Wasserstein Distributionally Robust Optimization for Coordinated AI Datacenter Scheduling and Power Distribution Network Operation

Jiebao Zhang, Siqi Yan, Zhichao Sheng, Hongwen Yu, Haoyu Wang, Yu Liu and Ye Shi

***Abstract*—The rapid growth of artificial intelligence (AI) workloads intensifies the coupling between AI datacenters and local distribution grids. We propose a distributionally robust day-ahead co-optimization model to jointly schedule multi-class, delay-tolerant workloads and operate a radial distribution network, explicitly incorporating power grid physical constraints. Our model handles operational risk from uncertain datacenter workloads and photovoltaic (PV) outputs using Wasserstein ambiguity sets and distributionally robust Conditional Value-at-Risk (CVaR) constraints. We develop a computationally efficient CVaR upper-bound approximation method that significantly reduces solution time while maintaining robustness guarantees for capacity exceedance and PV shortfall. Our simulation results demonstrate the effectiveness of the approach on a modified IEEE 33-bus system, showing cost reduction and robust grid security under worst-case distributions consistent with the data.**

***Index Terms*—Wasserstein Distributionally Robust Optimization, AI Datacenter, Power Distribution Network, Delay-tolerant Workloads.**

I. Introduction

The computational demands of modern deep learning and artificial intelligence (AI) have triggered an explosive growth in GPU-equipped hyperscale AI datacenters [1], [2]. This trend presents a horrible challenge to energy systems, with datacenter power consumption projected to rise dramatically [3], straining local grids and contributing significantly to carbon emissions [4]. A promising mitigation strategy lies in exploiting the inherent spatio-temporal flexibility of many computational workloads [5]. Batch processing, model training, and data analytics tasks are often not delay-sensitive and can be intelligently deferred to off-peak hours (temporal flexibility) or relocated to datacenters with lower electricity costs or higher renewable energy availability (spatial flexibility) [6]–[8].

However, the true distributions of multi-class datacenter loads are highly stochastic, and the integration of renewable resources like photovoltaic (PV) generation introduces further volatility. Deterministic optimization or simple Sample Average Approximation (SAA) methods are susceptible to out-of-sample failures, potentially leading to costly constraint violations such as grid congestion or voltage collapse.

To hedge against such ambiguity, the Wasserstein distributionally robust optimization (WDRO) framework [9], [10] for making decisions that are robust against the ambiguity of the true underlying data distribution. Instead of relying on a single empirical distribution, DRO optimizes against the worst-case distribution within a so-called ambiguity set [10]. Recent work has successfully applied DRO, particularly using the Wasserstein metric, to manage datacenter operations, providing tunable, probabilistic performance guarantees [11].

This paper builds on this state-of-the-art by proposing a WDRO model specifically for the coordination of AI datacenter scheduling and power grid operation. This work broadens the focus from solely managing datacenter-level metrics (like carbon cost [11]) to explicitly include the physical constraints of the power distribution network. Beyond an exact WDRO-CVaR reformulation, we introduce a computationally scalable CVaR upper-bound approximation that removes sample-dependent auxiliaries from the main problem, retaining convexity and delivering major speedups with negligible optimality loss. Our main contributions are:

- A novel joint day-ahead WDRO model that integrates the spatiotemporal scheduling of multi-class, delay-tolerant GPU workloads directly with the physical constraints of the local distribution grid.
- A computationally scalable, convex CVaR upper-bound approximation for Wasserstein DRO that eliminates sample-dependent auxiliary variables, dramatically reducing problem dimensionality and achieving order-of-magnitude speedups.

II. Problem Formulation

We formulate a day-ahead scheduling problem to co-optimize datacenter workload transfers and power grid dispatch over a 24-hour horizon T (indexed by t, s). The system comprises a set of datacenters K (indexed by k), which process different workload classes C (indexed by c), and a power grid with a set of nodes N (indexed by i). The grid includes nodes with conventional generators (N^{G}), PV generation (N^{PV}), and energy storage (N^{S}), connected by distribution lines $\mathcal{E}$. The uncertainty arises from submitted

Jiebao Zhang and Siqi Yan contributed equally to this work. Corresponding authors: Ye Shi.

Jiebao Zhang, Siqi Yan, Haoyu Wang, Yu Liu and Ye Shi are with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China. (email:zhangjb2023, yansq2024, wanghy, liuyu, shiye@shanghaitech.edu.cn)

Zhichao Sheng and Hongwen Yu are with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (email: zcsheng, hw_yu@shu.edu.cn)

This work was supported by National Natural Science Foundation of China (62303319), ShanghaiTech AI4S Initiative SHTAI4S202404, HPC Platform of ShanghaiTech University, and MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University).

