Energy Portfolio Optimization of Data Centers

Mahdi Ghamkhari, *Student Member, IEEE*, Adam Wierman,
Member, IEEE, and Hamed Mohsenian-Rad, *Senior Member, IEEE*

***Abstract*—Data centers have diverse options to procure electricity. However, the current literature on exploiting these options is very fractured. Specifically, it is still not clear how utilizing one energy option may affect selecting other energy options. To address this open problem, we propose a unified energy portfolio optimization framework that takes into consideration a broad range of energy choices for data centers. Despite the complexity and nonlinearity of the original models, the proposed analysis boils down to solving tractable linear mixed-integer stochastic programs. Using experimental electricity market and Internet workload data, various insightful numerical observations are reported. It is shown that the key to link different energy options with different short- and long-term profit characteristics is to conduct risk management at different time horizons. Also, there is a direct relationship between data centers' service-level agreement parameters and their ability to exploit certain energy options. The use of on-site storage and the deployment of geographical workload distribution can particularly help data centers in utilizing high-risk energy choices, such as offering ancillary services or participating in wholesale electricity markets.**

***Keywords*: Data centers, energy portfolio, day-ahead and real-time markets, reserve, renewable generation, energy storage.**

I. INTRODUCTION

As a major energy consumer, a data center has various options to procure electricity. For example, it may purchase electricity from a retailer (RET), e.g., a utility company [1] or a load serving entity [2]. It may also participate in wholesale electricity markets, including the day-ahead market (DAM) and real-time market (RTM) [3], [4]. Another option for data centers is to enroll in ancillary service (ANS) programs [5]–[7]. Data centers may also fully or partially operate by local renewable (REN) power generators such as wind turbines [8] and/or solar panels [9]. Some data centers also use on-site energy storage systems (ESS) [10]. Geographically dispersed data centers could also benefit from geographical workload distribution (GWD), where the Internet and cloud computing workload is routed towards data centers with lower electricity prices or higher renewable generation availability [11], [12].

The above options are summarized in Table I. The last two columns indicate whether the study takes into consideration service-level agreements (SLAs) [13] or risk management (RM) [14]–[16]. Note that, SLA is of importance in this context in order to maintain an acceptable trade-off between energy cost minimization and meeting the quality-of-service obligations for various Internet and cloud computing services.

TABLE I
SUMMARY OF REPRESENTATIVE RELATED LITERATURE

	RET	DAM	RTM	ANS	REN	ESS	GWD	SLA	RM
[3]	✗	✓	✓	✗	✗	✗	✗	✓	✓
[6]	✓	✗	✓	✓	✗	✗	✓	✗	✓
[7]	✓	✗	✗	✓	✗	✗	✗	✗	✗
[10]	✓	✗	✗	✗	✓	✓	✓	✗	✗
[14]	✗	✓	✓	✗	✗	✗	✓	✗	✓
[15]	✗	✓	✓	✗	✗	✗	✓	✗	✓
[16]	✓	✗	✓	✗	✗	✗	✓	✗	✓
[17]	✓	✗	✗	✓	✗	✓	✗	✗	✗
[18]	✗	✗	✗	✗	✗	✓	✗	✗	✗
[19]	✓	✗	✓	✗	✓	✓	✗	✗	✗
[20]	✗	✗	✗	✗	✗	✗	✗	✓	✗
[21]	✗	✗	✗	✗	✗	✗	✗	✓	✗
[22]	✗	✗	✗	✗	✓	✗	✗	✓	✗
[23]	✓	✗	✗	✗	✓	✓	✗	✗	✗
[24]	✓	✗	✗	✗	✓	✗	✓	✗	✗

From Table I, the literature on addressing data centers' energy options is *very fractured*. That is, most existing designs are specific to only a *small subset* of available energy options. Accordingly, it is still unclear how utilizing one energy option may affect selecting other energy options. Addressing these open problems is the focus of this paper, where we develop an *energy portfolio optimization framework* for data centers. The contributions in this paper can be summarized as follows:

1) *Comprehensive Energy Options*: The proposed energy portfolio optimization framework encompass a broad range of energy options for data centers, including all nine items in Table I. The RM and SLA models are particularly detailed in terms of the statistical characteristics of the Internet workload and other stochastic quantities.
2) *Computational Efficiency*: Despite the complexity and nonlinearity of the original models that are used in our comprehensive energy portfolio analysis, the proposed unified energy planning decision making process boils down to solving tractable linear mixed-integer programs.
3) *Insightful Numerical Results*: Using experimental electricity market and Internet workload data, the performance of the proposed energy portfolio optimization approach is evaluated in various case studies. It is observed that different energy options differ in their short-term and long-term profit characteristics. Accordingly, the key to link different energy options is to conduct RM at different time horizons. Also, there is a direct relationship between a data center's SLA parameters and its ability to exploit certain energy options, such as ANS. In this regard, the use of on-site ESS and the deployment of GWD can particularly help data centers in utilizing high-risk energy choices, such as ANS, REN, and RTM.

This paper is comparable also with the literature on energy

M. Ghamkhari and H. Mohsenian-Rad are with the Department of Electrical and Computer Engineering, University of California, Riverside, CA, USA, e-mails: {ghamkhari, hamed}@ece.ucr.edu. A. Wierman is with the Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA, e-mail: adamw@caltech.edu. This work is supported in part by the National Science Foundation via Grants CNS 1319798 and CNS 1319820. The corresponding author for this paper is H. Mohsenian-Rad.

portfolio management in contexts *other than data centers*, e.g., see [25], [26]. Here, the analysis includes energy options that are specific only to data centers, such as geographical workload distribution, which do *not* appear in other load types.

II. ENERGY MANAGEMENT OPTIONS

A. Retailer Market

A data center may procure its electricity power needs from a retail utility company at rates that are often *flat* and based on *long-term* bilateral contracts that are sometimes negotiable between the data center and the utility company. We denote the price and the quantity of power that is purchased at time t from the utility company by $\omega_{\text{RET}}[t]$ and $L_{\text{RET}}[t]$, respectively.

B. Electricity Wholesale Market

In most U.S. markets, power purchase is done in two settlements through *day-ahead* and *real-time* markets. The day-ahead market is settled at about one day before the operation time, while the real-time market is settled either a few minutes before or after operation [27]. We denote the amount of power that is purchased for operation at time t from the day-ahead market and the real-time market by $L_{\text{DAM}}[t]$ and $L_{\text{RTM}}[t]$, respectively. The price in these two markets at time t are denoted by $\omega_{\text{DAM}}[t]$ and $\omega_{\text{RTM}}[t]$, respectively.

By procuring electricity from the wholesale market instead of a local utility company, data centers can avoid the *insurance premiums*, *service charges*, and *mark-up* that utilities may include in retail rates. However, a key challenge in procuring power directly from the wholesale market is price uncertainly, especially in the real-time market. This can expose data centers to the *risk* of facing volatile electricity expenditure [3].

C. Local Renewable Generation

Depending on their locations, data centers can use various on-site renewable generation options, such as wind turbines [8] and/or solar panels [9]. However, renewable generation is a challenging power procurement option due to its intermittency and stochastic nature. We assume that the amount of local renewable generation at the data center at time t is denoted by random variable $G_{\text{REN}}[t]$ with a known probability distribution.

D. Offering Ancillary Services

Traditionally, ancillary services are offered by generators [27, Chapter 9]. However, large consumers, such as data centers, are also eligible to register as *load resources* to offer ancillary services [5], [28]. In this paper, our focus is on a data center that offers *spinning reserve* [29]. Spinning Reserve, also known as *responsive reserve*, is an on-line reserve capacity that is ready to be dispatched within 10 to 15 minutes of receiving a call signal from the power grid operator [30, Section 3].

For a data center that offers reserve service, the amount of power reduction or power injection at time t is $Y_{\text{ANS}}[t]L_{\text{ANS}}[t]$, where L_{ANS} is the reserve bid that is submitted to the day ahead reserve market and $Y_{\text{ANS}}[t]$ is a binary parameter that is 1 if the reserve capacity is actually called; and 0 otherwise. In the case of receiving a call signal, the data center is not allowed to purchase power from the real time market. The spinning reserve service that is offered by data center at time t is compensated by a *capacity* payment based on the total offered capacity $L_{\text{ANS}}[t]$ at rate $\omega_{\text{ANS}}[t]$, and a *call* payment at rate $\omega_{\text{CAL}}[t]$, only if the reserve is actually called [31].

E. Energy Storage

Data centers are often equipped with local energy storage to supply backup power in case of power disruption. Energy storage may also help data centers in lowering their energy expenditure, e.g., by storing energy at low price hours and releasing it at high price hours. We denote the energy storage level at the end of time t by $E_{\text{STR}}[t]$. We must always have

$$0 \le E_{\text{STR}}[t] \le E_{\text{STR}}^{max}, \tag{1}$$

where E_{STR}^{max} is the operational capacity of the storage units. The electricity that is stored at storage units can be injected into the data center to meet local demand, or into the power gird to satisfy the reserve service obligation of the data center once a reserve capacity call signal is received. In our model, unless a reserve capacity signal is received, the data center is not paid for the power that it may inject back to the grid [8].

F. Geographic Workload Distribution

As it is recently shown, e.g., in [8], [11], [12], [32], [33], a group of geographically dispersed data centers can cut their electricity bills by forwarding some of their workload to data centers that face lower regional electricity prices or have more available renewable generation. As we will see in this paper, geographic workload distribution can also help in improving service reliability in data centers, e.g., in case of regional power disruption, unexpected reduction in available renewable generation, or receiving a reserve capacity call signal.

III. ENERGY PORTFOLIO OPTIMIZATION

In this section, we seek to find the *best mix utilization* portfolio of the diverse available energy options that we listed in Section II. We divide the operating time of data center into T successive time slots of lengths τ minutes, e.g. $\tau = 15$. First, we address the case of a single data center. The case with *multiple data centers* is explained in Section III-I.

