An Online Algorithm for Data Center Demand Response

Shahab Bahrami, Yu Christine Chen, and Vincent W.S. Wong
Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, Canada
email: {bahramis, chen, vincentw}@ece.ubc.ca

***Abstract*—Data centers often support a range of delay-tolerant workloads with adjustable execution time under a prespecified service level agreement. This potential for workload management has motivated utility companies to deploy demand response programs to encourage data centers toward shifting workload execution away from peak load periods. In this paper, we focus on data centers' demand response considering the uncertainties in the arrival rates of the workloads, local renewable generation, and time-varying electricity prices. The centralized workload scheduling is shown to be a convex optimization problem. We deploy an online convex optimization framework to solve the centralized problem without any knowledge of the stochastic process that uncertain parameters follow. It also enables us to design a decentralized algorithm to address the high computational complexity of the centralized approach as well as the data centers' coupled decision making under the real-time pricing scheme. We perform extensive simulations to demonstrate the lower running time of the decentralized algorithm compared to the centralized approach. Data center demand response benefits the utility company by 12.4% reduction in the peak load demand. It also benefits a data center by 12.2% reduction in the average daily cost.**

I. Introduction

The issue of energy efficiency poses a crucial challenge to today's data centers owing to the growing requirements for data storage and analysis services. In this regard, practical energy management solutions are required for optimal workload execution while meeting the service level agreement (SLA) requirements [1]. The savings from reducing the electricity bill payment can be reinvested in information technology (IT) resources and expansion planning of the data center. Meanwhile, utility companies can benefit from lower power demand during peak time periods when the cost to produce or the price to purchase electricity is high. A demand response program with real-time pricing (RTP) scheme is an attractive solution for utility companies to steadily encourage data centers toward adjusting power demand through workload scheduling [2].

There are challenges in development and deployment of a demand response program for data centers. First, the uncertainty in the workloads' arrival rates forces a dynamic provisioning of virtual machines (servers) to optimally schedule the workloads execution under a prespecified SLA requirement. Second, most data centers are equipped with local renewable generators (e.g., photovoltaic (PV) panels, wind turbines) [3]. The uncertainty in the renewable generation makes it more challenging to achieve the optimal workload schedule. Third, the utility company and data centers require a real-time information exchange mechanism to deal with the dynamic prices changes in an RTP scheme.

Data center demand response is an active research area. Recent work has focused on the design of coordination mechanisms between the utility companies and data centers in a demand response program using different techniques such as Stackelberg game [4], reverse auction [5], and bargaining [6]. These papers assumed complete information about the workload characteristics and renewable generation. Other works focused on the uncertainty in workloads arrival rates and renewable genearion and proposed energy management solutions for data centers using robust optimization [7] and portfolio optimization [8], [9]. These studies relied on information about the stochastic process for uncertain parameters, which may not be available in practice. Moreover, there are works that deal with designing online energy management algorithms for data centers using stochastic optimization [10], reinforcement learning [11], and Lyapunov optimization [12]–[14]. These approaches often rely on limiting assumptions such as the Markovian characteristic of stochastic processes.

In this paper, we investigate the data center demand response problem by deploying an online convex optimization framework to deal with the uncertainty in the workload arrival rates and renewable generation. This paper is an extension of our previous work [15] by designing an online decentralized workload scheduling algorithm for data centers. The algorithm enables real-time workload scheduling in a parallel fashion based on the control signals from the utility company. This paper also extends our previous work [16] by considering a time-varying queue to model the workloads of a data center in the underlying online convex optimization framework. We also study the impact of a time-varying constraint associated with the required number of virtual machines on the design of an online workload scheduling algorithm. The key contributions of this paper are summarized as follows:

- *Addressing Uncertainty:* The stochastic processes for the workloads arrival rates and the renewable generation may not be available. We present an online convex optimization framework [17] to solve the centralized stochastic workloload scheduling problem. Thus, the utility company does not require information about the stochastic processes for the uncertain parameters. The utility company receives the realized values of the unknown parameters and determines the bill payment and the penalty for workload execution delay. Subsequently, the utility company schedules the virtual machines in the next time interval using a projected gradient method.
- *Decentralized Algorithm Design:* The proposed centralized workload scheduling method suffers from a high computational burden, and hence may not be viable for real-time implementation in a system with a large number of data centers. We address the computation complexity

by developing a decentralized algorithm, where the data centers solve their corresponding optimization problem in a parallel fashion. The proposed decentralized algorithm achieves the centralized problem's optimal solution with a significantly lower running time, especially as the number of data centers grows.

- *Reduction in Peak Load Demand and Average Daily Cost:* We conduct simulations on a test system with ten data centers participating in a demand response program. Simulation results for the scenarios with complete information show that the data center demand response benefits both the utility company by 12.4% decrease in the peak load demand, and a data center by 12.2% decrease in its average daily cost. The proposed algorithm with uncertainty enables the data centers to approximate the optimal number of virtual machines, such that they incur only a 2.15% higher average daily cost.

The remainder of this paper is organized as follows. Section II introduces the model of a data center. In Section III, we present an online convex optimization framework to solve the stochastic workload scheduling problem. We also develop a decentralized workload scheduling algorithm for data centers. The performance of the proposed decentralized algorithm is evaluated in Section IV. We conclude the paper in Section V.