979-8-3315-8138-1

datacenter workloads $\xi^{\mathrm{DC}} = (\xi^{\mathrm{DC}}_{c,t})_{c\in C,t\in T}$ and available PV output $\xi^{\mathrm{PV}} = (\xi^{\mathrm{PV}}_{i,t})_{i\in N^{\mathrm{PV}},t\in T}$.

A. Uncertainty and Distributionally Robust Formulation

The true probability distributions $\mathbb{P}$ of $\xi^{\mathrm{DC}}, \xi^{\mathrm{PV}}$ are unknown. We only have access to a set of M historical samples for both ξ^{DC} and ξ^{PV}, which together define an empirical distribution $\hat{\mathbb{P}}$. We build an ambiguity set $\mathcal{P}$ around $\hat{\mathbb{P}}$ using the Wasserstein metric.

Definition 1 (Wasserstein metric)**.** *The Wasserstein metric $d_W(\mathbb{P}_1, \mathbb{P}_2)$ is defined via*

$$d_W(\mathbb{P}_1, \mathbb{P}_2) = \inf_{\pi\in\Pi(\mathbb{P}_1,\mathbb{P}_2)} \left\{ \int_{\Xi^2} \|\xi_1 - \xi_2\| \pi(d\xi_1, d\xi_2) \right\}, \quad (1)$$

where $\Pi(\mathbb{P}_1, \mathbb{P}_2)$ is a joint distributions of ξ_1, ξ_2 with a marginal distribution of $\mathbb{P}_1$ and $\mathbb{P}_2$. Ξ denotes the support set. Here we use ℓ_1-norm.

The Wasserstein ambiguity set $\mathcal{P}$ is defined as:

$$\mathcal{P}^{\mathrm{DC}} = \{\mathbb{P} : W_1(\mathbb{P}, \hat{\mathbb{P}}^{\mathrm{DC}}) \le \epsilon^{\mathrm{DC}}\}, \mathcal{P}^{\mathrm{PV}} = \{\mathbb{P} : W_1(\mathbb{P}, \hat{\mathbb{P}}^{\mathrm{PV}}) \le \epsilon^{\mathrm{PV}}\}, \quad (2)$$

where W_1 is the 1-Wasserstein distance [9], [10] and ϵ is the robustness radius, a tuning parameter that controls the trade-off between robustness and optimality.

We replace the uncertain constraints with tractable, distributionally robust Conditional Value-at-Risk (CVaR) constraints, which enforce that the constraints hold with high probability even under the worst-case distribution in $\mathcal{P}$.

Definition 2. *Let $\ell(\xi)$ be a loss function that depends on a random vector ξ with distribution $\mathbb{P}$. For a given confidence level $\alpha \in (0, 1)$, the CVaR at level $1 - \alpha$ is defined as:*

$$\mathrm{CVaR}_{1-\alpha}(\ell(\xi)) = \inf_{\tau\in\mathbb{R}} \left\{ \tau + \frac{1}{\alpha} \mathbb{E}_{\mathbb{P}} \left[(\ell(\xi) - \tau)_+ \right] \right\}, \quad (3)$$

where $(\cdot)_+ = \max(\cdot, 0)$ and we have:

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathrm{CVaR}_{1-\alpha} (\ell(\xi)) \le 0 \Rightarrow \inf_{\mathbb{P}\in\mathcal{P}} \mathbb{P} (\ell(\xi) \le 0) \ge 1 - \alpha. \quad (4)$$

B. Optimization Objective and System Constraints

The objective is to minimize the total 24-hour generation cost. This cost is determined by the active power generation $P^{\mathrm{G}}_{i,t}$ (a decision variable) at each generator node $i \in N^{\mathrm{G}}$ and the corresponding electricity price $\lambda_{i,t}$:

$$\min_{x,P^{\mathrm{G}},Q^{\mathrm{G}},P^{\mathrm{S}},E,v,P^{\mathrm{PV}},P^{\mathrm{DC}},P^{\mathrm{DC,load}}} \sum_{t\in T} \sum_{i\in N^G} P^{\mathrm{G}}_{i,t} \cdot \lambda_{i,t}. \quad (5)$$

This objective is subject to the following grid and storage operational constraints.