A. Internet Workload and Service Rate

At each time slot t, suppose the Internet workload arrives at the data center with a general probability distribution with average $\lambda[t]$, variance $\sigma^2[t]$, and auto covariance function $\rho_l[t]$, where $l = 1, 2, \ldots$ is the lag time. Note that, these parameters may change significantly during the day [34]. We assume that each server can handle up to κ service requests per second, where κ is a fixed parameter that depends on the computation capability of the server and the type of service. Let $M[t] \le M^{\max}$ denote the number of servers that are switched on at time slot t. We assume that the service requests that arrive to the data center are queued upon their arrival, until they are pulled out from the queue in a first come-first-served order to be handled by one of the switched on computer servers. The rate at which service requests are pulled out of the queue to be handled by a computer server is

$$\mu[t] = M[t]\kappa. \tag{2}$$

Due to the wear and tear cost associated with switching computer servers on and off, we assume that $\mu[t]$ is changed only at the beginning of each time slot t, not on a moment-by-moment basis. If the duration of time slots τ is around 10 to 15 minutes, then this arrangement also meets the response time requirement in most practical responsive reserve services.

B. Service Level Agreement

To satisfy the quality-of-service (QoS) requirements, the queue waiting time for each service request must be limited according to its SLA [13]. An SLA is identified by three parameters D, δ, and γ. Parameter D indicates the maximum queue waiting time that a service request can tolerate. Parameter δ indicates the service money that the data center receives when it handles a single service request *before* deadline D. Parameter γ indicates the money that the data center must pay to its customers every time it *cannot* handle a service request before deadline D and consequently drops the request.

C. Power Consumption

For a data center, *power usage effectiveness* (PUE), denoted by E_{usage}, is as the ratio of the data center's total power usage to the power usage at servers [35]. Let P_{server} denote the average power usage of a switched on computer server, while it is handling a service request. Assuming full CPU utilization for all switched on servers, the total power consumption of the data center at time slot t is calculated as [8], [36]:

$$\text{Power Consumption} = \phi\,\mu[t], \tag{3}$$

where $\phi = E_{\text{usage}} P_{\text{server}}/\kappa$ and the equality is due to (2).

D. Operational Energy Cost

The operational energy cost of a data center depends on the realizations of various *random parameters*, ranging from the output of its local renewable generators to the cleared market prices and whether or not the data center receives a reserve capacity call signal. At each time slot t, we assume that the statistical characteristics of random variables $\omega_{\text{DAM}}[t]$, $\omega_{\text{RTM}}[t]$, $G_{\text{REN}}[t]$, $\omega_{\text{ANS}}[t]$, $Y_{\text{ANS}}[t]$ and $\omega_{\text{CAL}}[t]$ are modeled by K scenarios. These scenarios can be generated, e.g., from historical data, or from a joint probability distribution, say, using the Monte Carlo method [37]. For each scenario $k = 1, \ldots, K$, we denote the realizations of the random variables as $\omega^k_{\text{DAM}}[t]$, $\omega^k_{\text{RTM}}[t]$, $G^k_{\text{REN}}[t]$, $\omega^k_{\text{ANS}}[t]$, $Y^k_{\text{ANS}}[t]$ and $\omega^k_{\text{CAL}}[t]$. Recall that the retail electricity price $\omega_{\text{RET}}[t]$ is a known and fixed parameter. Also note that, since the real-time market bids are selected at the time of operation, they too depend on the realizations of random scenarios. Accordingly, we denote them as $L^k_{\text{RTM}}[t]$.

Under random scenario k and during time slot t, the total power draw of the data center from the grid is calculated as

$$L_{\text{RET}}[t] + L_{\text{DAM}}[t] + (1 - I_{\text{ANS}}[t] Y^k_{\text{ANS}}[t]) L^k_{\text{RTM}}[t], \tag{4}$$

where

$$I_{\text{ANS}}[t] = \mathbb{I}\left(L_{\text{ANS}}[t] > 0\right). \tag{5}$$

Here, $\mathbb{I}(\cdot)$ is a 0-1 indicator function. If $L_{\text{ANS}}[t] = 0$, then $I_{\text{ANS}}[t] = 0$. If $L_{\text{ANS}}[t] > 0$, then $I_{\text{ANS}}[t] = 1$. To understand the last term in (4), recall from Section II-D that if the data center offers reserve service, i.e., $I_{\text{ANS}}[t] = 1$, and it receives a reserve call signal under scenario k, i.e., $Y^k_{\text{ANS}}[t] = 1$, then the data center must not procure power from the real-time market.

Similarly, the operational energy cost of the data center during time slot t and under random scenario k is obtained as

$$\begin{aligned} &L_{\text{RET}}[t]\omega_{\text{RET}}[t] + L_{\text{DAM}}[t]\omega^k_{\text{DAM}}[t] \\ &\quad + (1 - I_{\text{ANS}}[t] Y^k_{\text{ANS}}[t]) L^k_{\text{RTM}}[t]\omega^k_{\text{RTM}}[t]. \end{aligned} \tag{6}$$

E. Service Rate Allocation

From (3) and (4), at each random scenario k and each time slot t, the following *power balance* equation must hold:

$$\begin{aligned} &L_{\text{RET}}[t] + L_{\text{DAM}}[t] + (1 - I_{\text{ANS}}[t] Y^k_{\text{ANS}}[t]) L^k_{\text{RTM}}[t] \\ &\quad = \phi\,\mu^k[t] - G^k_{\text{REN}}[t] + (E_{\text{STR}}[t] - E_{\text{STR}}[t-1])/\tau, \end{aligned} \tag{7}$$

where $\mu^k[t]$ is the service rate at time slot t under scenario k. Note that, the second and the third terms on the right hand side in (7) incorporate the impact of local renewable generator and energy storage unit, respectively. We can rewrite (7) as

$$\begin{aligned} \mu^k[t] = \frac{1}{\phi}\Big[& L_{\text{RET}}[t] + L_{\text{DAM}}[t] \\ & + (1 - I_{\text{ANS}}[t] Y^k_{\text{ANS}}[t]) L^k_{\text{RTM}}[t] \\ & + G^k_{\text{REN}}[t] - (E_{\text{STR}}[t] - E_{\text{STR}}[t-1])/\tau \Big]. \end{aligned} \tag{8}$$

F. Operational Revenue

The operational revenue of a data center may come from two sources: (a) the revenue due to offering Internet and cloud computing services, and (b) the revenue due to offering reserve service to the power grid. At each time slot t and under random scenario k, these revenue streams are calculated as

$$\tau\lambda[t](\delta - (\delta + \gamma) q(\mu^k[t])) \tag{9}$$

and

$$L_{\text{ANS}}[t]\omega^k_{\text{ANS}}[t] + L_{\text{ANS}}[t] Y^k_{\text{ANS}}[t]\omega^k_{\text{CAL}}, \tag{10}$$

respectively. First, we explain (9). Here, $q(\cdot)$ denotes the probability that an arriving service request is *not* handled before its SLA-required deadline. This probability is a function of service rate $\mu^k[t]$. From the analysis in [38], we have

$$q(\mu) = \alpha(\mu)\ \exp\left(-\frac{1}{2}\min_{n\geq 1}\ m_n(\mu)\right), \tag{11}$$

where

$$\alpha(\mu) = \frac{1}{\lambda\sqrt{2\pi}\sigma} e^{\frac{(\mu-\lambda)^2}{2\sigma^2}} \int_{\mu}^{\infty} (r-\mu) e^{-\frac{(r-\lambda)^2}{2\sigma^2}}\, dr, \tag{12}$$

$$m_n(\mu) = \frac{(D\mu + n(\mu - \lambda))^2}{n\sigma^2 + 2\sum\limits_{l=1}^{n-1} \rho_l[t](n-l)}, \quad \forall n \geq 1. \tag{13}$$

The above model is based on the assumption that service rate is higher than average service request arrival rate. An extension of (11) when this assumption is relaxed is given in [39]. As for the model in (10), the first term is the reserve capacity payment and the second term is the reserve call payment.

G. Risk Management

The *profit* for a data center can be calculated as the data center's revenue minus its cost. In presence of uncertainty, it is natural to seek to maximize the *expected* profit. However, such average-sense profit maximization approach does not take into consideration the distribution of the profit under different realizations of the random parameters in the system. Accordingly, it would still be possible that the data center faces very low profit under certain random scenarios. In this section, we address this shortcoming by restraining the average profit of a data center above a specified threshold, for the random scenarios where the data center's profit takes low values. We note that, the total profit of a data center over T time slots is a stochastic variable with the following sample space:

$$\Psi = \left\{ \sum_{t=1}^{T} \text{Profit}^{k}[t] \;\middle|\; 1 \leq k \leq K \right\} \tag{14}$$

where $\text{Profit}^{k}[t]$ is the profit of data center at time slot t under the kth random scenario. A model for $\text{Profit}^{k}[t]$ will be provided later in Section III-H. Note that, from the discussions in Section III-D, some of the elements in Ψ may be repeated.

In order to restrain the risk of low profit, we seek to keep the expected value of the total profit within the β fractile *lowest profit* random scenarios, above a design threshold Γ:

$$\begin{gathered}\text{Average of } \beta \text{ Fractile Lowest Total Profit Values} \geq \Gamma \\ \Longleftrightarrow \\ \text{Average of } \beta \text{ Fractile Lowest Elements in } \Psi \geq \Gamma,\end{gathered} \tag{15}$$

where $\beta \in [0, 1]$ is a design parameter. A typical value for β is 0.1. A higher Γ indicates a *risk averse* design while a lower Γ indicates a *risk seeking* design [40]–[43]. The choice of parameter Γ depends on the financial obligations that one faces in operating a data center. For example, even though a data center operator's ultimate goal is to maximize annual profit; it may face financial obligations to make monthly, weekly, or daily payments corresponding to facility charges or equipment mortgages. As a result, the operator needs a mechanism to assure a minimum short-term revenue to cover these charges in presence of uncertainty. The amounts of such short-term charges would directly translate to parameter Γ.