II. System Model

Consider a network comprising a set $\mathcal{D} = \{1, \ldots, D\}$ of D data centers. We assume that the utility company and a data center can exchange information about the electricity price and power consumption through a bi-directional communication infrastructure. The time period is divided into T equal time slots. Let $\mathcal{T} = \{1, \ldots, T\}$ denote the set of time slots. Let $\mathcal{C}_d = \{1, \ldots, C_d\}$ denote the set of C_d classes of services supported by data center $d \in \mathcal{D}$. For data center d, we consider the exponential distribution with mean $1/\lambda_{c,d}(t)$ for the inter-arrival time of the workloads requesting service $c \in \mathcal{C}_d$ in time slot $t \in \mathcal{T}$. We also consider the exponential distribution for workload execution time in a data center [4], [9]. We apply the approach in [15] to obtain the workload's execution rates. Consider data center $d \in \mathcal{D}$. We define parameter $\sigma_{c,d}$ as the average time that a single virtual machine spends to execute a workload requesting service c. Hence, with $n_d(t)$, $t \in \mathcal{T}$ virtual machines, the workloads of class c are executed with a rate $\overline{\mu}_{c,d}(t) = n_d(t)/\sigma_{c,d}$ per time slot. As it is shown in [15], we can use an $M/M/1$ queue with the workloads' average arrival rate $\lambda_{c,d}(t)$ and execution rate $\mu_{c,d}(t) = \overline{\mu}_{c,d}(t)\big(1 - \sum_{c' \in \mathcal{C}_d} \lambda_{c',d}(t)/\overline{\mu}_{c',d}(t) + \lambda_{c,d}(t)/\overline{\mu}_{c,d}(t)\big)$ to characterize the workloads requesting service $c \in \mathcal{C}_d$ in time slot $t \in \mathcal{T}$. Using $\overline{\mu}_{c,d}(t) = n_d(t)/\sigma_{c,d}$, we can express the execution rate $\mu_{c,d}(t)$, $c \in \mathcal{C}_d$ as follows:

$$\mu_{c,d}(t) = \frac{1}{\sigma_{c,d}} \Big(n_d(t) - \textstyle\sum_{c' \in \mathcal{C}_d} \sigma_{c',d}\, \lambda_{c',d}(t) \Big) + \lambda_{c,d}(t). \quad (1)$$

The underlying $M/M/1$ queueing system is stable if $\mu_{c,d}(t) > \lambda_{c,d}(t)$. That is, the first term in (1) should be positive. Hence, for small $\varepsilon > 0$ (e.g., $\varepsilon = 1$), we obtain

$$\sum_{c \in \mathcal{C}_d} \sigma_{c,d}\, \lambda_{c,d}(t) + \varepsilon \leq n_d(t), \quad d \in \mathcal{D},\, t \in \mathcal{T}. \quad (2)$$

Considering the maximum number of available virtual machines $n_d^{\max}$ in data center $d \in \mathcal{D}$, in time slot t, we have [15]

$$0 \leq n_d(t) \leq n_d^{\max}. \quad (3)$$

The offered services in a data center can be divided into interactive services and delay-tolerant flexible services according to the workload maximum *sojourn* time under the SLA between a data center and its customers. The interactive services should be executed in a timely fashion. Whereas, the delay-tolerant flexible services can tolerate a relatively large execution time (e.g., several minutes) [2]. For data center d, let $\delta_{c,d}$ denote the maximum sojourn time for the workloads requesting service c. Without loss of generality, we assume that $\delta_{c,d}$, $c \in \mathcal{C}_d$ is less than one time slot. We can apply the workload model in [15] to deal with the execution time greater than one time slot. Let $\tau_{c,d}(t)$ denote the sojourn time of a workload requesting service c from data center d in time slot t. We define the decision variable $p_{c,d}(t) \in [0,\, 1]$ as the *upper bound* for the probability that $\tau_{c,d}(t)$ exceeds $\delta_{c,d}$. We have

$$\Pr\left(\tau_{c,d}(t) \geq \delta_{c,d}\right) \leq p_{c,d}(t), \;\; c \in \mathcal{C}_d,\, d \in \mathcal{D},\, t \in \mathcal{T}, \quad (4)$$

where $\Pr(\cdot)$ is the probability function. Consider the queue for the workloads requesting service c in time slot t. The sojourn time of the workloads follows an exponential distribution with mean $\mu_{c,d}(t) - \lambda_{c,d}(t)$ [18]. Then (4) for $c \in \mathcal{C}_d$, $d \in \mathcal{D}$, $t \in \mathcal{T}$, can be expressed as

$$\exp\left(-\delta_{c,d}\big(\mu_{c,d}(t) - \lambda_{c,d}(t)\big)\right) \leq p_{c,d}(t), \quad (5)$$

where $\exp(\cdot)$ is the exponential function. By substituting (1) into (5) and performing some algebraic manipulations, for $c \in \mathcal{C}_d$, $d \in \mathcal{D}$, $t \in \mathcal{T}$, we obtain

$$\frac{\sigma_{c,d}}{\delta_{c,d}} \ln\left(\frac{1}{p_{c,d}(t)}\right) + \sum_{c' \in \mathcal{C}_d} \sigma_{c',d}\, \lambda_{c',d}(t) \leq n_d(t), \quad (6)$$

where $\ln(\cdot)$ is the natural logarithm function. To express (6) as a linear inequality, we define an auxiliary variable $\alpha_{c,d}(t) = \ln\big(1/p_{c,d}(t)\big)$, $c \in \mathcal{C}_d$, $t \in \mathcal{T}$. Constraint (6), for $c \in \mathcal{C}_d$, $d \in \mathcal{D}$, $t \in \mathcal{T}$, can be rewritten as follows:

$$\frac{\sigma_{c,d}}{\delta_{c,d}}\, \alpha_{c,d}(t) + \sum_{c' \in \mathcal{C}_d} \sigma_{c',d}\, \lambda_{c',d}(t) \leq n_d(t). \quad (7)$$

