UNIVERSITY OF CALIFORNIA
RIVERSIDE

Optimal Operation of Data Centers in Future Smart Grid

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering

by

Seyed Mahdi Ghamkhari

June 2016

Dissertation Committee:

Dr. Hamed Mohsenian Rad, Chairperson
Dr. Nael Abu-Ghazaleh
Dr. Yingbo Hua

The Dissertation of Seyed Mahdi Ghamkhari is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I would like to thank Dr. Hamed Mohsenian Rad for his excellent guidance and full support throughout my PhD studying. The content of this thesis is a reprint of the material that are appeared in the following publications:

- M. Ghamkhari and H. Mohsenian-Rad, "Data Centers to Offer Ancillary Services", in Proc. of the IEEE Conference on Smart Grid Communications, Tainan City, Taiwan, October 2012.
- 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 the IEEE INFOCOM Workshop on Smart Data Pricing, Toronto, ON, Canada, May 2014.
- M. Ghamkhari and H. Mohsenian-Rad, "A Convex Optimization Framework for Service Rate Allocation in Finite Communications Buffers", IEEE Communications Letters, vol. 20, no. 1, January 2016.
- M. Ghamkhari, A. Wierman, and H. Mohsenian-Rad, "Energy Portfolio Optimization of Data Centers", Accepted for Publication in IEEE Transactions on Smart Grid, 2016.

To my parents.

ABSTRACT OF THE DISSERTATION

Optimal Operation of Data Centers in Future Smart Grid

by

Seyed Mahdi Ghamkhari

Doctor of Philosophy, Graduate Program in Electrical Engineering
University of California, Riverside, June 2016
Dr. Hamed Mohsenian Rad, Chairperson

The emergence of cloud computing has established a growing trend towards building massive, energy-hungry, and geographically distributed data centers. Due to their enormous energy consumption, data centers are expected to have major impact on the electric grid by significantly increasing the load at locations where they are built. However, data centers also provide opportunities to help the grid with respect to robustness and load balancing. For instance, as data centers are *major* and yet *flexible* electric loads, they can be proper candidates to offer ancillary services, such as voluntary load reduction, to the smart grid. Also, data centers may better stabilize the price of energy in the electricity markets, and at the same time reduce their electricity cost by exploiting the diversity in the price of electricity in the day-ahead and real-time electricity markets. In this thesis, such potentials are investigated within an analytical profit maximization framework by developing new mathematical models based on queuing theory. The proposed models capture the trade-off between quality-of-service and power consumption in data centers. They are not only accurate, but also they posses convexity characteristics that facilitate joint optimization of data centers' service rates, demand levels and demand bids to different electricity markets.

The analysis is further expanded to also develop a unified comprehensive energy portfolio optimization for data centers in the future smart grid. Specifically, it is shown how utilizing one energy option may affect selecting other energy options that are available to a data center. For example, we will show that the use of on-site storage and the deployment of geographical workload distribution can particularly help data centers in utilizing high-risk energy options such as renewable generation. The analytical approach in this thesis takes into account service-level-agreements, risk management constraints, and also the statistical characteristics of the Internet workload and the electricity prices. Using empirical data, the performance of our proposed profit maximization models for data centers are evaluated, and the capability of data centers to benefit from participation in a variety of Demand Response programs is assessed.

Contents

List of Figures

Chapter 1

Data Centers to Offer Ancillary Services

1.1 Introduction

In an interconnected power system, Independent System Operators (ISOs) are responsible for coordinating, controlling, and monitoring the operation of the power grid for generation, transmission and distribution. An ISO is also responsible for maintaining a required level of power quality and reliability, e.g. by making a constant balance between supply and demand across the power grid. For this purpose, the ISOs rely on receiving different types of ancillary services from a variety of entities that are involved in the power system. Examples of ancillary services include frequency and voltage regulation, spinning reserves, and non-spinning reserves [72].

Ancillary services are usually procured from generators that are online and can increase or decrease their generation in response to the requests sent by ISOs. However, there is a growing interest towards procuring ancillary services from not only generators but also load resources. Reduced transmission and distribution losses,

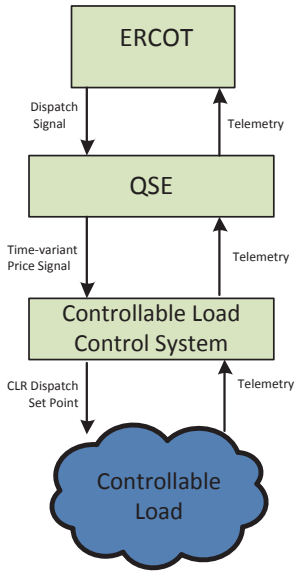

Figure 1.1: The ancillary service market model for controllable load in ERCOT [33]. A data center can act as a load resource to offer vulnerary load reduction.

increased transmission capacity and increased margin to voltage collapse are among the benefits in supplying ancillary services from load resources [49]. Load resources are utilized to offer *voluntary load reduction*, with a compensation value equivalent to what a power supplier is paid for generating the same amount of electricity [34].