1) Datacenter Workload Model: The core of the model captures the spatio-temporal flexibility of multi-class datacenter workloads. Here we follow the definition in [11] and the primary decision variable is $x_{c,t,k,s}$, representing the fraction of workload class c submitted at time t that is transferred to datacenter k for execution at time s. These transfer decisions must ensure all submitted workload is processed:

$$\sum_{k\in K} \sum_{s\in T} x_{c,t,k,s} = 1, \quad \forall c \in C, t \in T, \quad (6)$$

Transfers must respect the maximum delay tolerance h_c (a parameter for class c) and cannot be processed before submission:

$$x_{c,t,k,s} = 0, \quad \text{if } s < t \text{ or } s > t + h_c \quad (7a)$$
$$0 \le x_{c,t,k,s} \le 1. \quad (7b)$$

The realized workload at datacenter k and time s is determined by the sum of all workloads transferred to it, weighted by the random submitted loads ξ^{DC}:

$$P^{\mathrm{DC}}_{k,s} = \sum_{c\in C} \sum_{t\in T} x_{c,t,k,s}\, \xi^{\mathrm{DC}}_{c,t}, \quad (8)$$

which is random through ξ^{DC}. In contrast, we introduce the decision variable $P^{\mathrm{DC,load}}_{k,s}$ to denote the amount of datacenter workload committed to the grid operator. This commitment must respect the physical capacity:

$$0 \le P^{\mathrm{DC,load}}_{k,s} \le \overline{P}^{\mathrm{DC}}_{k,s}, \quad \forall k \in K, s \in T. \quad (9)$$

2) Storage Constraints: Storage dynamics are modeled via the energy state $E_{i,t}$ and the charging/discharging power $P^{\mathrm{S}}_{i,t}$ (where $P^{\mathrm{S}}_{i,t} > 0$ is charging). The model is constrained by the storage energy capacity $\overline{E}_i$, the power capacity $\overline{P}^S_i$, and the charge/discharge efficiency η_i. Daily periodicity is enforced ($E_{i,0} = E_{i,T}$).

$$E_{i,t} = E_{i,t-1} - \eta_i P^{\mathrm{S}}_{i,t}, \quad \forall i \in N^{\mathrm{S}}, t \in T, \quad (10a)$$
$$0 \le E_{i,t} \le \overline{E}_i, \; -\overline{P}^S_i \le P^{\mathrm{S}}_{i,t} \le \overline{P}^S_i, \quad (10b)$$
$$E_{i,0} = E_{i,T}. \quad (10c)$$

3) LinDistFlow Constraints: We use the linearized DistFlow model for radial distribution networks [12]. This model relates net power injections to the active and reactive power flows, $p_{i,j,t}$ and $q_{i,j,t}$, and the squared voltage magnitude $v_{i,t}$. The model uses the line resistance $r_{i,j}$ and reactance $x_{i,j}$.

$$\sum_{j:(i,j)\in\mathcal{E}} p_{i,j,t} - \sum_{j:(j,i)\in\mathcal{E}} p_{j,i,t} = P^{\mathrm{G}}_{i,t} - P^{\mathrm{load}}_{i,t} + P^{\mathrm{PV}}_{i,t} - P^{\mathrm{S}}_{i,t} - P^{\mathrm{DC,load}}_{i,t}, \quad \forall i \in N, t \in T, \quad (11a)$$

$$\sum_{j:(i,j)\in\mathcal{E}} q_{i,j,t} - \sum_{j:(j,i)\in\mathcal{E}} q_{j,i,t} = Q^{\mathrm{G}}_{i,t} - Q^{\mathrm{load}}_{i,t}, \quad \forall i \in N, t \in T, \quad (11b)$$

$$V_{j,t} - V_{i,t} = 2(r_{i,j} p_{i,j,t} + x_{i,j} q_{i,j,t}), \quad \forall (i,j) \in \mathcal{E}, t \in T, \quad (11c)$$
$$V_{0,t} = (v_0)^2, \quad \underline{v_i}^2 \le V_{i,t} \le \overline{v_i}^2, \quad \forall i \in N, t \in T, \quad (11d)$$
$$\underline{P}^G \le P^G_{i,t} \le \overline{P}^G, \quad \underline{Q}^G \le Q^G_{i,t} \le \overline{Q}^G. \quad (11e)$$

In these equations, $P^{\mathrm{G}}_{i,t}$ and $Q^{\mathrm{G}}_{i,t}$ are the generator active/reactive power dispatch. $P^{\mathrm{PV}}_{i,t}$ is the committed PV power based on ξ^{PV}. The nominal voltage V_0 is set at the slack bus (node 0). Grid security is enforced via bounds on generation and voltage.