The power demand $P_d^{\mathrm{w}}(t)$, $t \in \mathcal{T}$ for workload execution in data center $d \in \mathcal{D}$ can be obtained in terms of the average idle power rating P_d^{idle} and the peak power rating P_d^{peak} of a virtual machine. For data center d, let $\eta_d(t) > 1$ denote the power usage effectiveness in time slot $t \in \mathcal{T}$ [3]. We have [15]

$$P_d^{\mathrm{w}}(t) = \eta_d(t) \Big(P_d^{\text{idle}} n_d(t) + (P_d^{\text{peak}} - P_d^{\text{idle}}) \sum_{c \in \mathcal{C}_d} \sigma_{c,d} \lambda_{c,d}(t) \Big). \quad (8)$$

Suppose that data center d has local renewable generation [3] with the output power $P_d^{\mathrm{r}}(t)$, $t \in \mathcal{T}$. The net power

consumption of data center d can be expressed as:

$$P_d^{\text{net}}(t) = \left[P_d^{\text{w}}(t) - P_d^{\text{r}}(t)\right]^+, \quad d \in \mathcal{D},\, t \in \mathcal{T}, \tag{9}$$

where $[\cdot]^+ = \max\{0, \cdot\}$. Data center d incurs the electricity bill payment $c_d^{\text{b}}(t)$ and the penalty $c_d^{\text{p}}(t)$ for the delay in executing the incoming workloads. The expected total cost per time slot for a data center can be obtained as follows:

$$c_d(t) = c_d^{\text{b}}(t) + c_d^{\text{p}}(t), \quad d \in \mathcal{D},\, t \in \mathcal{T}. \tag{10}$$

For the bill payment component in (10), o encourage data centers toward workload scheduling, an RTP scheme with inclining block tariffs structure using two block rates $\pi_1(t)$ and $\pi_2(t)$ can be deployed [2]. Let $l^{\text{th}}(t)$ denote the threshold value for the aggregate demand $P^{\text{net}}(t) = \sum_{d\in\mathcal{D}} P_d^{\text{net}}(t)$. If $P^{\text{net}}(t) \leq l^{\text{th}}(t)$, then the bill payment in time slot $t \in \mathcal{T}$ for data center $d \in \mathcal{D}$ in (10) is obtained as

$$c_d^{\text{b}}(t) = P_d^{\text{net}}(t)\pi_1(t). \tag{11}$$

If $P^{\text{net}}(t) > l^{\text{th}}(t)$, then for $d \in \mathcal{D}$ and $t \in \mathcal{T}$, we have

$$c_d^{\text{b}}(t) = \frac{l^{\text{th}}(t)}{P^{\text{net}}(t)} P_d^{\text{net}}(t)\pi_1(t) + \frac{P^{\text{net}}(t) - l^{\text{th}}(t)}{P^{\text{net}}(t)} P_d^{\text{net}}(t)\pi_2(t). \tag{12}$$

For the penalty component in (10), suppose that data center d is charged a penalty $\pi_{c,d}^{\text{p}}(t)$, $c \in \mathcal{C}_d$ if the workload's request is delayed more than $\delta_{c,d}$. The values of $\pi_{c,d}^{\text{p}}(t)$, $c \in \mathcal{C}_d$, $t \in \mathcal{T}$ are set according to the SLA requirements between the data center d and its customers. For service c in data center d, the average number of workloads that their requests are delayed more than $\delta_{c,d}$ is $\lambda_{c,d}(t)\exp(-\delta_{c,d}(\mu_{c,d}(t) - \lambda_{c,d}(t)))$. Hence, the *expected* penalty of data center $d \in \mathcal{D}$ for not meeting the SLA requirements in time slot $t \in \mathcal{T}$ is obtained as follows:

$$c_d^{\text{p}}(t) = \sum_{c\in\mathcal{C}_d} \pi_{c,d}^{\text{p}}(t)\lambda_{c,d}(t)\exp\big(-\delta_{c,d}\big(\mu_{c,d}(t) - \lambda_{c,d}(t)\big)\big). \tag{13}$$

Substituting (1) into (13), for $d \in \mathcal{D}$, $t \in \mathcal{T}$, we obtain

$$c_d^{\text{p}}(t) = \sum_{c\in\mathcal{C}_d} \pi_{c,d}^{\text{p}}(t)\lambda_{c,d}(t)\exp\Big(-\frac{\delta_{c,d}}{\sigma_{c,d}}\Big(n_d(t) - \sum_{c'\in\mathcal{C}_d}\sigma_{c',d}\,\lambda_{c',d}(t)\Big)\Big). \tag{14}$$

The exponential function is strictly increasing. Considering inequality (7), for cost minimization, we can rewrite (14) as

$$c_d^{\text{p}}(t) = \sum_{c\in\mathcal{C}_d} \pi_{c,d}^{\text{p}}(t)\,\lambda_{c,d}(t)\,\exp\big(-\alpha_{c,d}(t)\big). \tag{15}$$

III. Problem Formulation

In this section, we formulate the workload scheduling problem for data centers. The intermittent renewable generation and workloads' arrival rates lead to uncertain net power demand in a data center. Furthermore, the data centers' workload scheduling decisions are coupled through the bill payment in (11) and (12).