The ancillary service market model for controllable load in ERCOT ISO is shown in Fig. 1.1. A key entity to facilitate load participation in ancillary service market is Qualified Scheduling Entity (QSE). It acts as an interface between ERCOT and controllable load resources (CLRs). The QSE coordinates the operation of CLRs based on the commands it receives from ERCOT. The QSE also aggregates the ancillary services that CLRs may offer. The current total CLR capacity in ERCOT is about 36 MW [75]. They are eligible to participate in both *regulation* and *voluntary load reduction* services [33]. However, our focus in this chapter is on the latter.

In this chapter, we would like to investigate the potential for Internet and cloud

computing data centers to participate in an ancillary service market to offer voluntary load reduction. Data centers have two important features that make them very good candidates to efficiently offer such services. First, data centers, such as those built and operated by Google, Microsoft, and Amazon, have significant daily and peak power consumption. For example, the peak power consumption of Microsoft's data center in Quincy, WA is 48 megawatts, which is enough to power 40,000 homes [78]. Second, data centers have flexible load and they are able to respond to the QSE's signals quickly and reduce their power consumption by switching off a group of computer servers or by migrating a portion of their workload to another data center [70]. These features suggest that data center participation in the ancillary service market can be promising. To the best of our knowledge, this chapter is the first to study the capability of data centers in offering ancillary services, in particular in form of voluntary load reduction. Our contributions in this chapter can be summarized as follows.

- We develop a mathematical model for data center's profit when it offers ancillary services. Our model includes elements with respect to the data center's *revenue* obtained from the Internet services it offers, the data center's *cost* of electricity, and the *compensation* that the data center receives for offering ancillary services. We take into account server's power consumption profiles, data center's power usage effectiveness, price of electricity, workload statistics, and service-level agreements (SLAs).

- We propose an optimization-based *profit maximization strategy* for data centers, when they offer ancillary services in form of voluntary load reduction. To gain insights, we also provide a *geometric interpretation* of the optimal solution of the profit maximization problem.

- Using experimental data, e.g., for workload, price of electricity, and SLA parameters, we assess the performance of the proposed optimization-based profit maximization strategy via computer simulations. We show that a data center can noticeably increase its profit by participating in a voluntary load reduction ancillary service program.

The rest of this chapter is organized as follows. The system model is described in Section 1.2. The mathematical expressions for revenue, cost, and ancillary service compensation are derived in Section 1.3. Our proposed profit maximization design framework is discussed in Section 1.4. Simulation results are presented in Section 1.5. The chapter is concluded in Section 1.6.

1.2 System Model

1.2.1 Power Consumption

Consider an Internet or cloud computing data center with $M_{\max}$ computer servers. The total power consumption in a data center is obtained by adding the total power consumption at computer servers to the total power consumption at the facility, e.g., for cooling, lighting, etc. For a data center, *power usage effectiveness* (PUE), denoted by E_{usage}, is defined as the ratio of the data center's total power consumption to the data center's power consumption at the computer servers [91]. The PUE is considered as a measure for data center's energy efficiency. Currently, the typical value for most data centers is around 2.0. However, recent studies have suggested that many data centers can soon reach a PUE of 1.7. A few state-of-the art facilities have reached a PUE of 1.2 [91].

Let P_{idle} denote the average idle power draw of a single server and P_{peak} denote the

average peak power when a server is handling a service request. The ratio $P_{\text{peak}}/P_{\text{idle}}$ denotes the power elasticity of servers. Higher elasticity means less power consumption when the server is idle, not handling any service request. Let $M \leq M_{max}$ denote the number of servers that are 'on' at data center. The total electric power consumption associated with the data center can be obtained as [12]:

$$\begin{aligned} P = M[P_{\text{idle}} + (E_{\text{usage}} - 1)P_{\text{peak}} + \\ (P_{\text{peak}} - P_{\text{idle}})U], \end{aligned} \tag{1.1}$$

where U is the CPU utilization of servers. From (1.1), the power consumption at data center increases as we turn on more computer servers or run servers at higher utilization.

1.2.2 Electricity Price

The electricity pricing models that are deployed for each region usually depend of whether the electricity market is regulated or deregulated in that region. In ERCOT, the electricity market is mostly deregulated. Therefore, the prices may vary during the day due to the fluctuations in the wholesale market. Some of the common non-flat retail pricing tariffs in deregulated electricity markets include: Day-ahead pricing (DAP), time-of-use pricing (TOUP), critical-peak pricing (CPP), and real-time pricing (RTP). In our system model, the instantaneous price of electricity is denoted by ω which is assumed to be known at least 15 minutes in advance.