4) Datacenter Load CVaR Constraint: We must ensure that the realized workload, defined by (8), in the worst-case distribution $\mathbb{P} \in \mathcal{P}$, does not exceed the committed workload in the worst $(1-\alpha)\%$ scenarios. We formulate this as a worst-case CVaR constraint with risk level α (e.g., $\alpha = 0.05$). For each datacenter k and time s:

$$\sup_{\mathbb{P}\in\mathcal{P}^{\mathrm{DC}}} \mathrm{CVaR}_{1-\alpha} \left(P^{\mathrm{DC}}_{k,s} - P^{\mathrm{DC,load}}_{k,s} \right) \le 0, \quad (12)$$

5) PV Generation CVaR Constraint: Similarly, we must ensure that the amount of dispatched PV power P^{PV} is deliverable with high probability. We formulate a robust shortfall constraint, ensuring the committed power $P^{\mathrm{PV}}_{i,t}$ does not exceed the actual realized PV generation $\xi^{\mathrm{PV}}_{i,t}$:

$$\sup_{\mathbb{P}\in\mathcal{P}^{\mathrm{PV}}} \mathrm{CVaR}_{1-\alpha} \left(P^{\mathrm{PV}}_{i,t} - \xi^{\mathrm{PV}}_{i,t} \right) \le 0. \quad (13)$$

This constraint ensures that the committed PV power is less than or equal to the realized PV generation with high probability $(1-\alpha)$ under the worst-case distribution, preventing

979-8-3315-8138-1

the system from over-relying on uncertain renewable energy.

C. Complete Formulation

The complete day-ahead co-optimization problem seeks to find the set of decision variables $\mathcal{X} = \{x, P^G, Q^G, P^S, E, v, P^{PV}, P^{DC}, P^{\text{DC,load}}\}$ that minimizes the objective function (5) subject to the deterministic and distributionally robust constraints. The full problem can be summarized compactly as:

$$\begin{aligned} \min_{\mathcal{X}} \quad & \sum_{t \in T} \sum_{i \in N^G} P^G_{i,t} \cdot \lambda_{i,t}, \\ \text{s.t.} \quad & (6) - (13), \end{aligned} \tag{14}$$

where (6)–(9) represent the datacenter workload requirements, (10) is storage operating limits, and (11) is the local distribution physical constraints. The CVaR constraints for the datacenter and PV generation are (12) and (13), respectively. The primary computational challenge stems from the semi-infinite CVaR constraints (12) and (13), which must hold for all distributions in the Wasserstein ambiguity sets.

III. Wasserstein DRO Formulation

The robust CVaR constraints in (12) and (13) are semi-infinite optimization problems, as they must hold for every distribution $\mathbb{P}$ within the infinite ambiguity sets $\mathcal{P}^{\text{DC}}$ and $\mathcal{P}^{\text{PV}}$. This section details the tractable reformulation of these constraints and proposes a scalable approximation to address computational challenges.

A. Tractable Reformulation of CVaR Constraints

A key result in distributionally robust optimization is that when the loss function $\ell(\xi)$ is affine in the random vector ξ, the worst-case CVaR constraint over a Wasserstein ball can be reformulated as a finite-dimensional convex problem. Both of our loss functions are affine. For the datacenter constraint (12), the loss $\ell(\xi^{\text{DC}}) = P^{\text{DC}}_{k,s} - P^{\text{DC,max}}_{k,s}$ is affine in ξ^{DC} via (8) (where $a = x_{c,t,k,s}$ and $b = -P^{\text{DC,max}}_{k,s}$). For the PV constraint (13), the loss $\ell(\xi^{\text{PV}}) = P^{\text{PV}}_{i,t} - \xi^{\text{PV}}_{i,t}$ is also affine in ξ^{PV}. Therefore, we can apply the following standard lemma, which provides an exact, tractable reformulation.