A. Centralized Workload Scheduling

Suppose that the utility company has complete information about the output power $P_d^{\text{r}}(t)$ of renewable generator and the workload's arrival rate $\lambda_{c,d}(t)$, $c \in \mathcal{C}_d$ for data center $d \in \mathcal{D}$. The objective function of the utility company is the *expected social cost*. Using (11) and (12), the total bill payment $\sum_{d\in\mathcal{D}} c_d^{\text{b}}(t)$, $t \in \mathcal{T}$ can be expressed as a piecewise linear function of the aggregate demand. We express the objective function of the utility company in time slot t as follows:

$$f^{\text{obj}}(t) = \max\Big\{\pi_1(t)\sum_{d\in\mathcal{D}} P_d^{\text{net}}(t),\, \pi_2(t)\sum_{d\in\mathcal{D}} P_d^{\text{net}}(t) - \omega(t)\Big\} + \sum_{d\in\mathcal{D}} c_d^{\text{p}}(t), \tag{16}$$

where $\omega(t) = l^{\text{th}}(t)(\pi_2(t) - \pi_1(t))$. We introduce auxiliary variables $\theta_d(t)$, $d \in \mathcal{D}$, $t \in \mathcal{T}$ associated with the projection in (9). We also introduce the auxiliary variable $\vartheta(t)$, $t \in \mathcal{T}$ for the piecewise linear term in (16). With these in place, the objective function (16) can be rewritten as follows:

$$\widetilde{f}^{\,\text{obj}}(t) = \vartheta(t) + \sum_{d\in\mathcal{D}} c_d^{\text{p}}(t). \tag{17}$$

The following constraints are included into the constraint set of the centralized workload scheduling problem:

$$\pi_1(t)\textstyle\sum_{d\in\mathcal{D}}\theta_d(t) \leq \vartheta(t), \qquad t \in \mathcal{T}, \tag{18a}$$

$$\pi_2(t)\textstyle\sum_{d\in\mathcal{D}}\theta_d(t) - \omega(t) \leq \vartheta(t), \qquad t \in \mathcal{T}, \tag{18b}$$

$$P_d^{\text{w}}(t) - P_d^{\text{r}}(t) \leq \theta_d(t), \quad d \in \mathcal{D},\, t \in \mathcal{T}, \tag{18c}$$

$$0 \leq \theta_d(t), \quad d \in \mathcal{D},\, t \in \mathcal{T}. \tag{18d}$$

In time slot t, we denote the decision vector of data center d by $\boldsymbol{\phi}_d(t) = \big((\alpha_{c,d}(t),\, c \in \mathcal{C}_d),\, n_d(t)\big)$. The centralized workload scheduling problem in time slot $t \in \mathcal{T}$ is as follows:

$$\boldsymbol{\mathcal{P}}_1(t): \underset{\vartheta(t),\,\boldsymbol{\phi}_d(t),\,\theta_d(t),\,d\in\mathcal{D}}{\text{minimize}} \quad \widetilde{f}^{\,\text{obj}}(t)$$
$$\text{subject to constraints (2), (3), (7), (8), and (18a)}-\text{(18d)}.$$

The utility company can solve the convex optimization problem $\boldsymbol{\mathcal{P}}_1(t)$ with complete information about the uncertain parameters. In practice, however, the utility company has uncertainty about the workloads' arrival rates and renewable generation in a data center. We use an online convex optimization framework [17] to deal with the uncertainty in the output power $P_d^{\text{r}}(t)$, $t \in \mathcal{T}$ of the renewable generator and the workload's average arrival rates $\lambda_{c,d}(t)$, $c \in \mathcal{C}_d$, $t \in \mathcal{T}$ for data center $d \in \mathcal{D}$. The utility company schedules the number of virtual machines $n_d(t)$, $t \in \mathcal{T}$ on behalf of data center d. The *realized* values of the uncertain parameters are revealed to the utility company during time slot t. Under the given number of virtual machines $n_d(t)$, $d \in \mathcal{D}$, the utility company solves the following convex optimization problem to obtain $\theta_d(t)$, $\alpha_{c,d}(t)$, $c \in \mathcal{C}_d$, $d \in \mathcal{D}$, and $\vartheta(t)$:

$$\boldsymbol{\mathcal{P}}_2(t): \underset{\vartheta(t),\,\theta_d(t),\,\alpha_{c,d}(t),\,c\in\mathcal{C}_d,\,d\in\mathcal{D}}{\text{minimize}} \quad \widetilde{f}^{\,\text{obj}}(t)$$
$$\text{subject to constraints (7), (8), and (18a)}-\text{(18d)}.$$

Subsequently, the utility company computes the number of virtual machines $n_d(t+1)$, $d \in \mathcal{D}$ for time slot $t+1$ in two steps. First, it determines the gradient of the objective

function with respect to $n_d(t)$. Let $\gamma_{1,d}(t)$ and $\gamma_{2,c,d}(t)$, $c \in \mathcal{C}_d$, and $\gamma_{3,d}(t)$ denote the dual variables associated with constraints (2), (7), and (18c) in problem $\boldsymbol{\mathcal{P}}_1(t)$, respectively. Substituting (8) into (18c), the utility company can compute the gradient of the objective function with respect to $n_d(t)$ as $\nabla_{n_d(t)} \widetilde{f}^{\,\text{obj}}(t) = -\gamma_{1,d}(t) - \sum_{c\in\mathcal{C}_d} \gamma_{2,c,d}(t) + \gamma_{3,d}(t)\, \eta_d(t) P_d^{\text{idle}}$. Next, the utility company can update $n_d(t)$ using the following projected gradient-based update rule [19]:

$$n_d(t+1) = \Big[n_d(t) - \xi_d(t) \nabla_{n_d(t)} \widetilde{f}^{\,\text{obj}}(t) \Big]_{\wp}, \tag{19}$$

where $\xi_d(t) > 0$ is a diminishing step size in time slot t, and $[\cdot]_{\wp}$ is the projection onto the interval defined by (3). The utility company can compute dual variables $\gamma_{2,c,d}(t)$, $c \in \mathcal{C}_d$ and $\gamma_{3,d}(t)$ by solving problem $\boldsymbol{\mathcal{P}}_2(t)$. However, the *time-varying* constraint (2) is not included in the constraints set of problem $\boldsymbol{\mathcal{P}}_2(t)$. To determine $\gamma_{1,d}(t)$, we rewrite (2) in the form $g_d(n_d(t)) \le 0$, where $g_d(n_d(t)) = \sum_{c'\in\mathcal{C}_d} \sigma_{c',d}\, \lambda_{c',d}(t) - n_d(t)$. The first-order approximation of function $g(n_d(t+1))$ around $n_d(t)$ can be obtained as

$$\begin{aligned} \Gamma_d(t) &= g_d(n_d(t)) + \nabla_{n_d(t)}\, g_d\big(n_d(t)\big)\big(n_d(t+1) - n_d(t)\big) \\ &= \sum_{c\in\mathcal{C}_d} \sigma_{c,d}\, \gamma_{c,d}(t) - n_d(t+1). \end{aligned} \tag{20}$$