To obtain a mathematical expression for the risk management constraint in (15), we first sort the elements in set Ψ in an ascending order to obtain the following set:

$$\overline{\Psi} = \text{Sort } (\Psi), \qquad \overline{\Psi}^{1} \leq \cdots \leq \overline{\Psi}^{K}. \tag{16}$$

From (16), the constraint in (15) is equivalent to

$$\sum_{k=1}^{\beta K} \frac{\overline{\Psi}^{k}}{\beta K} \geq \Gamma. \tag{17}$$

Next, we note that, from [44, Definition 3], we have

$$\begin{aligned}\text{CVaR}_{1-\beta}\left(-\sum_{t=1}^{T}\text{Profit}[t]\right) &= \sum_{k=\beta K}^{1} \frac{-\overline{\Psi}^{k}}{\beta K} \\ &= \sum_{k=1}^{\beta K} \frac{-\overline{\Psi}^{k}}{\beta K} \\ &= \sum_{k=1}^{K} \frac{-\overline{\Psi}^{k}}{\beta K} + \sum_{k=\beta K}^{K} \frac{\overline{\Psi}^{k}}{\beta K} \\ &= -\sum_{k=1}^{\beta K} \frac{\overline{\Psi}^{k}}{\beta K},\end{aligned} \tag{18}$$

where CVaR denotes the standard operator for *conditional value at risk* [44], [45]. From (17) and (18), we can express the risk restrain constraint in (15) as

$$-\text{CVaR}_{1-\beta}\left(-\sum_{t=1}^{T}\text{Profit}[t]\right) \geq \Gamma. \tag{19}$$

Note that, since CVaR is a combinatorial operator that takes the expected value of a sorted set, the minus signs inside and outside the CVaR function in (19) *do not* cancel out each other.

H. Risk-aware Profit Maximization Problem

From the expressions in (6), (9), and (10), the data center's profit at time slot t under scenario k is calculated as

$$\begin{aligned}\text{Profit}^{k}[t] = \;& \tau\lambda[t](\delta - (\delta+\gamma)q(\mu^{k}[t])) \\ & + L_{\text{ANS}}[t]\omega^{k}_{\text{ANS}}[t] \\ & + L_{\text{ANS}}[t]Y^{k}_{\text{ANS}}[t]\omega^{k}_{\text{CAL}} \\ & - L_{\text{RET}}[t]\omega_{\text{RET}}[t] - L_{\text{DAM}}[t]\omega^{k}_{\text{DAM}}[t] \\ & - (1 - I_{\text{ANS}}[t]Y^{k}_{\text{ANS}}[t])L^{k}_{\text{RTM}}[t]\omega^{k}_{\text{RTM}}[t].\end{aligned} \tag{20}$$

Therefore, the risk-aware energy portfolio optimization problem for a data center over T time slots is formulated as

$$\begin{aligned}\textbf{max} \quad & \sum_{t=1}^{T} \frac{1}{K} \sum_{k=1}^{K} \text{Profit}^{k}[t] \\ \textbf{s.t.} \quad & \text{Eqs. } (1), (5), (8), (11), (20), \qquad t = 1, \ldots, T \\ & -\text{CVaR}_{1-\beta}\left(-\sum_{t=1}^{T}\text{Profit}[t]\right) \geq \Gamma,\end{aligned} \tag{21}$$

From [44, Theorem 16], the last constraint in (21) can be reformulated and equivalently expressed as

$$\begin{aligned} & \sum_{t=1}^{T} \text{Profit}^{k}[t] + \zeta + \eta_{k} \geq 0, && k = 1, \ldots, K, \\ & \eta_{k} \geq 0, && k = 1, \ldots, K, \\ & \zeta + \frac{1}{\beta}\frac{1}{K}\sum_{k=1}^{K} \eta_{k} \leq -\Gamma, \end{aligned} \tag{22}$$

where ζ and η_k for all $k = 1, \ldots, K$ are auxiliary variables. By replacing the last constraint in optimization problem (21) with the set of inequalities in (22), the optimization problem (21) can be equivalently expressed as

$$\begin{aligned}
\textbf{max} \quad & \sum_{t=1}^{T} \frac{1}{K} \sum_{k=1}^{K} \text{Profit}^k[t] \\
\textbf{s.t.} \quad & \text{Eqs. } (1), (5), (8), (11), (20), && t = 1, \ldots, T \\
& \sum_{t=1}^{T} \text{Profit}^k[t] + \zeta + \eta_k \geq 0, && k = 1, \ldots, K, \\
& \eta_k \geq 0, && k = 1, \ldots, K, \\
& \zeta + \frac{1}{\beta} \frac{1}{K} \sum_{k=1}^{K} \eta_k \leq -\Gamma.
\end{aligned} \tag{23}$$

By solving optimization problem (23), we maximize the expected value of the profit subject to risk management constraints and several other operational constraints with respect to the diverse energy options that we listed in Section II.

I. Coordinated Geographically Dispersed Data Centers

In this section, we assume that the Internet and cloud computing workload is handled by $N \geq 2$ geographically distributed but coordinated data centers. The Internet workload is first received by a front-end web server and then distributed among data centers. For the case with multiple data centers, notations L_{RET}, L_{DAM}, L_{RTM}, G_{REN}, I_{ANS}, L_{ANS}, Y_{ANS} and E_{STR} are replaced with $L_{i,\text{RET}}$, $L_{i,\text{DAM}}$, $L_{i,\text{RTM}}$, $G_{i,\text{REN}}$, $I_{i,\text{ANS}}$, $L_{i,\text{ANS}}$, $Y_{i,\text{ANS}}$ and $E_{i,\text{STR}}$ corresponding to data center i. Precisely, we denote the electricity price at retail market, day-ahead market and real-time market at the location of ith data center within time slot t by $\omega_{i,\text{RET}}[t]$, $\omega_{i,\text{DAM}}[t]$ and $\omega_{i,\text{RTM}}[t]$ respectively. Also, the amount of available renewable generation at the location of ith data center within time slot t is denoted by $G_{i,\text{REN}}[t]$. Moreover, $Y_{i,\text{ANS}}[t] \in \{0, 1\}$ indicates whether a reserve capacity call signal is received at ith data center within time slot t. For the time slot t, $Y_{i,\text{ANS}}[t] = 1$ means a reserve capacity call signal is received by the ith data center, while $Y_{i,\text{ANS}}[t] = 0$ means no capacity call signal is received by the ith data center. The reserve capacity price and reserve call price at the location of data center i and within the time slot t are denoted by $\omega_{i,\text{ANS}}$ and $\omega_{i,\text{CAL}}$ respectively.

Let $I_{i,\text{ANS}}[t] \in \{0, 1\}$ denote whether data center i participates in the reserve market at time slot t. Specifically, for each time slot t, $I_{i,\text{ANS}}[t] = 1$ means that data center i does participate in the reserve market, while $I_{i,\text{ANS}}[t] = 0$ means that data center i does not participate in the reserve market. Similar to the discussion in Section III-D we have

$$I_{i,\text{ANS}}[t] = \mathbb{I}\left(L_{i,\text{ANS}}[t] > 0\right) \tag{24}$$

where, $\mathbb{I}(\cdot)$ is defined in Section III-D.

At each time slot t, we assume that the statistical characteristics of random variables $\omega_{i,\text{DAM}}[t]$, $\omega_{i,\text{RTM}}[t]$, $G_{i,\text{REN}}[t]$, $\omega_{i,\text{ANS}}[t]$, $Y_{i,\text{ANS}}[t]$ and $\omega_{i,\text{CAL}}[t]$ are modeled by K random scenarios. For each random scenario $k = 1, \ldots, K$, we denote the realizations of the random variables as $\omega^k_{i,\text{DAM}}[t]$, $\omega^k_{i,\text{RTM}}[t]$, $G^k_{i,\text{REN}}[t]$, $\omega^k_{i,\text{ANS}}[t]$, $Y^k_{i,\text{ANS}}[t]$ and $\omega^k_{i,\text{CAL}}[t]$. Also, let $L_{i,\text{RET}}[t]$, $L_{i,\text{DAM}}[t]$ and $L_{i,\text{ANS}}[t]$ denote the bid of data center i within time slot t at retail electricity market, day ahead electricity market and reserve market, respectively. Let $L^k_{i,\text{RTM}}[t]$ denote the ith data center bid at real time market at time slot t and under scenario k. Finally, $E_{i,\text{STR}}$ is the charging/discharging schedule of data center i within time slot t.

Let $\lambda^k_i[t]$ denote the average of Internet workload that is forwarded toward data center i from the front-end web server within time slot t and under the realization of kth scenario. Under each random scenario k, the total outgoing traffic at the front-end server must match the total arriving workload:

$$\sum_{i=1}^{N} \lambda^k_i[t] = \lambda[t] \qquad k = 1, \ldots, K. \tag{25}$$

Moreover, we assume that the service requests that are forwarded to the ith data center from the front-end web server are selected randomly from all arriving service requests to the front-end web server. Therefore, based on basic Statistics [46, Theorem 6.14], the variance and autocovarince of the Internet workload that is received by ith data center within time slot t and under kth scenario are obtained as

$$\sigma^k_i[t]^2 = \left(\frac{\lambda^k_i[t]}{\lambda[t]}\right)^2 \sigma^2[t], \qquad \rho^k_{i,l}[t] = \left(\frac{\lambda^k_i[t]}{\lambda[t]}\right)^2 \rho_l[t]. \tag{26}$$

Next, let $\mu^k_i[t]$ denote the service rate of data center i at time slot t and under random scenario k. Similar to the discussion in Section III-E, we have

$$\begin{aligned}
\mu^k_i[t] = \frac{1}{\phi} \Big[& L_{i,\text{RET}}[t] + L_{i,\text{DAM}}[t] \\
& + (1 - I_{i,\text{ANS}}[t] Y^k_{i,\text{ANS}}[t]) L^k_{i,\text{RTM}}[t] \\
& + G^k_{i,\text{REN}}[t] - (E_{i,\text{STR}}[t] - E_{i,\text{STR}}[t-1])/\tau \Big].
\end{aligned} \tag{27}$$