1.2.3 Voluntary Load Reduction as Ancillary Service

The ERCOT ancillary service market is designed with a number of features to reward consumers that are willing to curtail their load when needed to help maintain

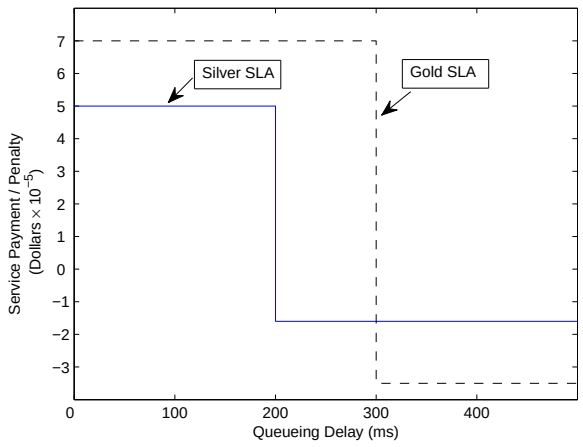

Figure 1.2: Two sample service-level agreements (SLAs) in data centers [28].

system reliability. In particular, consumers that offer voluntary load reduction are compensated in dollars per megawatt hour of load reduction at rates that are set based on the market clearing prices within the ERCOT-operated markets [34]. Given the data centers' major and flexible load, they can register as load resources and respond to the load reduction request signals that are sent by the QSEs in their region. The exact type of contracts to compensate such data centers would depend on the mutual agreements between the QSEs and the data centers [34]. Here, we assume that the QSE may send out a load reduction request periodically, e.g., once every 15 minutes. The request contains a compensation function $\tau(\cdot)$, which is calculated by the QSE based on a market clearing price analysis and indicates the dollars to be paid to the data center for each megawatt hour voluntary load reduction. The compensation function may or may not be linear. More details on compensation function will be discussed in Section 1.3.3.

1.2.4 Quality-of-Service Offered by Data Center

Because of the limited computation capacity of data centers and given the stochastic nature of most practical workload, data centers cannot process the incoming service requests immediately after they arrive. Therefore, all arriving service requests are first placed in a queue until they can be handled by an available server. In order to satisfy quality-of-service requirements, the waiting time/queuing delay for each incoming service request should be limited within a certain range which is determined by the *Service Level Agreement* (SLA). The exact SLA depends on the type of service offered which may range from cloud-based computational tasks to video streaming and HTML web services. Examples of two typical SLAs are shown in Fig. 1.2 [28]. In this figure, each SLA is identified by three non-negative parameters D, δ, and γ. Parameter D indicates the maximum 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 penalty that the data center has to pay every time it *cannot* handle a service request before deadline D. For the Gold SLA in Fig. 1.2, we have $D = 300$ ms, $\delta = 7 \times 10^{-5}$ dollars, and $\gamma = 3.5 \times 10^{-5}$ dollars. For the Silver SLA, we have $D = 200$ ms, $\delta = 5 \times 10^{-5}$ dollars, and $\gamma = 1.6 \times 10^{-5}$ dollars.

1.2.5 Service Rate

Let μ denote the rate at which service requests are removed from the queue and handled by a server. The service rate depends on the number of servers that are switched on. Let S denote the time it takes for a server to finish handling a service request. Each server can handle $\kappa = 1/S$ service requests per second. Therefore, the

total service rate is obtained as

$$\mu = \kappa M \quad \Rightarrow \quad M = \frac{\mu}{\kappa}. \tag{1.2}$$

As we increase the number of switched on servers and thus the service rate, more service requests can be handled before the SLA-deadline D, which in turn increases the payments that the data center receives. However, it also increases the data center's power consumption and thus the data center's energy expenditure. Furthermore, turning on more computer servers degrades the data center's ability to offer load reduction as an ancillary service. Therefore, there is a *trade-off* in selecting the data center's service rate, as we will discuss next.

1.3 Revenue, Cost, and Compensation Models

The rate at which *service requests* arrive at a data center can vary over time. To improve data center's performance, the number of switched on servers M should be adjusted in proportional to demand. More servers should be turned on when service requests are received at higher rates. However, because of the tear-and-wear cost of switching servers on and off, and also due to the delay in changing the status of a computer, M cannot be changed rapidly. It is rather desired to be updated only every few minutes. Therefore, we divide running time of data center into a sequence of time slots $\Lambda_1, \Lambda_2, \cdots, \Lambda_N$, each one with length T. The number of switched on servers are updated only at the beginning of each time slot. We assume that $T = 15$ minutes, so that making decision about the number of switched on computers is periodic and is done at the same time that the data center receives a load reduction request from the QSE. For the rest of this section, we focus on mathematically modeling the energy

cost and revenue of the data center of interest at each time slot $\Lambda \in \Lambda_1, \Lambda_2, \cdots, \Lambda_N$ as a function of service rate μ and consequently as a function of M, based on (1.2).

1.3.1 Revenue Modeling

Let $q(\mu)$ denote the probability that the waiting time for a service request exceeds the SLA-deadline D. Obtaining an analytical model for $q(\mu)$ requires a queueing theoretic analysis that we will provide in Section 1.3.4. Next, assume that λ denotes the average rate of receiving service requests within time slot Λ of length T. The total revenue collected by the data center at the time slot of interest can be calculated as

$$Revenue = (1 - q(\mu))\delta\lambda T - q(\mu)\gamma\lambda T, \tag{1.3}$$

where $(1 - q(\mu))\delta\lambda T$ denotes the total payment received by the data center within interval T, for the service requests that are handled before the SLA-deadline, while $q(\mu)\gamma\lambda T$ denotes the total penalty paid by the data center within interval T for the service requests that are not handled before the SLA-deadline.