The utility company updates $\gamma_{1,d}(t)$ as follows [20]:

$$\gamma_{1,d}(t+1) = \Big[\gamma_{1,d}(t) + \zeta_d(t)\, \Gamma_d(t) \Big]^{+}, \tag{21}$$

where $\zeta_d(t) > 0$ is a diminishing step size in time slot t.

Remark 1: Consider the optimal value $\widetilde{f}^{\,\text{obj,opt}}(t)$ of problem $\boldsymbol{\mathcal{P}}_1(t)$. We can use the results in [17, Ch. 5] and [20] for a Lipschitz and convex function $\widetilde{f}^{\,\text{obj}}(t)$ and linear function $g_d(n_d(t))$ as well as diminishing step sizes $\xi_d(t)$ and $\zeta_d(t)$ to guarantee that solving problem $\boldsymbol{\mathcal{P}}_2(t)$ and performing the updates in (19) and (21) achieve a sublinear regret $R(T) = \sum_{t\in\mathcal{T}} (\widetilde{f}^{\,\text{obj}}(t) - \widetilde{f}^{\,\text{obj,opt}}(t))$ and sublinear feasibility functions $F_d(T) = [\sum_{t\in\mathcal{T}} g_d(n_d(t))]^{+}$, $d \in \mathcal{D}$. In other words, $R(T)/T$ and $F_d(T)/T$, $d \in \mathcal{D}$ tend to zero as T approaches infinity.

B. Decentralized Algorithm Design

In a centralized approach, the utility company requires complete information about the *realized* values of the renewable generation and workloads' arrival rate for all data centers. In practice, however, these information may not be available to the utility company. Under the given number of virtual machines $n_d(t)$, $d \in \mathcal{D}$, we can decompose problem $\boldsymbol{\mathcal{P}}_2(t)$ into D optimization problems corresponding to data centers $d \in \mathcal{D}$. This enables us to design the decentralized Algorithm 1 in order to solve $\boldsymbol{\mathcal{P}}_2(t)$ along with the updates (19) and (21) in a distributed fashion. In Algorithm 1, Line 1 describes the initiation phase. In Line 2, data center $d \in \mathcal{D}$ observes the realized $\lambda_{c,d}(t)$, $c \in \mathcal{C}_d$ and computes $P_d^{\text{w}}(t)$ using (8). Data center $d \in \mathcal{D}$ also observes the realized renewable generation $P_d^{\text{r}}(t)$ and determines $\theta_d(t)$ as follows:

$$\theta_d(t) = \Big[P_d^{\text{w}}(t) - P_d^{\text{r}}(t) \Big]^{+}. \tag{22}$$

Algorithm 1 Interactions between Data Center $d \in \mathcal{D}$ and Utility Company in Time Slot $t \in \mathcal{T}$.

1: Data center d randomly sets $n_d(1)$ and sets $\gamma_{1,d}(t) = 0$ in time slot $t := 1$.
2: Data center d determines $\theta_d(t)$ according to (22) and sends it to the utility company.
3: Data center d solves problem $\boldsymbol{\mathcal{P}}_{2,d}(t)$ to obtain $\alpha_{c,d}(t), c \in \mathcal{C}_d$.
4: Utility company determines $\vartheta(t)$ according to (23).
5: Utility company sends the control signal $\pi(t)$ to the data centers.
6: Data center d determines $n_d(t+1)$ using (19).
7: Data center d computes $\gamma_{1,d}(t+1)$ using (21).

Data center $d \in \mathcal{D}$ sends $\theta_d(t)$ to the utility company. It solves the following optimization problem to determine $\alpha_{c,d}(t)$, $c \in \mathcal{C}_d$ in Line 3:

$$\begin{aligned} \boldsymbol{\mathcal{P}}_{2,d}(t): \quad & \underset{\alpha_{c,d}(t) \ge \varepsilon\, \delta_{c,d}/\sigma_{c,d},\ c\in\mathcal{C}_d}{\text{minimize}} \quad c_d^{\text{p}}(t) \\ & \text{subject to constraint (7).} \end{aligned}$$

Constraint $\alpha_{c,d}(t) \ge \varepsilon\, \delta_{c,d}/\sigma_{c,d}$, $c \in \mathcal{C}_d$ is necessary to guarantee the stability of the queue associated with the workloads requesting service $c \in \mathcal{C}_d$ in time slot t. In Line 4, the utility company determines $\vartheta(t)$ as follows:

$$\vartheta(t) = \max\left\{ \pi_1(t) \sum_{d\in\mathcal{D}} \theta_d(t),\ \pi_2(t) \sum_{d\in\mathcal{D}} \theta_d(t) - \omega(t) \right\}. \tag{23}$$

In Line 5, the utility company sets the control signal $\pi(t)$ to $\pi_1(t)$ if $\sum_{d\in\mathcal{D}} \theta_d(t) \le l^{\text{th}}(t)$ or to $\pi_2(t)$ if $\sum_{d\in\mathcal{D}} \theta_d(t) > l^{\text{th}}(t)$. It broadcasts $\pi(t)$ to data center $d \in \mathcal{D}$. In Line 6, data center $d \in \mathcal{D}$ computes $\gamma_{2,c,d}(t)$, $c \in \mathcal{C}_d$ (from solving problem $\boldsymbol{\mathcal{P}}_{2,d}(t)$) and sets $\gamma_{3,d}(t)$ to the control signal $\pi(t)$. It determines the updated number of virtual machines in the next time slot using (19). Data center $d \in \mathcal{D}$ uses (21) to update the dual variable $\gamma_{1,d}(t)$ in Line 7. Using Algorithm 1, data centers achieve the solution to problem $\boldsymbol{\mathcal{P}}_2(t)$.