Suppose the communication cost to transmit the workload from the front-end web server to data center i is $\xi_i \lambda_i[t]$. Similar to the discussion in Sections III-F and III-H, the total profit of the data centers under scenario k is obtained as

$$\sum_{i=1}^{N} \sum_{t=1}^{T} \text{Profit}^k_i[t], \tag{28}$$

where

$$\begin{aligned}
\text{Profit}^k_i[t] = \; & \tau \lambda^k_i[t] (\delta - (\delta + \gamma) q_i(\mu^k_i[t]), \lambda^k_i[t]) - \xi_i \lambda^k_i[t] \\
& + L_{i,\text{ANS}}[t] \omega^k_{i,\text{ANS}}[t] \\
& + L_{i,\text{ANS}}[t] Y^k_{i,\text{ANS}}[t] \omega^k_{i,\text{CAL}} \\
& - L_{i,\text{RET}}[t] \omega_{i,\text{RET}}[t] - L_{i,\text{DAM}}[t] \omega^k_{i,\text{DAM}}[t] \\
& - (1 - I_{i,\text{ANS}}[t] Y^k_{i,\text{ANS}}[t]) L^k_{i,\text{RTM}}[t] \omega^k_{i,\text{RTM}}[t].
\end{aligned} \tag{29}$$

and

$$q_i(\mu^k_i[t], \lambda^k_i[t]) = \alpha(\mu^k_i[t], \lambda^k_i[t]) \exp\left(-\frac{1}{2} \min_{n \geq 1} \; m_n(\mu^k_i[t])\right). \tag{30}$$

As in (12) and (13), we have

$$\alpha(\mu_i^k[t], \lambda_i^k[t]) = \left(exp\left(\frac{(\mu_i^k[t] - \lambda_i^k[t])^2}{2\sigma_i^k[t]^2} \right) \Big/ \lambda_i^k[t]\sqrt{2\pi}\sigma_i^k[t] \right) \int\limits_{\mu_i^k[t]}^{\infty} (r - \mu_i^k[t]) e^{-\frac{(r-\lambda_i^k[t])^2}{2\sigma_i^k[t]^2}} dr, \tag{31}$$

and

$$m_n(\mu_i^k[t]) = \frac{(D\mu_i^k[t] + n(\mu_i^k[t] - \lambda_i^k[t]))^2}{n\sigma_i^k[t]^2 + 2\sum_{l=1}^{n-1} \rho_{i,l}^k[t](n-l)}, \quad \forall n \geq 1. \tag{32}$$

Similar to the discussion in Section III-G, the average total profit over β fractile lowest profit random scenarios is kept above a threshould Γ, if the following inequality holds:

$$-\text{CVaR}_{1-\beta}\left(-\sum_{i=1}^{N}\sum_{t=1}^{T} \text{Profit}_i^k[t] \right) \geq \Gamma. \tag{33}$$

We seek to maximize the *aggregated* expected profit of *all* data centers. Different from the single data center case in Section III-H, here, $\lambda_i^k[t]$ for $i = 1, \dots, N$ is an optimization variable. The following risk-aware energy portfolio optimization problem gives the optimum operation variables of data centers:

$$\begin{aligned} \textbf{max} \quad & \sum_{t=1}^{T} \frac{1}{K} \sum_{k=1}^{K}\sum_{i=1}^{N} \text{Profit}_i^k[t] \\ \textbf{s.t.} \quad & \text{Eqs. } (24), (25), (27), (29), (30) \qquad \forall t, i, k, \\ & -\text{CVaR}_{1-\beta}\left(-\sum_{i=1}^{N}\sum_{t=1}^{T} \text{Profit}_i^k[t] \right) \geq \Gamma. \end{aligned} \tag{34}$$

Similar to the discussion in Section III-H, from [44], [47], the optimization problem (34) can be equivalently expressed as

$$\begin{aligned} \textbf{max} \quad & \sum_{t=1}^{T} \frac{1}{K} \sum_{k=1}^{K}\sum_{i=1}^{N} \text{Profit}_i^k[t] \\ \textbf{s.t.} \quad & \text{Eqs. } (24), (25), (27), (29), (30) \qquad \forall t, i, k, \\ & \sum_{t=1}^{T}\sum_{i=1}^{N} \text{Profit}_i^k[t] + \zeta + \eta_k \geq 0, \qquad \forall k, \\ & \eta_k \geq 0, \qquad \forall k, \\ & \zeta + \frac{1}{\beta}\frac{1}{K}\sum_{k=1}^{K} \eta_k \leq -\Gamma. \end{aligned} \tag{35}$$

If $N = 1$, then problem (35) reduces to problem (23).

IV. SOLUTION METHOD

Problem (35) is a mixed-integer *nonlinear* program, which is hard to solve. Notice that, even if we relax the binary constraints in (24), problem (35) is still hard to solve due to the non-convex bilinear terms $L_{i,\text{RTM}}^k[t] I_{i,\text{ANS}}^k[t], \forall i, \forall k$ in (27) and (29). In this Section, we first propose a solution approach based on combining convex programming with the branch-and-bound method [48]. This approach is *guaranteed* to give the optimal solution of the problem in (35). After that, we will also propose an approximate solution for the problem in (35) which is based on mixed integer linear programming (MILP) and can be solved efficiently, e.g., using CPLEX [49].

We start by pointing out that we can replace the expression

$$(1 - I_{i,\text{ANS}}[t] Y_{i,\text{ANS}}^k[t]) L_{i,\text{RTM}}^k[t]$$

with $L_{i,\text{RTM}}^k[t]$ in (27) and (29) by introducing the following inequality as a new constraint to the problem in (35):

$$0 \leq L_{i,\text{RTM}}^k[t] \leq (1 - I_{i,\text{ANS}}[t] Y_{i,\text{ANS}}^k[t])(\kappa M^{\max}\phi), \tag{36}$$

where $\kappa M^{\max}\phi$ is the maximum value that $L_{i,\text{RTM}}^k[t]$ can take. To see this, we note that from (36), if $I_{i,\text{ANS}}[t] Y_{i,\text{ANS}} = 1$, then $L_{i,\text{RTM}}^k[t]$ is forced to zero. Hence, the value of $L_{i,\text{RTM}}^k[t]$ is the same as that of $(1 - I_{i,\text{ANS}}[t] Y_{i,\text{ANS}}^k[t]) L_{i,\text{RTM}}^k[t]$, as long as $I_{i,\text{ANS}}[t] Y_{i,\text{ANS}} = 1$. Furthermore, if $I_{i,\text{ANS}}[t] Y_{i,\text{ANS}} = 0$, then $1 - I_{i,\text{ANS}}[t] Y_{i,\text{ANS}}^k[t] = 1$ and the value of $L_{i,\text{RTM}}^k[t]$ is again the same as that of $(1 - I_{i,\text{ANS}}[t] Y_{i,\text{ANS}}^k[t]) L_{i,\text{RTM}}^k[t]$. After replacing $(1 - I_{i,\text{ANS}}[t] Y_{i,\text{ANS}}^k[t]) L_{i,\text{RTM}}^k[t]$ in (27) and (29) with $L_{i,\text{RTM}}^k[t]$, and adding (36) as a new constraint to (35), the following optimization problem is obtained which is equivalent to (35):

$$\begin{aligned} \textbf{max} \quad & \sum_{t=1}^{T} \frac{1}{K} \sum_{k=1}^{K}\sum_{i=1}^{N} \text{Profit}_i^k[t] \\ \textbf{s.t.} \quad & \text{Eqs. } (25), (30) \qquad \forall t, i, k, \\ & \sum_{t=1}^{T}\sum_{i=1}^{N} \text{Profit}_i^k[t] + \zeta + \eta_k \geq 0, \qquad \forall k, \\ & \eta_k \geq 0, \qquad \forall k, \\ & \zeta + \frac{1}{\beta}\frac{1}{K}\sum_{k=1}^{K} \eta_k \leq -\Gamma \\ & 0 \leq L_{i,\text{RTM}}^k[t] \leq (1 - I_{i,\text{ANS}}[t] Y_{i,\text{ANS}}^k[t]) \kappa M^{\max}\phi, \end{aligned} \tag{37}$$

where

$$\begin{aligned} \text{Profit}_i^k[t] = & \tau\lambda_i^k[t](\delta - (\delta+\gamma) q_i(\mu_i^k[t]), \lambda_i^k[t]) - \xi_i \lambda_i^k[t] \\ & + L_{i,\text{ANS}}[t]\omega_{i,\text{ANS}}^k[t] + L_{i,\text{ANS}}[t] Y_{i,\text{ANS}}^k[t] \omega_{i,\text{CAL}}^k \\ & - L_{i,\text{RET}}[t]\omega_{i,\text{RET}}[t] - L_{i,\text{DAM}}[t]\omega_{i,\text{DAM}}^k[t] \\ & - L_{i,\text{RTM}}^k[t]\omega_{i,\text{RTM}}^k[t], \end{aligned} \tag{38}$$

and

$$\begin{aligned} \mu_i^k[t] = \frac{1}{\phi}\Big[& L_{i,\text{RET}}[t] + L_{i,\text{DAM}}[t] + L_{i,\text{RTM}}^k[t] \\ & + G_{i,\text{REN}}^k[t] \\ & - (E_{i,\text{STR}}[t] - E_{i,\text{STR}}[t-1])/\tau \Big]. \end{aligned} \tag{39}$$

Note that, from [39, Theorem 2], $q_i(\mu_i^k[t], \lambda_i^k[t])$ in (30) is a convex function of $\mu_i^k[t]$. Also, from (26), $q_i(\mu_i^k[t], \lambda_i^k[t])$ in (30) is a function of $\mu_i^k[t]/\lambda_i^k[t]$, i.e., it depends on only the ratio of $\mu_i^k[t]$ and $\lambda_i^k[t]$. Therefore, from [50, Proposition 4], $\lambda_i^k[t] q_i(\mu_i^k[t], \lambda_i^k[t])$ is jointly convex over $\mu_i^k[t]$ and $\lambda_i^k[t]$. As a result, the profit model in (38) is convex and therefore the optimization problem (37) is a mixed-integer convex program.

It can be solved with *guaranteed optimality* using convex programming and branch-and-bound method [48].

