represents a cost-focused solution, and alpha = 0.5 represents the balanced case used for the main KPI reporting. A tight frontier cluster indicates limited tradeoff space because deadlines, capacity limits, the carbon cap, and battery constraints restrict the feasible set [13], [14].

$$J_\alpha = \alpha\, cost_{norm} + (1 - \alpha)\, carbon_{norm} \text{ (Eq. 26)}$$

The Pareto frontier should be interpreted as a diagnostic tool, not the final investment decision. It shows how the optimizer trades normalized cost against normalized carbon for a given scenario. Because BESS changes the feasible set, Pareto plots should ideally be shown separately for each BESS size or combined with clear labels; otherwise, the effect of storage size can be hidden. Appendix B shows the Pareto frontier for the battery-off scenario across three seasons, where each point represents one value of α swept from 0 (carbon only) to 1 (cost only) and the star marks the unoptimized baseline. Summer achieves the largest cost reduction, shifting leftward from its baseline, due to a steep intraday LMP spread that gives the optimizer clear incentive to defer flexible workloads away from expensive evening hours. Shoulder season shows the strongest simultaneous cost and carbon reduction as the near-zero midday MOER from high solar penetration creates a clean charging window that the optimizer exploits. Winter frontiers cluster tightly around the baseline because both price and carbon signals are flat across hours, leaving little room for improvement through workload shifting alone. Across all BESS sizing scenarios the frontier shape remains similar to battery-off, confirming that the primary optimization lever in this model is workload scheduling rather than storage capacity.

4.5 Next Steps

Future work should first improve the emissions data pipeline by replacing the cached MOER profiles with fully refreshed 30-day WattTime marginal emissions data for each season. This would make the carbon results more reliable and would allow the optimizer to be tested under real historical grid conditions rather than partially cached profiles.

A second next step is to expand the battery value model. In this study, BESS is valued only through cost-carbon arbitrage inside the data center scheduling problem. Future teams should include additional revenue and reliability streams such as demand-charge reduction, frequency regulation, reserves, capacity value, backup power, and resilience. This would provide a more realistic investment case because real battery projects are rarely justified by energy arbitrage alone.

Third, the workload model could be made more realistic by using measured data center or cloud traces instead of synthetic workload profiles. Future work could also test different data center mixes, such as AI-heavy facilities with more batch training demand or enterprise facilities with a larger interactive load floor. This would show how sensitive the results are to workload composition and flexibility assumptions.

Finally, the model could be extended from one data center to a multi-region setting. Combining temporal shifting, geographic load migration, renewable procurement, and storage would better represent how large cloud providers actually operate. This would also allow future teams to study when moving computation across regions is more effective than shifting it in time or storing energy locally.

5. Summary

This project shows that carbon- and cost-aware workload scheduling can reduce weekly operating cost and emissions while preserving inflexible interactive demand. The main sustainability insight is that data centers already have useful flexibility in batch, ETL, ML training, and backup workloads, so some benefits can be achieved through scheduling rather than immediate infrastructure investment. Systems and control methods are well suited to this problem because they can coordinate workload deadlines, latency constraints, site power limits, battery limits, and grid price/carbon signals at the same time [9], [13], [14].

The four-class workload model is important because it avoids two unrealistic extremes: treating all demand as flexible or treating all demand as fixed. Instead, it keeps the latency-critical interactive load unchanged while allowing bounded shifting of ETL, batch ML, and cold-backup tasks. This matches prior carbon-aware scheduling work showing that real benefits depend on workload type, deadlines, and the timing of grid emissions [9], [14], [17].

The results show strong seasonal differences. Summer gives the largest cost-saving opportunity because CAISO prices rise sharply in the evening. Shoulder season gives the largest carbon-saving opportunity because midday solar oversupply can make marginal emissions very low. Winter has fewer opportunities because both price and carbon signals are flatter.

BESS improves gross operating performance by charging during lower-cost or lower-carbon hours and discharging during less favorable hours. However, after ownership and degradation costs, all tested BESS sizes have negative net incremental value. This means batteries are operationally useful, but cost-carbon arbitrage alone is not enough to justify them in this model. Prior storage economics work similarly shows that batteries often need value stacking, such as demand-charge reduction, resilience, ancillary services, reserves, and capacity value, to become economically viable [2], [8], [18].

A key modeling lesson is that carbon reduction should not rely only on a weighted cost-carbon objective. Since low-price hours are not always low-carbon hours, the carbon-cap constraint helps ensure emissions do not rise while costs fall. Overall, workload shifting is the most cost-effective lever in this study, while BESS needs a broader investment case that includes full-year data, demand charges, resilience, ancillary services, renewable integration, and lifecycle emissions from both batteries and IT infrastructure [1], [4], [12], [13], [14], [17], [18].

References

[1] Acun, B., et al. (2023). *Carbon Explorer: A Holistic Framework for Designing Carbon Aware Datacenters*. ASPLOS.

[2] Bae, J., et al. (2025). *Energy Stacking in Battery Energy Storage Systems*. *Management Science*.

[3] Cheng, M., et al. (2023). *Optimal dispatch for second-life batteries with online SoH estimation*. *Renewable & Sustainable Energy Reviews*, 173, 113053.

[4] Cole, W., Ramasamy, V., & Turan, M. (2025). *Cost Projections for Utility-Scale Battery Storage: 2025 Update*. National Renewable Energy Laboratory.

[5] Ghatikar, G., et al. (2012). *Demand Response Opportunities for Data Centers*. LBNL-5763E.