IV. Performance Evaluation

The system level performance evaluation of Algorithm 1 consists of ten data centers, and a time period of 10 days, each of which is divided into 96 equal time slots with duration of 15 minutes. Fig. 1(a) shows the electricity price block rates $\pi_1(t)$ and $\pi_2(t)$ during one day. Parameter l^{th} is set to 9 MW in each time slot. Penalties $\pi_{c,d}^{\text{p}}(t)$, $c \in \mathcal{C}_d$, $d \in \mathcal{D}$, $t \in \mathcal{T}$ are chosen uniformly within the interval [5 cents, 15 cents]. The nominal power rating of the PV plant for each data center is chosen uniformly within the interval [0.5 MW, 1 MW]. To obtain the PV plant's daily generation pattern for each data center, the historical PV generation data for Ontario, Canada power grid database from July 1, 2018 to July 10, 2018 [21] is used. Fig. 1(b) depicts the average PV generation. A data center offers 5 classes with the average workload arrival rates given in the World Cup 98 web hits dataset [22]. We choose $\delta_{c,d}$, $c \in \mathcal{C}_d$, $d \in \mathcal{D}$ uniformly within the interval [0.01 sec, 600 sec]. In a data center, the maximum number of virtual machines is chosen uniformly within the interval [3500, 5000] with power ratings $P_d^{\text{idle}} = 100\,\text{W}$ and $P_d^{\text{peak}} = 200\,\text{W}$ per time slot.

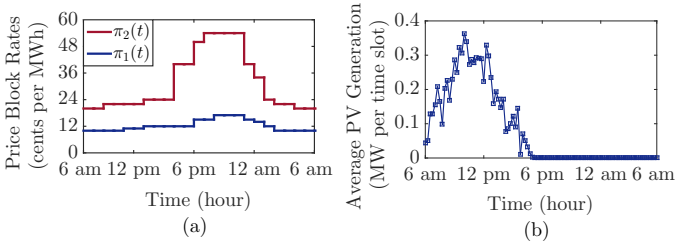

Figure 1. (a) Tariff block rates; (b) Average PV generation during one day.

Parameters $\sigma_{c,d}$, $c \in \mathcal{C}_d$, $d \in \mathcal{D}$ are chosen uniformly within the interval $[1\,\text{sec}, 100\,\text{sec}]$ under the condition that $\sigma_{c,d}$ is smaller for class c with smaller $\delta_{c,d}$.

We evaluate the performance of Algorithm 1 in executing the workload of a data center. In the scenario without workload scheduling, data center $d \in \mathcal{D}$ sets the probabilities $p_{c,d}(t)$ to 0.01 for class $c \in \mathcal{C}_d$ and time slot $t \in \mathcal{T}$. That is, with probability of 99%, the SLA for the workload maximum execution time is met. In the scenario with workload scheduling and complete information, the utility company solves problem $\boldsymbol{\mathcal{P}}_1(t)$. In the scenario with workload scheduling and incomplete information, data centers use Algorithm 1 in a decentralized fashion. The step sizes in (19) and (21) are set to $\xi_d(t) = 400/\sqrt{t}$ and $\zeta_d(t) = 10/\sqrt{t}$, respectively. Figs. 2(a) and (b) show the number of virtual machines in data center 1 in the aforementioned scenarios during day 1 and day 10, respectively. With workload scheduling, the number of virtual machines decreases during the time period with high price block rates (between 5 pm and 10 pm). Using Algorithm 1, the number of virtual machines tightly approximates the optimal number of virtual machines obtained from solving problem $\boldsymbol{\mathcal{P}}_1(t)$ with complete information. The approximation is tighter in day 10 compared to day 1. That is, Algorithm 1 can better follow the fluctuations in the optimal number of virtual machines. Fig. 2(c) shows that the probabilities $p_{c,d}(t)$ becomes larger (up to 0.1) during the time period with high price block rates. That is, reducing the number of virtual machines causes the delay in workload execution exceeds the threshold value with higher probability. The data center achieves a trade-off between decreasing the power consumption during the time period with high prices and the penalty for not meeting the SLA requirements workload execution. Fig. 2(d) shows that the data centers' aggregate demand during peak hours is reduced by about 12.4% (from 11.5 MW to 10.2 MW) on average with workload scheduling. Decreasing the number of virtual machines during off-peak hours is not beneficial, since the penalty (15) becomes large for not meeting the SLA requirements. Additionally, during 6 am to 5 pm, a data center benefits from PV plant to partially supply its load demand.