In practice, a complicated convex function such as $\lambda_i^k[t]q_i(\mu_i^k[t], \lambda_i^k[t])$ is often approximated by *piecewise linear* or *piecewise quadratic* functions to facilitate applying numerical convex programming algorithms, c.f. [51, Section 10.4], [52, Section 13.5], and [53]–[58]. Similarly, in this paper, we replace $\lambda_i^k[t]q_i(\mu_i^k[t], \lambda_i^k[t])$ in (38) with its *two-dimensional piece-wise outer-linearized approximation* [59]:

$$z_i^k[t] = \max_p \{A_{p,i}[t]\mu_i^k[t] + B_{p,i}[t]\lambda_i^k[t] + C_{p,i}[t]\}. \tag{40}$$

where $A_{p,i}[t]$, $B_{p,i}[t]$ and $C_{p,i}[t]$ are the parameters of the tangent plane to $\lambda_i^k[t]q_i(\mu_i^k[t], \lambda_i^k[t])$ at $(\mu_p^\star[t], \lambda_p^\star[t])$. Here, linearization is done at P different points $(\mu_p^\star[t], \lambda_p^\star[t])$, where $p = 1, \ldots, P$. Note that, any desirable accuracy can be reached if P is large enough. In fact, from [60, Proposition 6.4.1], we have:

$$\lambda_i^k[t]q_i(\mu_i^k[t], \lambda_i^k[t]) = \lim_{P\to\infty} z_i^k[t]. \tag{41}$$

From (38), minimizing the objective function in (37) involves minimizing $z_i^k[t]$; accordingly, we can replace (40) with

$$z_i^k[t] \geq \{A_{p,i}[t]\mu_i^k[t] + B_{p,i}[t]\lambda_i^k[t] + C_{p,i}[t]\} \quad \forall p. \tag{42}$$

After substituting the term $\lambda_i^k[t]q_i(\mu_i^k[t], \lambda_i^k[t])$ in (38) with $z_i^k[t]$ and adding the constraint in (42) to the problem (37), the problem (37) becomes a mixed integer linear program and can be solved with existing software such as CPLEX and MOSEK.

V. CASE STUDIES

A. *Simulation Setting*

Unless stated otherwise, we consider a data center with $M^{\max} = 50{,}000$ servers, $P_{\text{server}} = 150$ watts, $E_{\text{usage}} = 1.2$, and $\kappa = 0.1$. The SLA parameters are set as in [38], where $\delta = 7 \times 10^{-5}$, $\gamma = 3.5 \times 10^{-5}$ and $D = 0.3$. The service rate is updated every $T = 15$ minutes. The default risk parameters are $\beta = 0.1$ and $\Gamma = 80$. The day-ahead and real-time market prices are from PJM at [61]. The data for the reserve capacity call signal is from PJM, based on its historical synchronized reserve events [62]. The data for reserve capacity price is from PJM [63]. We set $L_{\text{CAL}} = L_{\text{RTM}}$ [64]. The PJM datasets are from January 1, 2004 to January 30, 2004. The data for wind speed is from [65], and the wind turbine power-versus-wind-speed curve is from [66]. The statistical data of the workload is from the web hits of Wikipedia on 9/19/2007 [67]. For the case studies that involve only one time slot, the data is from 3:30 PM to 3:45 PM, which is one of the ten time intervals at which PJM sent out a reserve capacity signal during the studied period. For simulations that include one data center, we use the loss probability model in [39], which is an extension of the model in (11) to the entire range of service rate.

B. *Impact of Risk Management Constraint*

The optimum bids and the resulted optimal expected profit over one time slot versus parameter Γ are shown in Fig. 1. For a data center that bids in the reserve market, the lowest profit values occur in scenarios where a reserve capacity call signal is received. In such scenarios, although the data center gains a payment of $L_{\text{ANS}}\omega_{\text{RTM}}$ that is not gained in other scenarios without a reserved capacity call signal, such payment is still much lower than the SLA revenue that the data center loses due to dropping its service requests to lower its power consumption. As the risk parameter Γ increases the data center becomes more risk averse and lowers its reserve capacity bids.

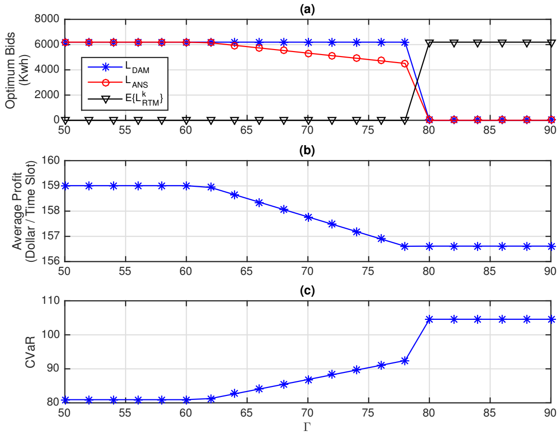

Fig. 1. The impact of risk management parameter Γ: (a) optimal day-ahead energy and reserve market bids, (b) expected profit, (c) CVaR of profit.

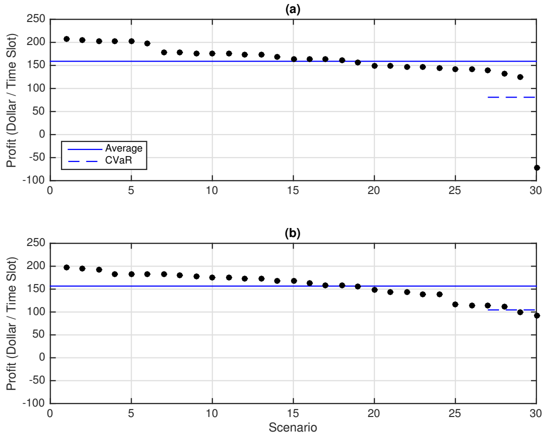

Fig. 2. The profit values over 30 scenarios for a design that is: (a) risk seeking, (b) risk averse. The profits are sorted in a descending order.

Next, we compare a risk averse design with $\Gamma = 50$ and a risk seeking design with $\Gamma = 80$. The results are shown in Fig. 2, where the profit values over the considered 30 random scenarios are sorted in a descending order. The average per-time slot profit is 156.61 and 159.00 for the risk averse and risk seeking designs, respectively. However, the average profit across the 10% lowest profit scenarios is 104.57 and 80.93 for the risk-averse and risk seeking designs, respectively. There is one scenario with *negative* profit under a risk seeking design, while the profit is always positive under a risk averse design.

C. *Impact of Renewable Generation*

Suppose some wind turbines are installed at a data center. Each turbine has a rated power output of 50 kW. The expected profit and the optimal bids versus the number of wind turbines are shown in Fig. 3. The profit increases as we increase the

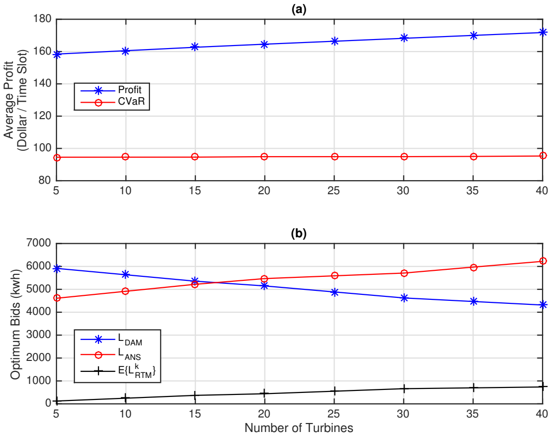

Fig. 3. Operation with renewable generators based on the number of wind turbines: (a) average profit, (b) optimal day-ahead energy and reserve bids.

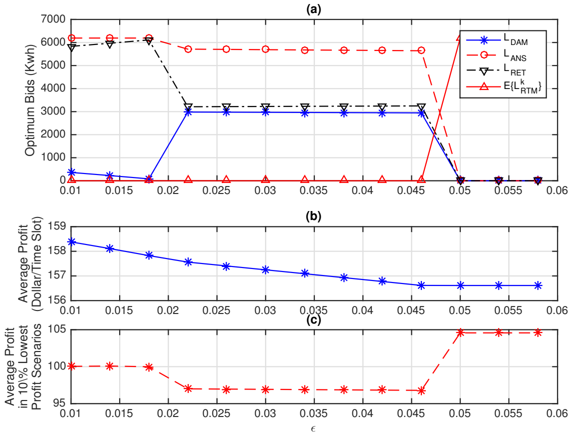

Fig. 4. The impact of power purchase from Retail Market (a) optimum bids, (b) average profit and (c) Average of profit in 10% lowest profit scenarios

number of wind turbines. Also, as the amount of turbines increases, the total electricity purchase, i.e., the summation of day-ahead market bid and real-time market bid, reduces in order to lower the electricity cost of the data center. Furthermore, increasing the number of wind turbines allows the data center to increase its real-time market and reserve bids, because it can now rely on its local generation during the time slots where it receives a reserve capacity call signal.

D. Impact of Power Purchase from Retail Market

Suppose the data center can purchase electricity also from a retailer at fixed price $\omega_{\text{RET}} = (1+\epsilon)E\{\omega_{\text{DAM}}\}$, where $\epsilon > 0$. Fig. 4 shows the optimum bids and the average profit versus parameter ϵ. When ϵ is low, the data center procures a portion of its energy needs from the retailer while it also bids to the reserve market. This is because, by obtaining electricity from the retailer at a flat rate, the data center is not exposed to high prices. Consequently, the average profit in the 10% lowest profit scenarios is kept above Γ even when the data center bids in the reserve market. As ϵ increases, more electricity is procured from the wholesale market than the retailer in order to decrease the cost. The reserve bid is lowered so as to increase the average profit in the 10% lowest profit scenarios.

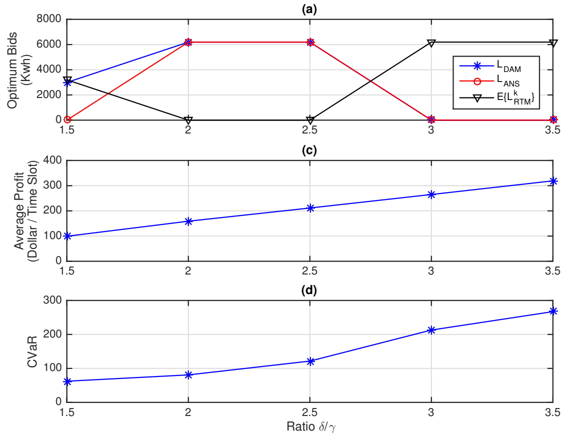

Fig. 5. The impact of changing SLA parameters on the data center operation: (a) optimal bids, (b) average L^k_{RTM} and (c) Average of profit and (d) CVaR are shown for different ratios of δ/γ, where γ is fixed and δ is changing.