1.3.2 Cost Modeling

Within time interval T, each turned on server handles

$$\frac{T(1 - q(\mu))\lambda}{M} \tag{1.4}$$

service requests. This makes each server busy for $T(1 - q(\mu))\lambda/\kappa M$ seconds. By dividing the total CPU busy time by T, the CPU utilization for each server is obtained

as

$$U = \frac{(1 - q(\mu))\lambda}{\kappa M}. \tag{1.5}$$

Replacing (1.2) and (1.5) in (1.1), the power consumption associated with the data center at the time slot of interest is obtained as

$$P(\mu) = \frac{a\mu + b\lambda(1 - q(\mu))}{\kappa}, \tag{1.6}$$

where $a \triangleq P_{\text{idle}} + (E_{\text{usage}} - 1)P_{\text{peak}}$ and $b \triangleq P_{\text{peak}} - P_{\text{idle}}$. Multiplying (1.6) by the electricity price ω, the total energy cost at the time interval of interest is obtained as

$$Cost = T\omega \left[\frac{a\mu + b\lambda(1 - q(\mu))}{\kappa} \right]. \tag{1.7}$$

1.3.3 Ancillary Service Compensation Model

Let Λ_0 denote the time slot right before the current time slot Λ. Also let P_0 denote the total power consumption at time slot Λ_0. If the data center reduces its load by $\Delta P = P_0 - P(\mu)$ compared to the previous time slot, then the QSE pays

$$Compensation = \tau(\Delta P) \tag{1.8}$$

to the data center in order to compensate for the voluntary load reduction ancillary service that is offered by the data center. The choice of compensation function $\tau(\cdot)$ is set by the QSE. If no load reduction ancillary service is needed at a time slot, then we simply have $\tau(\Delta P) = 0$. If the compensation function is linear, then we have $\tau(\Delta P) = c\,\Delta P$, where higher c indicates higher compensation rates, e.g., due to a

more severe need for load reduction. Other forms of compensation functions may include quadratic or piece-wise linear functions. It is worth mentioning that once the compensation function is announced by the QSE, it is up to the data center to decide whether offering load reduction ancillary service is beneficial.

1.3.4 Probability Model of $q(\mu)$

Consider a new service request that arrives within time slot Λ. Let Q denote the number of service requests waiting in the service queue right before the arrival of the new service request. Since the data center's service rate is μ, it takes Q/μ seconds until all existing requests are removed from the queue. Hence, the new service request can be handled after Q/μ seconds since its arrival. According to the SLA, if $Q/\mu \leq D$, then the request is handled before the deadline D. If $Q/\mu > D$, the request is not handled before the deadline D and it is dropped. Therefore, we can model the SLA-deadline by a *finite-size queue* with the length $\mu\, D$. A service request can be handled before the SLA-deadline, if and only if it enters the aforementioned finite size queue. We assume that the service request rate has an arbitrary and *general* probability distribution function. On the other hand, since the service rate μ is fixed over each time interval of length T, $q(\mu)$ can be modeled as the *loss probability* of a G/D/1 queue. Therefore, following the queuing theoretic analysis in [46], we can obtain

$$q(\mu) = \alpha(\mu)\; e^{-\frac{1}{2}\min\limits_{n\geq 1}\; m_n(\mu)}, \tag{1.9}$$

where

$$\alpha(\mu) = \frac{1}{\lambda\sqrt{2\pi}\sigma} e^{\frac{(\mu-\lambda)^2}{2\sigma^2}} \int_{\mu}^{\infty} (r-\mu) e^{-\frac{(r-\lambda)^2}{2\sigma^2}} dr \tag{1.10}$$

and for each $n \geq 1$ we have

$$m_n(\mu) = \frac{(D\mu + n(\mu - \lambda))^2}{nC_\lambda(0) + 2\sum_{l=1}^{n-1} C_\lambda(l)(n-l)}. \tag{1.11}$$

Here, $\sigma = C_\lambda(0)$ and C_λ denotes the auto-covariance of the service request rate's probability distribution function [69].

1.4 Profit Maximization

1.4.1 The Case without Offering Ancillary Service

For the case where the data center does not offer ancillary service, its profit at each time slot Λ is obtained as

$$Profit = Revenue - Cost, \tag{1.12}$$

where revenue is as in (1.3) and cost is as in (1.7). We seek to choose the data center's service rate μ to maximize profit. This can be expressed as the following optimization problem:

$$\begin{aligned} \underset{\lambda \leq \mu \leq \kappa M_{max}}{\textbf{Maximize}} \; & T\lambda\left[(1 - q(\mu))\delta - q(\mu)\gamma\right] - \\ & T\omega\left(\frac{a\mu + b\lambda(1 - q(\mu))}{\kappa}\right), \end{aligned} \tag{1.13}$$

where the probability $q(\mu)$ is as in (1.9). We note that the service rate μ is lower bounded by λ. This is necessary to assure stabilizing the service request queue [101, 46]. We also note that problem (1.15) needs to be solved separately for every time slot $\Lambda \in \{\Lambda_1, \ldots, \Lambda_N\}$, i.e., once every T minutes.