[6] He, G., et al. (2021). *Power system dispatch with marginal degradation cost of battery storage*. *IEEE Transactions on Power Systems*, 36(4), 3578–3590.

[7] Hu, Q., Sun, P., Yan, S., Wen, Y., & Zhang, T. (2021). *Characterization and prediction of deep learning workloads in large-scale GPU datacenters*. ACM/IEEE SC.

[8] International Renewable Energy Agency. (2020). *Electricity Storage Valuation Framework*.

[9] Lechowicz, A., et al. (2025). *Carbon- and Precedence-Aware Scheduling for Data Processing Clusters*. arXiv:2502.09717.

[10] Liu, Z., et al. (2011). *Greening geographical load balancing*. *ACM SIGMETRICS*, 39(1), 193–204.

[11] Liu, Z., et al. (2012). *Renewable and cooling aware workload management*. *ACM SIGMETRICS*, 40(1), 175–186.

[12] National Renewable Energy Laboratory. (2025). *2024 Annual Technology Baseline: Utility-Scale Battery Storage*.

[13] Radovanovic, A., et al. (2023). *Carbon-Aware Computing for Datacenters*. *IEEE Transactions on Power Systems*, 38(2), 1270–1280.

[14] Sukprasert, T., Souza, A., Bashir, N., Irwin, D., & Shenoy, P. (2024). *On the Limitations of Carbon-Aware Temporal and Spatial Workload Shifting in the Cloud*. EuroSys.

[15] Verma, A., Pedrosa, L., Korupolu, M., Oppenheimer, D., Tune, E., & Wilkes, J. (2015). *Large-scale cluster management at Google with Borg*. EuroSys.

[16] Wierman, A., et al. (2014). *Opportunities and Challenges for Data Center Demand Response*. International Green Computing Conference, 1–10.

[17] Wiesner, P., et al. (2021). *Let's Wait Awhile: How Temporal Workload Shifting Can Reduce Carbon Emissions in the Cloud*. ACM Middleware.

[18] Yamujala, S., et al. (2022). *Multi-service based economic valuation of grid-connected battery energy storage systems*. *Journal of Energy Storage*.

[19] Yang, L., et al. (2022). *Workload consolidation in Alibaba clusters: The good, the bad, and the ugly*. ACM SoCC.

[20] Zheng, J., Chien, A. A., & Suh, S. (2022). *Mitigating Curtailment and Carbon Emissions through Load Migration*. *Joule*, 6(10), 2208–2222.

[21] He, G., Kar, S., Mohammadi, J., & Moutis, P. (2020). *Power System Dispatch With Marginal Degradation Cost of Battery Storage. IEEE Transactions on Power Systems*, PP(99), 1–1. https://doi.org/10.1109/TPWRS.2020.3048401

[22] California ISO. (2016). What the Duck Curve Tells Us About Managing a Green Grid. CAISO Fast Facts.

[23] Lawrence Berkeley National Laboratory. (2024). United States Data Center Energy Usage Report.

[24] Takci, M.T., et al. (2025). Data centres as a source of flexibility for power systems. Energy Reports, 13, 3661–3671.

Appendix A. KPI Table at alpha = 0.5 (weekly values only)

The table reports weekly operating results for the representative simulated week. Cost and carbon operating savings are not annualized. Battery ownership is reported as a weekly equivalent of annualized CAPEX recovery plus fixed O&M.

Scenario	Season	Cost saved ($/wk)	Carbon avoided (tons/wk)	Throughput (MWh/wk)	Incremental savings ($/wk)	Ownership ($/wk)	Net value ($/wk)	Decision
Battery OFF	Summer	$302,326	78.6	0.0	$0	$0	$0	BASELINE
Battery OFF	Winter	$31,725	39.7	0.0	$0	$0	$0	BASELINE
Battery OFF	Shoulder	$76,813	371.6	0.0	$0	$0	$0	BASELINE
BESS 50 MW / 100 MWh	Summer	$340,672	107.0	647.0	$38,345	$110,687	-$75,577	NO
BESS 50 MW / 100 MWh	Winter	$39,779	75.0	317.6	$8,053	$110,687	-$104,222	NO
BESS 50 MW / 100 MWh	Shoulder	$80,218	410.3	76.7	$3,404	$110,687	-$107,666	NO
BESS 100 MW / 200 MWh	Summer	$376,612	134.9	1,294.1	$74,286	$221,374	-$153,559	NO
BESS 100 MW / 200 MWh	Winter	$45,777	113.4	234.2	$14,052	$221,374	-$208,493	NO
BESS 100 MW / 200 MWh	Shoulder	$83,622	449.1	153.5	$6,809	$221,374	-$215,333	NO
BESS 200 MW / 500 MWh	Summer	$420,914	219.6	2,504.7	$118,588	$505,221	-$399,156	NO
BESS 200 MW / 500 MWh	Winter	$48,800	132.7	192.1	$17,075	$505,221	-$489,106	NO
BESS 200 MW / 500 MWh	Shoulder	$93,835	565.4	383.7	$17,022	$505,221	-$490,117	NO

BESS 200 MW / 800 MWh	Summer	$430,468	329.9	2,254.5	$128,142	$692,637	-$575,768	NO
BESS 200 MW / 800 MWh	Winter	$48,800	132.7	192.1	$17,075	$692,637	-$676,523	NO
BESS 200 MW / 800 MWh	Shoulder	$104,048	681.7	613.9	$27,235	$692,637	-$668,472	NO

Appendix B. Cost-Carbon Pareto Frontier by BESS Scenario