A data center's cost is reduced with workload scheduling. Fig. 3(a) depicts the average daily cost of data center 1 in the aforementioned cases during day 1 to day 10. When compared with the scenario without workload scheduling, the daily cost of data center 1 is reduced by 12.2% on average with workload scheduling and complete information. Using Algorithm 1 in the scenario with incomplete information, the daily cost is only

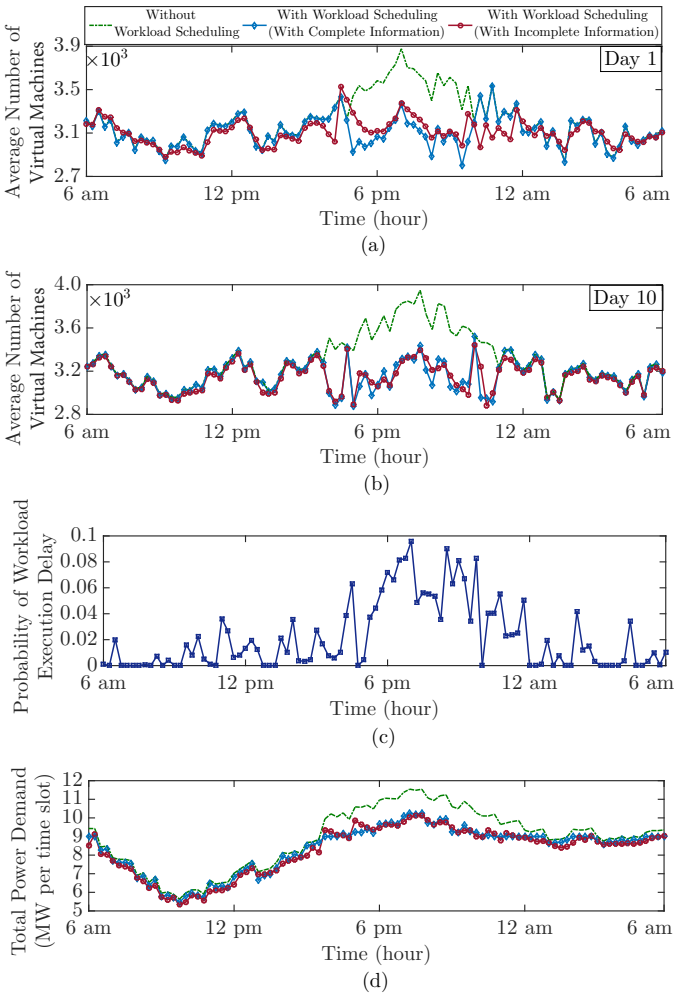

Figure 2. (a) Average number of virtual machines in data center 1 during day 1; (b) Average number of virtual machines in data center 1 during day 10; (c) The probability of workload execution delay for data center 1 during one day; (d) The average total power demand of all data centers during one day.

2.15% larger (on average) compared with the optimal daily cost in the scenario with complete information. Fig.3(b) shows the average cost per time slot of data center 1. It confirms that the gap between the cost obtained from Algorithm 1 and the optimal cost decreases gradually. In other words, the average regret $R(T')/T'$ converges to zero as T' increases from 1 to 960 time slots. Additionally, Fig. 3(c) depicts the average regret for data center 1, which confirms the result of Remark 1 that the regret of a data center grows sublinearly in T'.

Finally, we discuss the running time of Algorithm 1. The algorithm's running time per time slot is about 1.32 seconds on average. Data centers execute Algorithm 1 in a distributed and parallel manner. Thus, the running time per time slot of Algorithm 1 is independent of the number of data centers. In the centralized approach, however, the utility company solves problem $\boldsymbol{\mathcal{P}}_2(t)$ on behalf of all data centers. Although $\boldsymbol{\mathcal{P}}_2(t)$ is a convex optimization problem, it has a nonlinear objective function due the exponential penalty function in (15). Fig. 4 shows the required average running time to solve problem $\boldsymbol{\mathcal{P}}_2(t)$ versus the number of data centers. The results

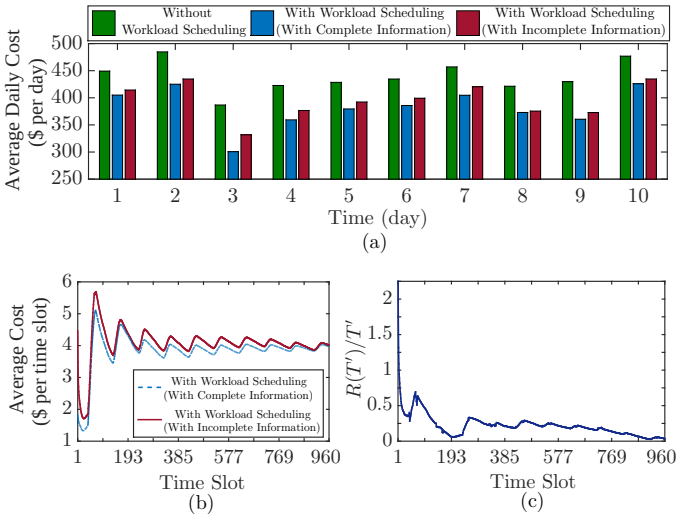

Figure 3. (a) Average daily cost of data center 1 during day 1 to day 10; (b) Average cost per time slot for data center 1; (c) Average regret of data center 1 in time slots 1 to 960.

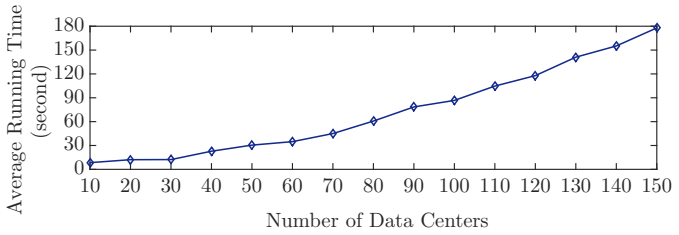

Figure 4. Average running time of the centralized approach. Note that the running time of Algorithm 1 is independent of the number of data centers.

indicate that implementing the centralized algorithm for *online* workload scheduling can be impractical for a system with a large number of data centers.