1.4.2 The Case with Offering Ancillary Service

For the case where data center does offer ancillary service, the data center's profit at each time slot Λ is obtained as

$$Profit = Revenue - Cost + Compensation, \tag{1.14}$$

where compensation is as in (1.8). We seek to choose the data center's service rate μ to maximize profit. This can be expressed as the following optimization problem:

$$\begin{aligned} \underset{\lambda \le \mu \le \kappa M_{max}}{\textbf{Maximize}} \; & T\lambda \left[(1 - q(\mu))\delta - q(\mu)\gamma \right] - \\ & T\omega \left(\frac{a\mu + b\lambda(1 - q(\mu))}{\kappa} \right) + \\ & T\tau \left(P_0 - \frac{a\mu + b\lambda(1 - q(\mu))}{\kappa} \right), \end{aligned} \tag{1.15}$$

where the last term is based on the definition of ΔP in Section 1.3.3 and the expression for power consumption in (1.6).

1.4.3 Geometric Interpretation

Let $Profit_{Base}$ denote the optimal objective value of problem (1.13). That is, the maximum profit that a data center can obtain, by properly selecting its service rate, when the data center does *not* offer any ancillary service. Clearly, offering ancillary service is beneficial if the profit when offering ancillary service is greater

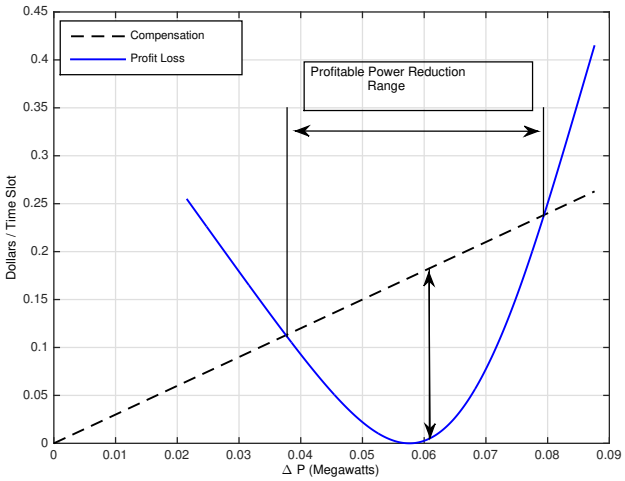

Figure 1.3: An example to compare profit loss due to load reduction versus the compensation obtained by the data center due to offering ancillary service. The arrow indicates the optimal amount of load reduction ΔP.

than $Profit_{Base}$. That is,

$$\begin{aligned} &T\lambda\left[(1-q(\mu))\delta - q(\mu)\gamma\right] - \\ &T\omega\left(\frac{a\mu + b\lambda(1-q(\mu))}{\kappa}\right) + \\ &T\tau\left(P_0 - \frac{a\mu + b\lambda(1-q(\mu))}{\kappa}\right) \geq Profit_{Base}. \end{aligned} \tag{1.16}$$

From the definition of ΔP, we have

$$P(\mu) = P_0 - \Delta P. \tag{1.17}$$

Therefore, we can write μ in terms of ΔP as follows:

$$\mu = P^{-1}(P_0 - \Delta P), \tag{1.18}$$

where $P^{-1}(\cdot)$ denotes the inverse of function $P(\mu)$ in (1.6). Note that, since $q(\mu)$ is a non-increasing function of μ [69, Theorem 1], $P(\mu)$ is an increasing function of service rate μ. Therefore, $P(\mu)$ is an invertible function. From (1.18) and after reordering the terms, we can rewrite condition (1.16) as

$$\begin{aligned}\tau(\Delta P) > Profit_{Base}/T - \lambda \big[(1 - q(P^{-1}(P_0 - \Delta P)))\delta \\ - q(P^{-1}(P_0 - \Delta P))\gamma\big] + \omega(P_0 - \Delta P).\end{aligned} \tag{1.19}$$

First, we note that both sides of condition (1.19) are written in terms of ΔP as the only variable. Second, while the left hand side in (1.19) is the ancillary service compensation function at load reduction level ΔP, the right hand side of (1.19) denotes the data center's *profit loss* due to reducing its load by ΔP if the QSE does *not* compensate the offered load reduction service. Therefore, we can conclude that offering voluntary load reduction at level ΔP is profitable if and only if

$$Compensation > Profit\,Loss, \tag{1.20}$$

where $Profit\,Loss$ is defined as the expression in the right hand side of (1.19). The geometric interpretation of the above condition is shown in Fig. 1.3. Here, the compensation function $\tau(\cdot)$ is assumed to be linear. We can see that offering load reduction ancillary service less than $\Delta P = 0.0378$ Megawatts or more than $\Delta P = 0.0793$ Megawatts is not profitable for the data center. Furthermore, we can see that the maximum profit is gained if the data center offers load reduction at $\Delta P = 0.0608$ Megawatts. This is essentially the same amount that the data center chooses to offer