E. Impact of SLA Parameters

The optimum bids and the resulted expected profit for different ratios of SLA parameters δ/γ are shown in Fig. 5, where $\Gamma = 60$, γ is fixed and δ changes. From the results in Fig. 5(a), as the ratio δ/γ increases from 1.5 to 2, the day-ahead and reserve market bids take increasing trends. This is because, with higher values of δ, the data center's SLA revenue in the 10% lowest profit scenarios can be kept above Γ even if fewer service requests are handled in scenarios where the data center receives a reserve capacity call signal. Therefore, without violating the risk management constraint, the reserve market bid is increased such that the revenue from the reserve service and consequently the average of profit increases. As δ/γ increases from 2 to 2.5, the optimum power purchase from the day-ahead market remains fixed and equal to an amount that is enough to handle all the service requests in the scenarios without a reserve capacity call signal. However, there is still one scenario in which all of the service requests are dropped. Finally, when the ratio δ/γ changes from 2.5 to 3, the SLA revenue is quite high, making it optimum not to bid in the reserve market. Also, from the results in Fig. 5(b), when $\delta/\gamma \geq 3$, the optimum reserve bid is zero and power is purchased from the real-time market, not the day-ahead market. From the results in Fig. 5(c), the average profit increases as the SLA parameter δ increases. Finally, from the results in Fig. 5(d), the CVaR is always kept above Γ.

F. Energy Portfolio Management Over Multiple Time Slots

In this section, we conduct energy portfolio management over $T = 8$ successive time slots, from 3:00 PM to 5:00 PM. The results are shown in Fig. 6. In single-time-slot energy portfolio management, problem (23) is solved T times for T time slots, where $\Gamma = \Gamma_{ST} = 80$. In contrast, under multiple-time-slots energy portfolio management, problem (23) is solved only *once* but across all time slots, where $\Gamma = \Gamma_{MT} = T\Gamma_{ST} = 8 \times 80 = 640$. As one would expect, the per time slot CVaR is always above Γ_{ST} in the single-time-slot design. However, the per time slot CVaR is below Γ_{ST} for two time slots in the multiple-time-slots design because such

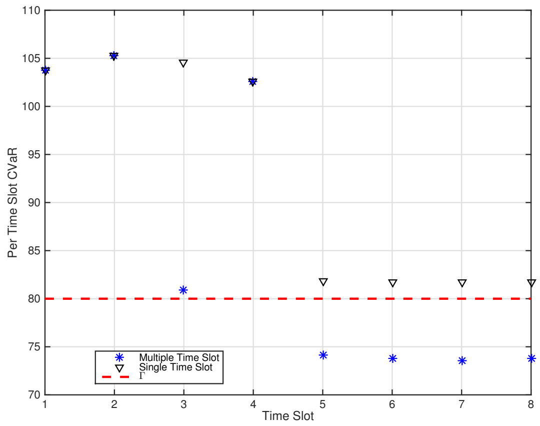

Fig. 6. Per-time-slot CVaR for profit based on two designs: risk management on a single time slot; and joint risk management across multiple time slots.

design only keeps the CVaR of the *total* profit above Γ_{MT} and does *not* impose any constraint on the per-time-slot CVaR. The CVaR of the total profit is 726.64 for the multiple-time-slots design which is above Γ_{MT}. The CVaR of the total profit for the single-time-slot design is 767.76. The average total profit across all scenarios is $11,602$ and $11,670$ for the single-time-slot and multiple-time-slots designs, respectively. Hence, single-time-slot energy portfolio management is more risk averse than multiple-time-slots energy portfolio management.

G. Impact of Local Electricity Storage

Suppose the data center is equipped with an energy storage system with capacity 100 KWh and also ten wind turbines of the type in [66]. Energy portfolio management is done in multiple-time-slots fashion over $T = 96$ time slots, i.e., an entire day. We set $\Gamma_{MT} = 96 \times 80 = 7680$. The storage unit can take one of the following states at each time slot: *charge*, *discharge*, and *idle*. Whenever a reserve capacity call signal is received, either consumption is reduced by L_{ANS}, or the storage unit is discharged at L_{ANS}. At those time slots where the storage unit is discharged, its electricity output *cannot* be injected into the grid, unless a signal for reserve capacity call is received. Here, the data center is allowed to submit reserve bids at time slots 28, 29, 30, 31, 38, 50, 63, 72, 73, 75 [62].

Fig. 7 shows the Internet workload, the average renewable power generation, and the average of the real-time and day-ahead market prices during the one day operation horizon. Fig. 8 shows the optimal bids and the optimal charge (positive) and discharge (negative) schedule for the energy storage unit. At optimality, the data center submits non-zero reserve bids at time slots 28, 38, 50, 63 72, 73, 75. From Fig. 7(a), the optimum real-time market bid L^k_{RTM} is always zero during scenarios where there is a received reserve capacity call signal, which is consistent with the reserve market rules, see Section II-D. Specially, at time slots 63, 72, 73 and 75, electricity is purchased only from the day-ahead market, even though the average RTM price is less than the average DAM price.

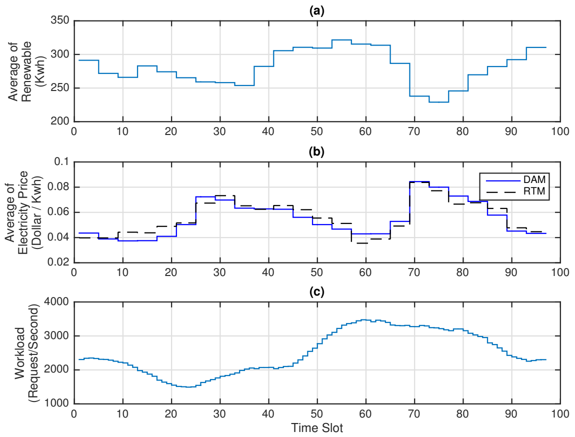

Fig. 7. The system parameters for the case study in Sections V-G and V-L: (a) the average local renewable power generation, (b) the average day-ahead and real-time market prices, (c) Internet workload.

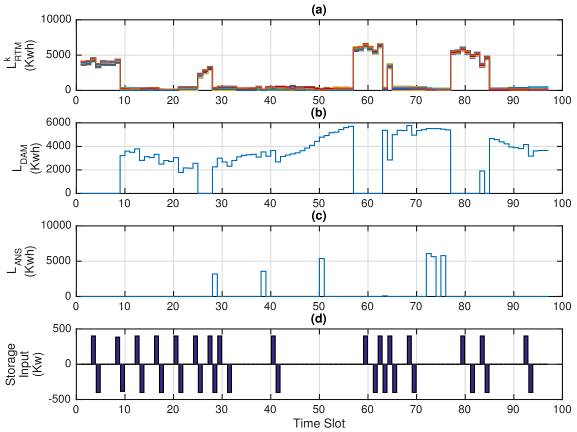

Fig. 8. The optimal operation results for the case study in Section V-G: (a) optimum bids to real-time market, (b) the optimal bids to the day-ahead market, (c) the optimum bid to reserve market, (d) the optimal charge and discharge schedule of the energy storage unit.

H. Geographical Workload Distribution

Consider two geographically distributed data centers. The first one is equipped with 50 wind turbines of the type in [66]. It can also purchase electricity from a retailer at price $(1+.01)E\{\omega_{\text{DAM}}\}$, where the data for ω_{DAM} is from [61]. The second data center can purchase electricity directly from the day-ahead and real-time wholesale markets [61]. It can also bid in the reserve market. Here, we have $\Gamma = 75$ and $\xi_i = 0$, for all $i = 1, \ldots, N$. Fig. 9 shows the results for one time slot from 5:45 PM to 6:00 PM, in which there is a received capacity call at the 18th scenario. Fig. 9(c) shows the optimum bids and Fig. 9(d) shows the optimum fraction of workload that is sent to the first data center at the case of each scenario. At the 18th scenario, the optimum real-time market bid is zero, which follows the reserve market participation rules. Thus, a high fraction of the workload is sent to the first data center at the case of the 18th scenario.

I. Impact of Communication Cost

Fig. 10 shows the optimum total profit of data centers over a time slot of length $T = 15$ minutes, and the optimum

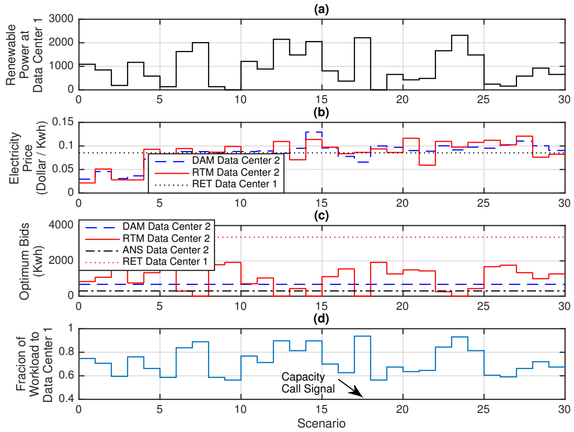

Fig. 9. Two coordinated data centers: (a) renewable generation at data center 1; (b) electricity prices; (c) optimum bids; (d) optimum workload distribution.

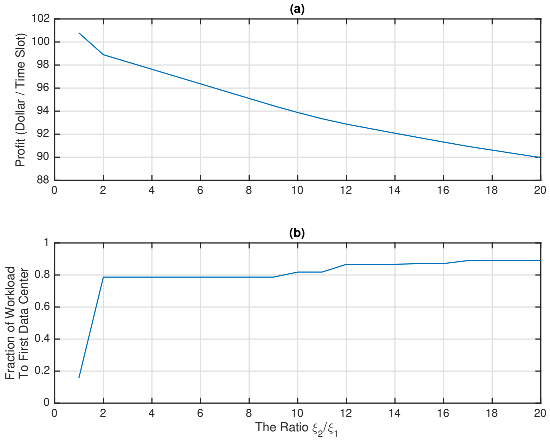

Fig. 10. The impact of communication cost on geographical workload distribution with two coordinated data centers: (a) the optimum profit; (b) optimum fraction of workload that is forwarded to the first data center

fraction of workload that is forwarded to the first data center, as a function of ξ_2/ξ_1, where ξ_1 is assumed to be fixed. In obtaining this figure, we assumed that the first data center submits bids to the day-ahead and real-time electricity markets. As for the second data center, we assumed that it is equipped with wind turbines and also procures electricity from a retail market. Here, we set $\Gamma = 10$. From Fig. 10, as the ratio ξ_2/ξ_1 increases, the total profit of data centers decreases and a higher fraction of workload is forwarded to the first data center.