Lemma 1 (Wasserstein-DRO CVaR Reformulation [9], [13])**.** *Consider a constraint of the form*

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathrm{CVaR}_{1-\alpha}\left(\ell(\xi)\right) \leq 0, \tag{15}$$

where $\ell(\xi) = a^\top \xi + b$ is affine in ξ, and $\mathcal{P} = \{\mathbb{P} : W_1(\mathbb{P}, \hat{\mathbb{P}}) \leq \epsilon\}$ with support $\Xi = \{\xi : G\xi \leq h\}$.

$$\begin{aligned} & \tau + \frac{1}{\alpha}\left(\lambda\epsilon + \frac{1}{M}\sum_{m=1}^{M} s_m\right) \leq 0, \\ & a^\top \hat{\xi}_m + b - \tau + \gamma_{m,1}^\top (h - G\hat{\xi}_m) \leq s_m, \quad s_m \geq 0, \forall m, \\ & \gamma_{m,2}^\top (h - G\hat{\xi}_m) \leq s_m, \quad \forall m, \\ & \|G^\top \gamma_{m,1} - a\|_* \leq \lambda, \quad \gamma_{m,1} \geq 0, \lambda \geq 0, \forall m, \\ & \|G^\top \gamma_{m,2}\|_* \leq \lambda, \quad \gamma_{m,2} \geq 0, \forall m, \end{aligned} \tag{16}$$

where λ and $\tau \in \mathbb{R}$, $s \in \mathbb{R}^M$, $\gamma_{i,1}$ and $\gamma_{i,2} \in \mathbb{R}^{2 \times K \times T \times C}$. They are all auxiliary variables. $\|\cdot\|_*$ is the dual norm of the ground 1-norm in W_1. Hence the worst-case CVaR constraint admits a finite convex linear programming representation.

B. Computationally Efficient CVaR Upper Bound

Applying Lemma 1 to our problem, while exact, leads to a significant increase in problem size. The constraint (12) must be applied for each datacenter $k \in K$ and each time $s \in T$. This gives $K \times T$ individual CVaR constraints. For *each* CVaR constraint, the reformulation in Lemma 1 introduces $O(M)$ constraints and $O(M \cdot K \cdot T \cdot C)$ auxiliary variables. When the time horizon T, number of datacenters K, and number of samples M are large, the resulting convex problem becomes computationally intractable. To address this computational burden, we propose a more conservative but highly scalable approximation. This approach leverages two fundamental properties of CVaR.

Lemma 2 (CVaR Properties [14])**.** *CVaR satisfies:*
(1) Subadditivity: For any two random losses A and B, $\mathrm{CVaR}_\alpha(A + B) \leq \mathrm{CVaR}_\alpha(A) + \mathrm{CVaR}_\alpha(B)$;
(2) Positive Homogeneity: For any non-negative scalar $\gamma \geq 0$ and random loss A, $\mathrm{CVaR}_\alpha(\gamma \cdot A) = \gamma \cdot \mathrm{CVaR}_\alpha(A)$.

We apply these properties to derive an upper bound for the datacenter constraint (12). The loss is $L = P^{\text{DC}}_{k,s} - P^{\text{DC,max}}_{k,s}$. The original constraint is:

$$\sup_{\mathbb{P} \in \mathcal{P}^{\text{DC}}} \mathrm{CVaR}_{1-\alpha}\left(\sum_{c,t} x_{c,t,k,s} \cdot \xi^{\text{DC}}_{c,t} - P^{\text{DC,load}}_{k,s}\right) \leq 0. \tag{17}$$

Since $P^{\text{DC,max}}_{k,s}$ is a constant and $\mathrm{CVaR}(A+b) = \mathrm{CVaR}(A) + b$ for a constant b, this is equivalent to:

$$\sup_{\mathbb{P} \in \mathcal{P}^{\text{DC}}} \mathrm{CVaR}_{1-\alpha}\left(\sum_{c,t} x_{c,t,k,s} \cdot \xi^{\text{DC}}_{c,t}\right) \leq P^{\text{DC,load}}_{k,s}. \tag{18}$$

We now apply subadditivity and positive homogeneity (since $x_{c,t,k,s} \geq 0$) to the left-hand side:

$$\begin{aligned} \mathrm{CVaR}_{1-\alpha}\left(\sum_{c,t} x_{c,t,k,s} \cdot \xi^{\text{DC}}_{c,t}\right) & \leq \sum_{c,t} \mathrm{CVaR}_{1-\alpha}\left(x_{c,t,k,s} \cdot \xi^{\text{DC}}_{c,t}\right) \\ & = \sum_{c,t} x_{c,t,k,s} \cdot \mathrm{CVaR}_{1-\alpha}\left(\xi^{\text{DC}}_{c,t}\right). \end{aligned} \tag{19}$$

This creates a more tractable, conservative approximation by replacing the complex CVaR of a sum with a simple linear sum of pre-computed CVaR values. It introduces zero auxiliary variables and zero new constraints into the main optimization problem and simply replaces the complex CVaR constraints with standard linear constraints using the empirically pre-computed $\mathrm{CVaR}_{1-\alpha}\left(\xi^{\text{DC}}_{c,t}\right)$. The main problem size is independent of the number of samples M.