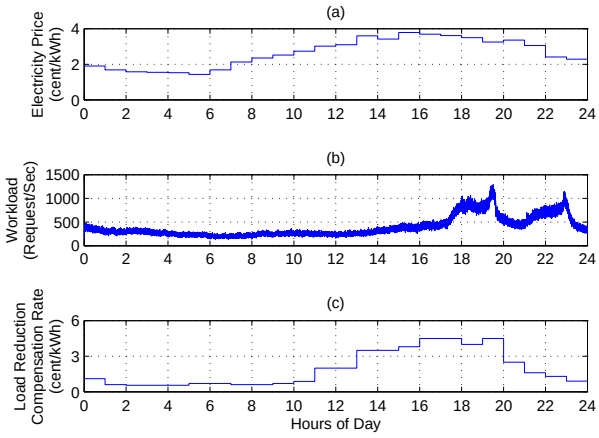

Figure 1.4: A sample set of data over 24 hours that is used for simulation studies. (a) Time-of-use electricity prices [1]. (b) Internet workload [2]. (c) The price of compensating voluntary load reduction as ancillary service [34].

after solving problem (1.15).

1.5 Case Studies

Consider a data center with a maximum of $M_{\max} = 50000$ servers. The exact number of switched on servers M is updated periodically for each time slot of length $T = 15$ minutes by solving the profit maximization approaches explained in Section 1.4. For each switched on server, we have $P_{\text{peak}} = 200$ watts and $P_{\text{idle}} = 100$ watts [70]. We assume that $E_{\text{usage}} = 1.2$ [91]. The electricity price information is based on the hourly real-time pricing tariffs currently practiced in Illinois Zone I, spanning from June 10, 2011 to July 9, 2011 [1]. We assume that $\kappa = 0.1$ and the Gold SLA is used. To simulate the total workload, we use the publicly available World Cup 98 web hits data, spanning from June 10, 1998 to July 1998, as the trend for the incoming service requests [2]. We assume that the ancillary service compensation function is linear and it is set according to the prices used in ERCOT [34]. A sample daily data

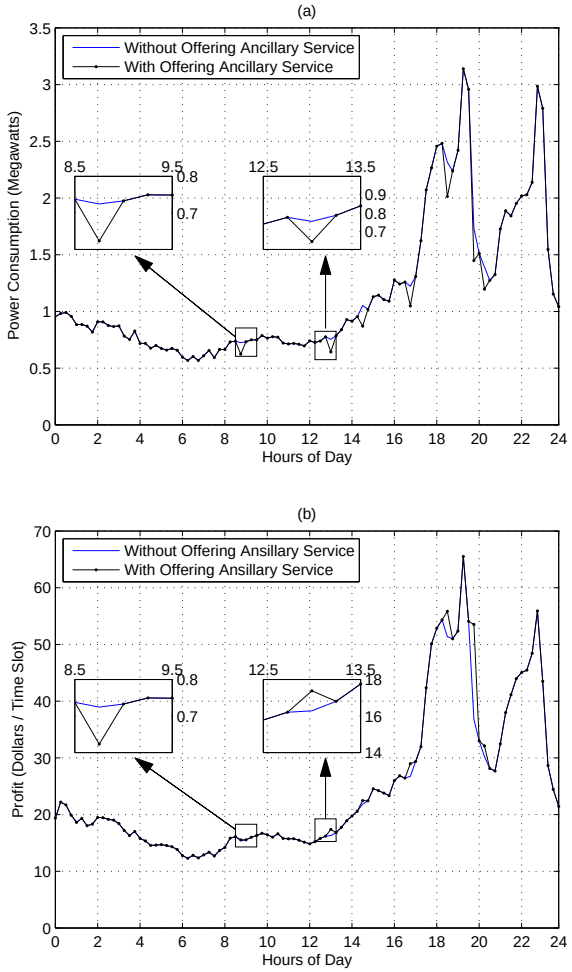

Figure 1.5: The data center receives load reduction requests from QSE in seven time slots. It accordingly reduces its load that results in additional profit.

set that we used in our simulation are shown in Fig. 1.4.