J. Comparison to other Profit Maximization Models

In this Section, we compare the performance of the proposed profit maximization models with the ones in [11] and [12] for the case of a data center that purchases electricity from both day-ahead and real-time markets. Fig. 11 shows that the proposed model in (23) gives a higher profit for all time slots during a 24 hours time interval. Specially, in time slots 25 to 28, the model in (23) significantly outperforms the models in [11] and [12], as the model in (23) considers procuring electricity at real-time electricity market which has lower electricity prices than the day-ahead market in time slots 25 to 28, while the models in [11] and [12] are solely based on procuring electricity from one single electricity market, i.e., day-ahead market. Also, in time slots 29 to 32, as the electricity price is cheaper in day-ahead market than in real-time market, all the approaches in (23), [11], and [12] procure electricity from the day-ahead market. However, our approach in (23) still outperforms the models in [11] and [12] in time slots 29 to 32, as the model in (23) takes into account the SLA, while the other two models do not.

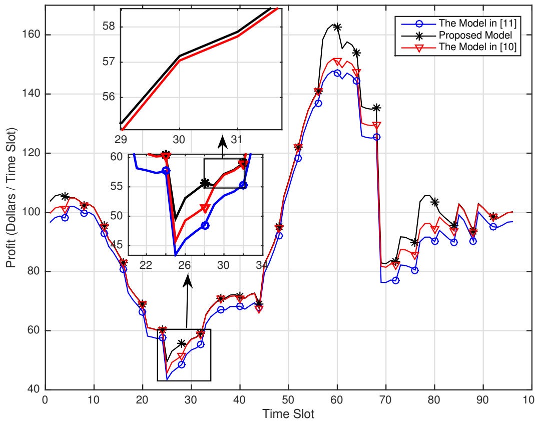

Fig. 11. The profit of data center over one single time slot for our proposed profit maximization design as well as for the designs in [12] and [11].

K. Computational Time and Optimality of Proposed Solution

Again consider the simulation setup in Section V-G. In this section, we examine the impact of changing the number of linearization segments P on the computation time and the optimization accuracy. Fig. 12(a) shows that the profit that is obtained from (23) increases as the number of segments increases, but it is saturated when the number of segments reaches 25. Also, Fig. 12(b) shows that the computation time in solving problem (23) has an increasing trend when the number of segments increases. In overall, from Figs. (12)(a) and (b), one can achieve reasonable optimality and computational time by using the optimization formulation in (23). We note that, the results in this section are obtained from a personal computer with 16 Gb of RAM and an Intel Core i5 CPU @ 2.6 GHz.

L. Flexibility in Decision Making Timing Horizon

In practice, the time interval for switching computer servers on or off could be longer than what we assumed in our case studies so far. However, changing the length of time intervals is easy. For example, suppose the number of switched on computer servers is changed once a day. The model in (23) gives the optimum operating variables for this setup, if we add the following constraint to the problem in (23):

$$\mu^k[1] = \cdots = \mu^k[96], \qquad \forall k \leq K. \tag{43}$$

Notice that a whole day constitutes of 96 time slots of length $T = 15$ minutes, and therefore the constraint in (43) indicates that the service rate is fixed over one whole day and under the kth scenario. Fig. 13 shows the optimal operating variables, for the simulation setup in Fig. 7, when the constraint (43) is added to (23). Here, the optimum reserve market bid is

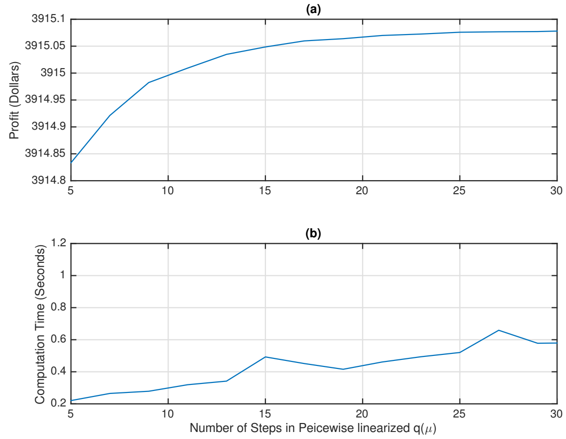

Fig. 12. The impact of number of line segments on the results on the case study in Section V-G: (a) the optimum profit; and (b) computational time.

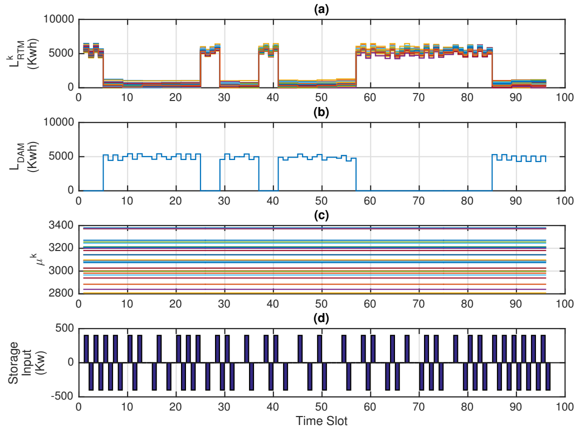

Fig. 13. The optimal operation results for the case study in Section V-L: (a) the optimal bids to real-time market, (b) the optimal bids to the day-ahead market, (c) The optimum service rates, (d) the optimal charge and discharge schedule of the energy storage unit.

obtained as zero over all time slots. Also, from Fig. 13(c), the optimum service rates at the realization of each random scenario is the same over all time slots.

VI. CONCLUSIONS

A comprehensive and unified energy portfolio optimization framework was presented in form of solving tractable linear mixed-integer programs for both single and coordinated multiple data centers. It takes into account a broad range of energy options and design factors. Using practical electricity market and practical Internet workload data, various case studies were presented to gain insights about the performance of the proposed energy portfolio optimization under different operating conditions, and also to gain insights on how utilizing one energy option may affect selecting other energy options.

REFERENCES

[1] W. Cheng, B. Urgaonkar, G. Kesidis, U. V. Shanbhag, and Q. Wang, “A case for virtualizing the electric utility in cloud data centers,” in *Proc. of the USENIX HotCloud*, Philadelphia, PA, June 2014.

[2] Z. Liu, I. Liu, S. Low, and A. Wierman, “Pricing data center demand response,” in *Proc. of ACM Sigmertics*, Austin, TX, June 2014.

[3] M. Ghamkhari, H. Mohsenian-Rad, and A. Wierman, “Optimal risk-aware power procurement for data centers in day-ahead and real-time electricity markets,” in *Proc. of IEEE INFOCOM Smart Data Pricing (SDP) Workshop*, Toronto, Canada, May 2014.

[4] A. Wierman, Z. Liu, I. Liu, and H. Mohsenian-Rad, “Opportunities and challenges for data center demand response,” in *Proc. of the IEEE International Green Computing Conference*, Dallas, TX, November 2014.

[5] M. Ghamkhari and H. Mohsenian-Rad, “Data centers to offer ancillary services,” in *Proc. of IEEE Smart Grid Comm*, November 2012.

[6] R. Wang, N. Kandasamy, C. Nwankpa, and D. R. Kaeli, “Datacenters as controllable load resources in the electricity market,” in *Proc. of the IEEE ICDCS*, Philadelphia, PA, July 2013.

[7] H. Chen, M. Caramanis, and A. K. Coskun, “The data center as a grid load stabilizer,” in *Proc. of ASPDAC*, Tokyo, Japan, January 2014.

[8] M. Ghamkhari and H. Mohsenian-Rad, “Optimal integration of renewable energy resources in data centers with behind-the-meter renewable generators,” in *Proc. of the IEEE ICC*, Ottawa, Canada, June 2012.

[9] I. Goiri, K. Le, T. Nguyen, J. Guitart, J. Torres, and R. Bianchini, “Greenhadoop: Leveraging green energy in data-processing frameworks,” in *Proc. of the ACM EuroSys*, Bern, Switzerland, April 2012.

[10] Y. Guo, Z. Ding, Y. Fang, and D. Wu, “Cutting down electricity cost in internet data centers by using energy storage,” in *IEEE Global Telecommunications Conference*, Houston, TX, December 2011.

[11] L. Rao, X. Liu, L. Xie, and W. Liu, “Minimizing electricity cost: Optimization of distributed internet data centers in a multi-electricity-market environment,” in *Proc. of IEEE INFOCOM*, San Diego, CA, March 2010.

[12] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. H. Andrew, “Geographical load balancing with renewables,” in *Proc. of the ACM GreenMetrics Workshop*, San Jose, CA, June 2011.

[13] D. Kusic, J.Kephart, J. Hanson, N. Kandasamy, and G. Jiang, “Power and performance management of virtualized computing environments via lookahead control,” in *Proc. of ICAC*, Chicago, IL, June 2008.

[14] L. Yu, T. Jiang, Y. Cao, and J. Wu, “Risk-constrained operation for internet data centers under smart grid environment,” in *Proc. of IEEE WCSP*, Hangzhou, China, October 2013.

[15] Q. Zhang, “Risk-constrained operation for internet data centers in deregulated electricity markets,” *IEEE Trans. on Parallel and Distributed Systems*, vol. 25, no. 5, pp. 1306 – 1316, May 2014.

[16] L. Rao, X. Liuand, L. Xie, and Z. Pang, “Hedging against uncertainty: A tale of internet data center operations under smart grid environment,” *IEEE Trans. on Smart Grid*, vol. 2, no. 3, pp. 555–563, 2011.

[17] B. Aksanli and T. Rosing, “Providing regulation services and managing data center peak power budgets,” in *Proc. of DATE*, Grenoble, France, Mar. 2014.

[18] R. Urgaonkar, B. Urgaonkar, M. J. Neely, and A. Sivasubramaniam, “Optimal power cost management using stored energy in data centers,” in *In Proc. of the ACM International Conference on Measurement and Modeling of Computer Systems*, San Jose, California, USA, June 2011.