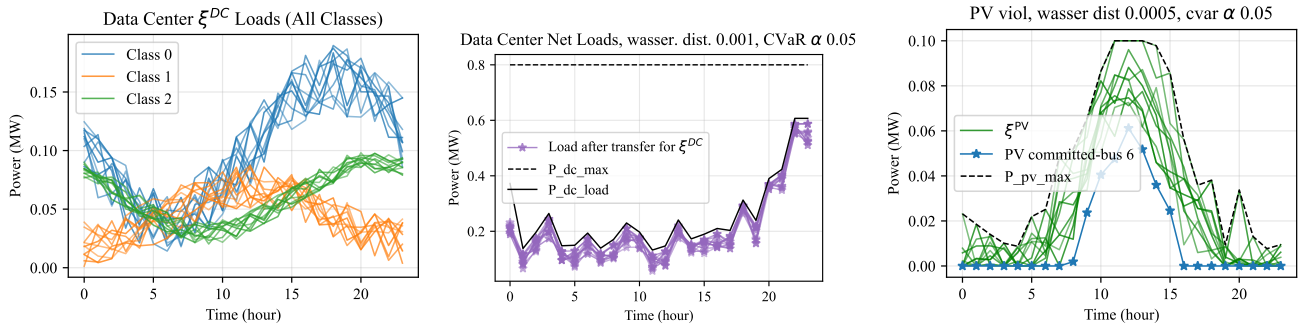
IV. Case Study

We validate the proposed WDRO model on a modified IEEE 33-bus distribution test system, implemented in Python using CVXPY [15] and solved with MOSEK [16].

A. System Configuration

We study the 33-bus radial distribution network over a 24-hour horizon (time slot = 1 h). The slack generator is at Bus 1; a single AI datacenter connects at Bus 24; one PV site is installed at Bus 6; and three storage units are located at Buses 9, 13, and 21.

979-8-3315-8138-1



(a) Datacenter stochastic demand by class (10 scenarios).