The results for a sample daily power consumption trend are shown in Fig. 1.5(a). It is assume that the QSE sends requests for load reduction in seven time slots. In all cases, the data center chooses to respond by reducing its load. Recall that the amount of load reduction is set based on the optimal solution of problem (1.15). The reduced power consumption for two time slots, one around 8:30 AM and one around 12:15 PM are zoomed in. The data center's corresponding profit is shown in 1.5(b). We can see that every time that the data center reduces its load in response to the

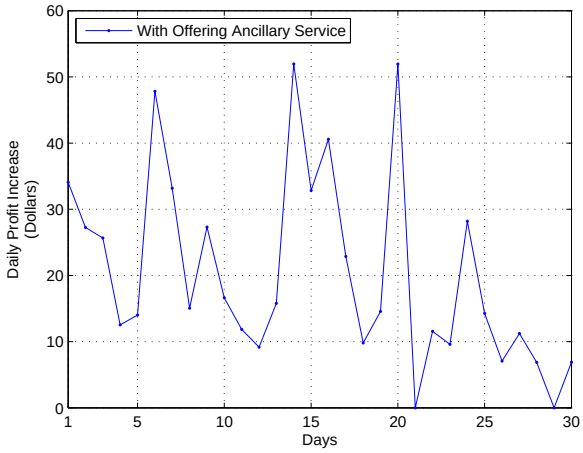

Figure 1.6: The daily additional profit the data center gains over 30 days due to offering voluntary load reduction as ancillary service to smart grid.

QSE's request its profit increases.

The daily increase in the data center's profit due to offering voluntary load reduction is shown in Fig. 1.6 over 30 days. We can see that, on average, the data center's daily profit increases by 22.7 dollars, summing up to a monthly increase of 681 dollars. For the results in this figure, we have assumed that the QSE sends load reduction requests in 20% of the time slots. Clearly, a higher number of load reduction requests to be sent by the QSE can provide the data center with more opportunities to further increase its profit. This is shown in Fig. 1.7. The results in this figure are based on repeating the simulations in Fig. 1.6 for different percentages of the number of time slots in which the QSE sends load reduction requests. We can see that, the data center's profit increases as the participates of time slots with voluntary load reduction opportunities increases. If a load reduction request is sent by the QSE in every time slot, then the data center's increased profit can grow up to 3287 dollars per month.

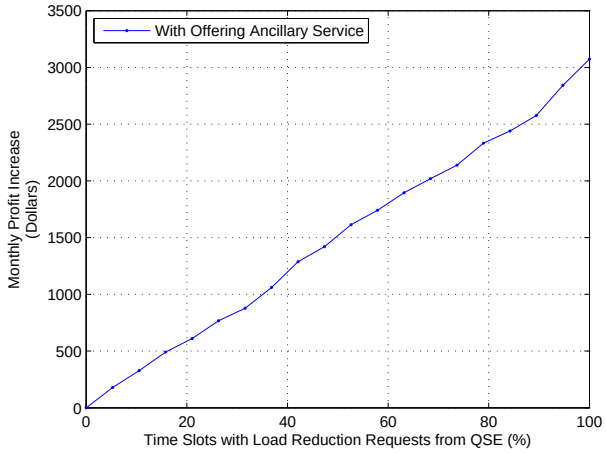

Figure 1.7: The monthly additional profit the data center gains versus the percentage of time slots in which the QSE sends out load reduction requests.

1.6 Conclusions and Future Work

This chapter represents the first step towards enabling Internet and cloud computing data centers to offer ancillary services to smart grid. We particularly focused on the scenario where a data center is registered as a load resource to offer voluntary load reduction ancillary service. We proposed an analytical profit maximization framework for the data center to set its service rate and the amount of load reduction service that it should offer. Our profit model includes elements with respect to the data center's Internet service revenue, the cost of electricity, and the compensation it may receive to offer ancillary services. The model takes into account server's power consumption profiles, data center's power usage effectiveness, price of electricity, workload statistics, and SLAs. Our simulation results show that data centers can benefit from offering voluntary load reduction by increasing their profit.

The results in this chapter can be extended in several directions. First, while our focus in this chapter is on voluntary load reduction, it is interesting to examine offering other forms of ancillary services such as frequency regulation. Second, other contract

models between the data center and the QSE can be considered. In particular, one can extend the analysis such that the data center can submit bids to the ancillary service market. Finally, in addition to adjusting the operation of each data center using the proposed optimization-based approach, a group of data centers can coordinate their operation to further increase their profit. In particular, the idea of migrating a portion of service workload from one data center to another which has already been studied in the literature for reducing data centers cost of electricity can be further adjusted to also be used for the purpose of offering ancillary services.

Chapter 2

Optimal Power Procurement for Data Centers in Day-Ahead and Real-Time Electricity Markets

2.1 Introduction

The energy demands of data centers have significantly increased over the past years. Accordingly, the cost of electricity to operate data centers have been skyrocketing. For example, it is estimated that Microsoft and Google each spent over $36 million on annual electricity bills for their data centers in 2007 [12]. The total annual electricity cost of servers and data centers in the United States is estimated at $7.4 billion [91].

The growing energy cost of data centers has motivated various studies to lower data centers' electricity bills. The prior work can be classified into at least five different classes. First, there have been studies to reduce the amount of power that

computing and memory devices consume, e.g., see [90, 42]. Second, different methods have been proposed to optimize the operation of hardware and software systems in data centers in response to changes in the workload, e.g., by conducting *dynamic cluster server configuration* [32, 53]. Third, there have been efforts to make the best use of *local energy recourses* at data centers, such as solar and wchapterind generators [77, 56], battery banks [17], and backup diesel generators [64]. Fourth, some recent studies have focused on *workload redistribution* across data centers to benefit from *geographical diversity* in both electricity prices [12, 102, 64] and renewable generation [101, 70]. Finally, there have been studies to manage the operation and available resources of data centers to better respond to the changes in the price of electricity, whether by lowering power consumption or by increasing the use of local energy resources. Examples include [69] for the case of time-of-use prices, [62] for the case of day-ahead prices when hedging is practiced, [100] for the case of coincidental peak prices, and [65] for the case of prediction-based prices.