[19] W. Deng, F. Liu, H. Jin, and X. Liao, “Online control of data center power supply under uncertain demand and renewable energy,” in *Proc. of IEEE ICC*, Budapest, Hungary, June 2013.

[20] C. Wang and M. Groot, “Enabling demand response in a computer cluster,” in *Proc. of IEEE Smart Grid Comm*, Vancouver, BC, October 2013.

[21] Y. Choi and Y. Lim, “A cost-efficient mechanism for dynamic VM provisioning in cloud computing,” in *Proc. of ACM RACS*, Towson, MD, October 2014.

[22] R. Buyya, A. Beloglazov, and J. Abawajy, “Energy-efficient management of data center resources for cloud computing: A vision, architectural elements, and open challenges,” in *Proc. of International Conference on Parallel and Distributed Processing Techniques and Applications*, Las Vegas, Nv, July 2010.

[23] Y. Guo, Y. Gong, Y. Fang, P. Khargonekar, and X. Geng, “Optimal power and workload management for green data centers with thermal storage,” in *Proc. of IEEE Globecom*, December 2013.

[24] Z. Zhou, F. Liu, Z. Li, and H. Jin, “When smart grid meets geo-distributed cloud: An auction approach to datacenter demand response,” in *Proc. of IEEE INFOCOM*, Hong Kong, April 2015.

[25] J. Xu, P. Luh, F. White, E. Ni, and K. Kasiviswanathan, “Power portfolio optimization in deregulated electricity markets with risk management,” *IEEE Trans. on Power Systems*, vol. 21, pp. 1653–1662, November 2006.

[26] K. Zare, M. P. Moghaddam, and M. K. Sheikh-El-Eslami, “Risk-based electricity procurement for large consumers,” *IEEE Trans. on Power Systems*, vol. 26, no. 4, November 2011.

[27] M. Shahidehpour, H. Yamin, and Z. Li, *Market Operations in Electric Power Systems*. New York, NY: John Wiley & Sons, 2002.

[28] S. Li, M. Brocanelli, W. Zhang, and X. Wang, "Integrated power management of data centers and electric vehicles for energy and regulation market participation," *IEEE Trans. on Smart Grid*, vol. 5, no. 5, September 2014.

[29] B. Kirby, "Spinning reserve from responsive loads," Oak Ridge National Laboratory, 2003.

[30] ——, "Ancillary services: technical and commercial insights," prepared for Wrtsil, July 2007.

[31] ERCOT, "Load participation in the ercot nodal market," June 2007.

[32] H. Mohsenian-Rad and A. Leon-Garcia, "Coordination of cloud computing and smart power grids," in *Proc. of IEEE International Conference on Smart Grid Communications*, Gaithersburg, MD, October 2010.

[33] ——, "Energy-information transmission tradeoff in green cloud computing," in *Proc. of IEEE Conference on Global Communications (Globecom)*, Miami, FL, December 2010.

[34] J. Dilley, "Web server workload characterization," *Hewlett-Packard Laboratories*, 1996.

[35] U.S. Environmental Protection Agency, "EPA report on server and data center energy efficiency," Final Report to Congress, August 2007.

[36] X. Fan, W. D. Weber, and L. A. Barroso, "Power provisioning for a warehouse-sized computer," in *Proc. of the ACM International Symposium on Computer Architecture*, San Diego, CA, June 2007.

[37] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo Method*, 2nd ed. Wiley, December 2007.

[38] M. Ghamkhari and H. Mohsenian-Rad, "Energy and performance management of green data centers: A profit maximization approach," *IEEE Trans. on Smart Grid*, vol. 4, no. 2, pp. 1017–1025, June 2013.

[39] ——, "A convex optimization framework for service rate allocation in finite communications buffers," *Accepted for publication in IEEE Communications Letters*, 2015.

[40] M. Dicorato, G. Forte, M. Trovato, and E. Caruso, "Risk-constrained profit maximization in day-ahead electricity market," *IEEE Transactions on Power Systems*, vol. 24, no. 3, pp. 1107–1114, August 2009.

[41] J. D. Molina, J. Contreras, and H. Rudnick, "Risk-constrained project portfolio in centralized transmission expansion planning," *IEEE Systems journal*, vol. PP, no. 99, pp. 1–9, 2014.

[42] Y. Zhang and G. B. Giannakis, "Robust optimal power flow with wind integration using conditional value-at-risk," in *Proc. of IEEE Conference on Smart Grid Communications*, Vancouver, BC, October 2013.

[43] D. Panda, S. N. Singh, and V. Kumar, "Risk constraint profit maximization in a multi-electricity market," in *Proc. of IEEE PES General Meeting,*, Washington, DC, July 2014.

[44] R. T. Rockafellar and S. Uryasev, "Conditional value-at-risk for general loss distributions," *Journal of Banking and Finance*, vol. 26, no. 7, pp. 1443–1471, 2002.

[45] ——, "Optimization of conditional value-at-risk," *Journal of Risk*, vol. 2, pp. 21–41, 2000.

[46] C. M. Grinstead and J. Snell, *Introduction to Probability*. American Mathematical Society, 1997.

[47] S. Sarykalin, G. Serraino, and S. Uryasev, "Value-at-risk vs. conditional value-at-risk in risk management and optimization." *Tutorials in Operations Research. INFORMS*, pp. 270–294, 2008.

[48] O. K. Gupta and A. Ravindran, "Branch and bound experiments in convex nonlinear integer programming," *Management Science*, vol. 31, no. 12, pp. 1533–1546, 1985.

[49] ftp://public.dhe.ibm.com/software/websphere/ilog/docs/optimization/cplex/ps_usrmancplex.pdf.

[50] A. Harel, "Convexity properties of the erlang loss formula," *Operation Research*, vol. 38, no. 3, pp. 499–505, May 1990.

[51] P. A. Jensen and J. F. Bard, *Operations Research Models and Methods*. Wiley, 2003.

[52] S. P. Bradley, A. C. Hax, and T. L. Magnanti, *Applied Mathematical Programming*. Addison-Wesley, 1977.

[53] A. Astorino, A. Frangioni, M. Gaudioso, and E. Gorgone, "Piecewise-quadratic approximations in convex numerical optimization," *SIAM Journal on Optimization*, vol. 21, no. 4, pp. 1418–1438, 2011.

[54] H. Zhang and S. Wang, "Linearly constrained global optimization via piecewise-linear approximation," *Journal of Computational and Applied Mathematics*, vol. 214, no. 1, pp. 111–120, April 2008.

[55] C. Y. Kao and R. R. Meyer, "Secant approximation methods for convex optimization," *Computer Sciences Technical Report 352*, April 1979.

[56] L. S. Thakur, "Error analysis for convex separable programs: The piecewise linear approximation and the bounds on the optimal objective value," *SIAM Journal on Applied Mathematics*, vol. 34, no. 4, pp. 704–714, June 1978.

[57] F. Guder and J. G. Morris, "Optimal objective function approximation for separable convex quadratic programming," *Journal of Mathematical Programming*, vol. 67, no. 3, pp. 133–142, October 1994.

[58] B. Feijoo and R. R. Meyer, "Piecewise-linear approximation methods for nonseparable convex optimization," *Management Science*, vol. 34, no. 3, pp. 411–419, March 1988.

[59] D. P. Bertsekas and H. Yu, "A unifying polyhedral approximation framework for convex optimization," *SIAM Journal on Optimization*, vol. 21, no. 1, pp. 333–360, 2011.

[60] D. P. Bertsekas, *Convex Optimization Theory*. Nashua, NH: Athena Scientific, 2009. [Online]. Available: http://www.athenasc.com/convexdualitychapter.pdf

[61] http://www.pjm.com/markets-and-operations/energy/real-time/monthlylmp.aspx.

[62] http://www.pjm.com/markets-and-operations/ancillary-services.aspx.

[63] http://www.pjm.com/markets-and-operations/rpm.aspx.

[64] "Frequency regulation compensation in the organized wholesale power markets," United States of America federal energy regulatory commission Docket Nos. RM11-7-001 & AD10-11-001, February 2012.

[65] http://www.windenergy.org/.

[66] http://www.endurancewindpower.com/e3120.html.

[67] http://www.wikibench.eu/wiki/2007-09/.

Mahdi Ghamkhari received his B.Sc. degree from Sharif University of Technology in 2011, and his M.Sc. degree from Texas Tech University in 2012, both in Electrical Engineering. He is now pursuing his PhD degree in University of California, Riverside. His research interests includes energy management of data centers, and application of convex optimization in power systems and smart power grids.

Adam Wierman is a Professor in the Department of Computing and Mathematical Sciences at the California Institute of Technology, where he is a founding member of the Rigorous Systems Research Group (RSRG) and maintains a popular blog called Rigor + Relevance. His research interests center around resource allocation and scheduling decisions in computer systems and services. He received the 2011 ACM SIGMETRICS Rising Star award, the 2014 IEEE Communications Society William R. Bennett Prize, and has been coauthor on papers that received of best paper awards at ACM SIGMETRICS, IEEE INFOCOM, IFIP Performance (twice), IEEE Green Computing Conference, IEEE Power & Energy Society General Meeting, and ACM GREENMETRICS.

Hamed Mohsenian-Rad is an Assistant Professor of Electrical Engineering at the University of California - Riverside. He received his Ph.D. degree in Electrical and Computer Engineering from The University of British Columbia - Vancouver in 2008. Dr. Mohsenian-Rad is the recipient of the NSF CAREER Award and the Best Paper Award from the IEEE International Conference on Smart Grid Communications 2012. He is also a coauthor of a Best Paper on System Operations and Market Economics in the IEEE Power and Energy Society General Meeting 2013. Dr. Mohsenian-Rad is an Associate Editor of the IEEE Communications Surveys and Tutorials, an Associate Editor of the IEEE Communication Letters, a Guest Editor of the ACM Transactions on Embedded Computing Systems - Special Issue on Smart Grid, and a Guest Editor of the KICS/IEEE Journal of Communications and Networks - Special Issue of Smart Grid. His research interests include the design, optimization, and game-theoretic analysis of power systems and smart power grids.