(b) Net committed DC load and capacity

(c) PV uncertainty and commitment

Fig. 1: Distributionally-robust feasibility

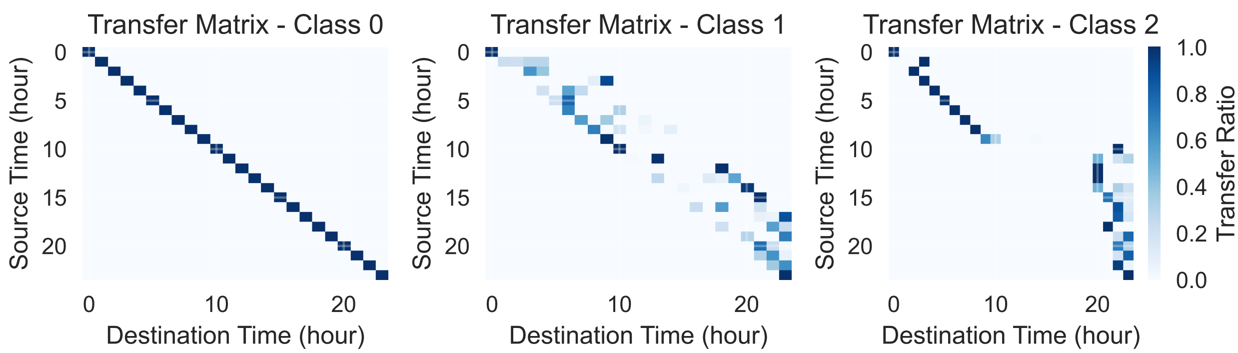


Fig. 2: Transfer matrices

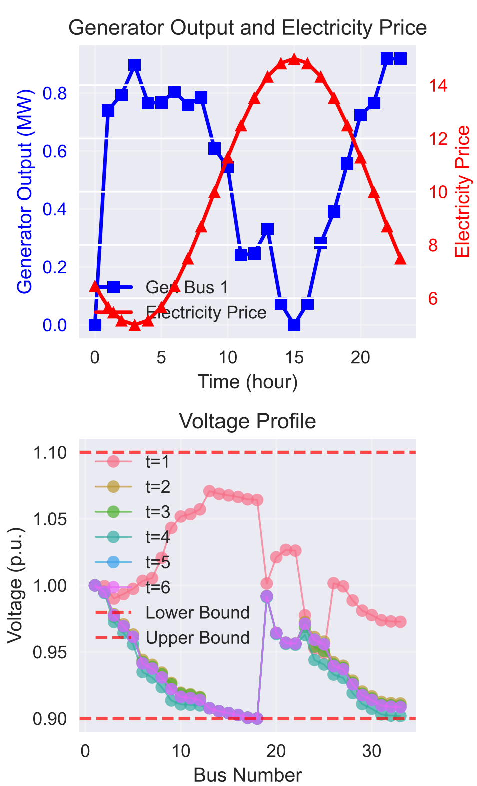
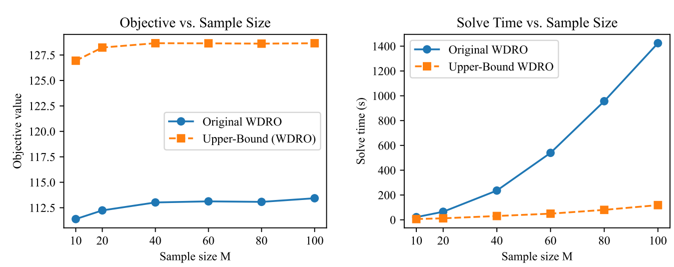


Fig. 3: Price, generator output, and voltages



(a) Objective vs. sample size M

(b) Solve time vs. sample size M

Fig. 4: Comparison between solve time and objective

B. Data and Uncertainty Modeling

We use $M = 10$ historical samples to construct the empirical distribution $\hat{\mathbb{P}}$. For Datacenter Loads (ξ^{DC}), we model $n_{\text{classes}} = 3$ task classes with a sinusoidal daily base workload split among them: Class 0 (Delay-sensitive, $h_0 = 0$ hours), Class 1 (Short delay-tolerant, $h_1 = 6$ hours), and Class 2 (Long delay-tolerant, $h_2 = 12$ hours). Stochastic scenarios are generated by adding random noise to this multi-class base profile. PV Generation (ξ^{PV}) scenarios are generated using a realistic bell-shaped profile peaking at 12:00 PM, with random noise applied to simulate weather-induced fluctuations. A dynamic electricity price signal λ is used, with prices lowest in the early morning (3:00 AM) and peaking in the afternoon (3:00 PM). The risk level for all CVaR constraints is set to

979-8-3315-8138-1

$\alpha = 0.05$, and the Wasserstein radii are set to $\epsilon_{\text{dc}} = 0.001$ for datacenter loads and $\epsilon_{\text{pv}} = 0.0005$ for PV generation.

C. Results and Analysis

Cost-Optimal Load Shifting. The primary objective of minimizing generation cost is successfully achieved through temporal workload shifting. The price plot in Fig. 3 shows high prices from t=12 to t=18. In response, the DC commit loads plot in Fig. 1b shows that the model schedules the vast majority of its workload during the low-price valleys, specifically t=0-6 (early morning) and t=20-24 (late night). The model actively minimizes workload during the expensive peak period, demonstrating effective cost arbitrage.

Exploitation of Temporal Flexibility. This workload shifting is enabled by the model's use of temporal flexibility. The transfer matrix heatmap in Fig. 2 visualizes the scheduling policy for the most flexible workloads (Class 2, 12h delay). The strong diagonal entries from t=2 to t=7 represent workload executed immediately upon submission since the low electricity price. Critically, the strong *off-diagonal* entries show workload being deferred. For example, workload submitted at t=8 (during rising prices) is clearly shifted for execution at t=20-22 (during the evening price drop). This explicitly visualizes the model exploiting the $h_c = 12$ hour delay tolerance.

Robust Grid-Aware Operation. The most significant finding is the model's ability to ensure robust grid security. Despite the high uncertainty in the input loads, the WDRO solution respects all critical constraints. Fig. 1b shows the realized workload (purple line), weighted by the random submitted loads ξ^{DC}, remains strictly below the final committed workload (black dashed line) at all times. This is a direct result of the robust CVaR constraint (12). Furthermore, Fig. 3 (bottom) shows the voltage magnitudes at all 33 buses for the first 6 hours of operation are within the secure operational bounds. The WDRO formulation, by optimizing against a worst-case distribution, successfully finds a scheduling policy that is not only low-cost but also prevents capacity and voltage violations.