V. Concluding Remarks

In this paper, we studied the workload scheduling problem for data centers in a demand response program. The deterministic centralized workload scheduling was formulated as a convex optimization problem, where the utility company aims to jointly minimize the expected bill payment and the penalty associated with delaying the workloads execution for all data centers. We deployed an online convex optimization framework to enable workload scheduling under uncertainty in the workloads arrival rates and renewable generation. We also developed an algorithm that enables data centers to schedule their workloads in a decentralized manner. By simulations, we showed that the data center demand response benefits the utility company by 12.4% reduction in the aggregate power demand during peak hours. A data center also can benefit from 12.2% reduction in its average daily cost. The proposed online decentralized algorithm can tightly approximate the optimal workload scheduling with complete information. In particular, the difference between the total daily cost of data centers with incomplete information and the optimal cost with complete information is 2.15%. Results also verify the sublinear growth of the data centers' regret. The proposed decentralized algorithm has the advantage of lower average running time compared with the centralized approach. For future work, we will consider the operating constraints imposed by the power network in the system model.

References

[1] W. Zhang, Y. Wen, Y. W. Wong, K. C. Toh, and C. Chen, "Towards joint optimization over ICT and cooling systems in data centre: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1596–1616, 3rd quarter 2016.

[2] Z. Liu, I. Liu, S. Low, and A. Wierman, "Pricing data center demand response," in *Proc. of ACM Int'l Conf. on Measurement and Modeling of Computer Systems*, Austin, TX, Jun. 2014.

[3] M. Dayarathna, Y. Wen, and R. Fan, "Data center energy consumption modeling: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 732–794, 1st quarter 2016.

[4] G. S. Aujla, M. Singh, N. Kumar, and A. Zomaya, "Stackelberg game for energy-aware resource allocation to sustain data centers using RES," accepted for publication in *IEEE Trans. on Cloud Computing*, Jun. 2017.

[5] L. Zhang, S. Ren, C. Wu, and Z. Li, "A truthful incentive mechanism for emergency demand response in colocation data centers," in *Proc. of IEEE Conf. on Computer Communications (INFOCOM)*, Kowloon, Hong Kong, April 2015.

[6] X. Cao, J. Zhang, and H. V. Poor, "Data center demand response with on-site renewable generation: A bargaining approach," *IEEE/ACM Trans. on Networking*, vol. 26, no. 6, pp. 2707–2720, Dec. 2018.

[7] L. Shi, Y. Shi, X. Wei, X. Ding, and Z. Wei, "Cost minimization algorithms for data center management," *IEEE Trans. on Parallel and Distributed Systems*, vol. 28, no. 1, pp. 60–71, Jan. 2017.

[8] D. Paul, W. Zhong, and S. K. Bose, "Demand response in data centers through energy-efficient scheduling and simple incentivization," *IEEE Systems Journal*, vol. 11, no. 2, pp. 613–624, Jun. 2017.

[9] M. Ghamkhari, A. Wierman, and H. Mohsenian-Rad, "Energy portfolio optimization of data centers," *IEEE Trans. on Smart Grid*, vol. 8, no. 4, pp. 1898–1910, Jul. 2017.

[10] S. Kwon, L. Ntaimo, and N. Gautam, "Demand response in data centers: Integration of server provisioning and power procurement," accepted for publication in *IEEE Trans. on Smart Grid*, Sept. 2018.

[11] J. Yang, S. Zhang, X. Wu, Y. Ran, and H. Xi, "Online learning-based server provisioning for electricity cost reduction in data center," *IEEE Trans. on Control Systems Technology*, vol. 25, no. 3, pp. 1044–1051, May 2017.

[12] Y. Guo, M. Pan, Y. Gong, and Y. Fang, "Dynamic multi-tenant coordination for sustainable colocation data centers," accepted for publication in *IEEE Trans. on Cloud Computing*, Apr. 2017.

[13] M. A. Islam, S. Ren, A. H. Mahmud, and G. Quan, "Online energy budgeting for cost minimization in virtualized data center," *IEEE Trans. on Services Computing*, vol. 9, no. 3, pp. 421–432, May 2016.

[14] L. Yu, T. Jiang, and Y. Zou, "Distributed real-time energy management in data center microgrids," *IEEE Trans. on Smart Grid*, vol. 9, no. 4, pp. 3748–3762, Jul. 2018.

[15] S. Bahrami, V.W.S. Wong, and J. Huang, "Data center demand response in deregulated electricity markets," accepted for publication in *IEEE Trans. on Smart Grid*, Mar. 2018.

[16] S. Bahrami, Y.C. Chen, and V.W.S. Wong, "An autonomous demand response algorithm based on online convex optimization," in *Proc. of IEEE SmartGridComm*, Aalborg, Denmark, Oct. 2018.

[17] E. Hazan, "Introduction to online convex optimization," *Foundations & Trends in Optimization*, vol. 2, no. 3-4, pp. 157–325, Aug. 2016.

[18] T. G. Robertazzi, *Computer Networks and Systems: Queueing Theory and Performance Evaluation*, 3rd ed. Springer, 2000.

[19] R. Jenatton, J. C. Huang, and C. Archambeau, "Adaptive algorithms for online convex optimization with long-term constraints," in *Proc. of Int'l Conf. on Machine Learning*, New York, NY, Jun. 2016.

[20] T. Chen and G. B. Giannakis, "Bandit convex optimization for scalable and dynamic IoT management," accepted for publication in *IEEE Internet of Things Journal*, May 2018.

[21] Ontario's Independent Electricity System Operator Dataset. [Online]. Available: http://www.ieso.ca/en/Power-Data

[22] M. Arlitt and T. Jin, "Workload characterization of the 1998 world Cup website," Internet Systems and Applications Laboratory, Tech. Rep. HPL-1999-35(R.1), Sept. 1999.