Computational Acceleration. We conducted a comprehensive comparison between the proposed CVaR upper-bound approximation method and the original WDRO algorithm across various sample sizes (M=10 to M=100). As shown in Fig. 4a, the objective values obtained by both methods remain consistently close across all sample sizes. More significantly, Fig. 4b demonstrates that the computational time of the upper-bound method is reduced by two orders of magnitude, decreasing from approximately 1400 seconds for M=100 with the original algorithm to just 120 seconds with the proposed method, achieving a 11× speedup with minor optimality loss. This substantial efficiency gain makes the CVaR upper-bound approach practical for real-time applications while maintaining the robustness guarantees of the original WDRO formulation.

V. Conclusion

This paper proposes a WDRO framework to address the critical coordination challenge between AI datacenters and power distribution grids under significant uncertainty. Our approach explicitly model the joint operational risks using distributionally robust CVaR constraints. These constraints are systematically formulated to hedge against the volatility of both multi-class datacenter workloads and PV generation under the worst-case distribution. To mitigate the computational challenge posed by this formulation, we introduce a CVaR upper-bound approximation that eliminates sample-dependent auxiliary variables. This approximation achieves a great speedup while preserving robustness guarantees and incurring only minor optimality loss. Future work may explore a receding-horizon implementation for real-time operation.

References

[1] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, "Carbon emissions and large neural network training," *arXiv preprint arXiv:2104.10350*, 2021.

[2] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green ai," *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, 2020.

[3] C. Debus, M. Piraud, A. Streit, F. Theis, and M. Götz, "Reporting electricity consumption is essential for sustainable AI," *Nature Machine Intelligence*, vol. 5, no. 11, pp. 1176–1178, 2023.

[4] A. Faiz, S. Kaneda, R. Wang, R. Osi, P. Sharma, F. Chen, and L. Jiang, "LLMCarbon: Modeling the end-to-end carbon footprint of large language models," *arXiv preprint arXiv:2309.14393*, 2023.

[5] Z. Wu, L. Chen, J. Wang, M. Zhou, G. Li, and Q. Xia, "Incentivizing the Spatiotemporal Flexibility of Data Centers Toward Power System Coordination," *IEEE Transactions on Network Science and Engineering*, vol. 10, no. 3, pp. 1766–1778, 2023.

[6] B. Acun, B. Lee, F. Kazhamiaka, K. Maeng, U. Gupta, M. Chakkaravarthy, D. Brooks, and C.-J. Wu, "Carbon Explorer: A Holistic Framework for Designing Carbon Aware Datacenters," ser. International Conference on Architectural Support for Programming Languages and Operating Systems, 2023.

[7] A. Radovanović, R. Koningstein, I. Schneider, B. Chen, A. Duarte, B. Roy, D. Xiao, M. Haridasan, P. Hung, N. Care *et al.*, "Carbon-Aware Computing for Datacenters," *IEEE Transactions on Power Systems*, vol. 38, no. 2, pp. 1270–1280, 2022.

[8] W. Zhang, L. Roald, and V. Zavala, "Exploring the Impacts of Power Grid Signals on Data Center Operations Using a Receding-Horizon Scheduling Model," *IEEE Transactions on Sustainable Computing*, vol. 8, no. 2, pp. 245–256, 2023.

[9] P. Mohajerin Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, no. 1, pp. 115–166, 2018.

[10] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh, "Wasserstein distributionally robust optimization: Theory and applications in machine learning," in *Operations research & management science in the age of analytics*. Informs, 2019, pp. 130–166.

[11] S. Hall, F. Micheli, G. Belgioioso, A. Radovanović, and F. Dörfler, "Carbon-aware computing for data centers with probabilistic performance guarantees," *arXiv preprint arXiv:2410.21510*, 2024.

[12] M. E. Baran and F. F. Wu, "Network reconfiguration in distribution systems for loss reduction and load balancing," *IEEE Transactions on Power delivery*, vol. 4, no. 2, pp. 1401–1407, 2002.

[13] A. Arrigo, C. Ordoudis, J. Kazempour, Z. De Grève, J.-F. Toubeau, and F. Vallée, "Wasserstein distributionally robust chance-constrained optimization for energy and reserve dispatch: An exact and physically-bounded formulation," *European Journal of Operational Research*, vol. 296, no. 1, pp. 304–322, 2022.

[14] R. T. Rockafellar, S. Uryasev *et al.*, "Optimization of conditional value-at-risk," *Journal of risk*, vol. 2, pp. 21–42, 2000.

[15] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *Journal of Machine Learning Research*, vol. 17, no. 83, pp. 1–5, 2016.

[16] M. ApS, *The MOSEK Python Fusion API manual. Version 11.0.*, 2025. [Online]. Available: https://docs.mosek.com/latest/pythonfusion/index.html

979-8-3315-8138-1