In this chapter, our approach is related to the fifth class above. Our focus is on procuring power for data centers in a *deregulated electricity market*, i.e., a market where prices are set by running bidding mechanisms among the electricity suppliers and consumers, c.f., [72]. Compared to the prior work, our study is unique in the sense that we consider a setting where data centers can buy electricity from both the *day-ahead market* and the *real-time market*. The day-ahead market is usually settled several hours or even a day in advance while the real-time market is settled only one hour or sometimes 15 minutes in advance [96, Chapter 2].

Our goal is to understand the cost reductions data centers can achieve by exploiting the *diversity in the price across day-ahead and real-time markets*. To see the potential for such cost reductions, consider the sample empirical price data in Fig.

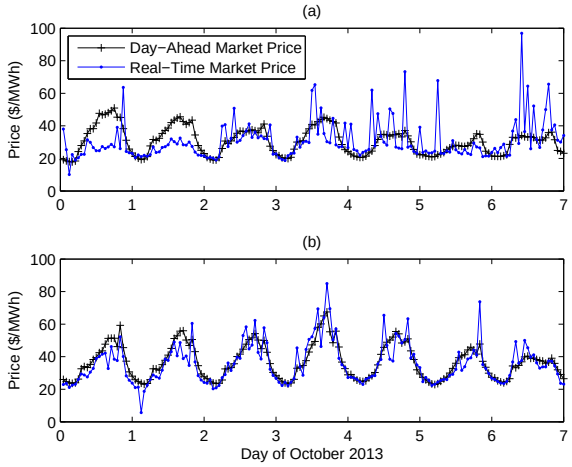

Figure 2.1: Two sample empirical price trends for day-ahead and real-time electricity markets during the first week of October 2013: (a) The Ameren retail price trends [1]. (b) The PJM wholesale market price trends [3].

2.1. Here, we have plotted the prices in the day-ahead market versus the real-time market in the Ameren retail market in Illinois [1] and the Pennsylvania, New Jersey, and Maryland (PJM) wholesale market [3]. Price diversity is evident in this figure. Based on the one-week sample price data in Fig. 2.1, and compared to purchasing electricity only from the day-ahead market, procuring electricity from *both* the day-ahead and real-time electricity markets may result in a saving in the cost of electricity up to 13.2% and 7.5% for the cases of the Ameren retail and PJM wholesale markets, respectively. An interesting observation in Fig. 2.1, which is consistent with other empirical price data that we have analyzed, is that while real-time prices are much more volatile, they are typically lower *on an average sense* than the day-ahead prices.

There are typically at least two challenges when it comes to procuring power directly (and jointly) from the day-ahead and real-time markets [34]. First, the uncertainty in the electricity price values. Second, the uncertainty and lack of accurate models with respect to the upcoming demand levels. The second item is particularly an issue for data centers due to the additional uncertainty that arises with respect to

their Internet and cloud-computing workload. To tackle these challenges, we propose a novel stochastic optimization approach that seeks to maximize the data center's profit, i.e., revenue minus cost, subject to the data center operator's risk management constraints as well as the constraints with respect to power consumption and service-level-agreements (SLAs), based on some elaborate queuing theoretic models.

Last but not least, it is worth pointing out that the large sizes of data centers make them *eligible* to directly participate in the day-ahead and real-time electricity markets that currently exist in the U.S., instead of purchasing electricity from regional utilities who charge "insurance premiums" to handle the variations in the wholesale price of electricity. For example, Microsoft's data center in Quincy, WA consumes over 48 megawatts of electricity which is enough to power 40,000 homes [78]. There are also data centers already operational or under construction that consume up to 100 megawatts [43].

2.2 System Model

2.2.1 Power Market and Cost of Electricity

In most deregulated electricity markets, electricity can be purchased both at the Day-Ahead Market (DAM) and the Real-Time Market (RTM). This is done by submitting demand bids L_{DAM} and L_{RTM} in megawatts to the day-ahead market and real-time market, respectively. The total amount of purchased power from the two markets combined is obtained as

$$\text{Power Purchase} = L_{DAM} + L_{RTM}. \tag{2.1}$$

The day-ahead market is usually settled several hours or even a day in advance while the real-time market is settled only one hour or even 15 minutes in advance [96, Chapter 2]. Let ω_{DAM} and ω_{RTM} denote the *market clearing prices* at the day-ahead and real-time markets, respectively. The total cost of power purchase for each bidding period is obtained as

$$\text{Cost} = L_{DAM}\omega_{DAM} + L_{RTM}\omega_{RTM}. \tag{2.2}$$

We note that, since the bidding process is done *before* the market is settled, the market clearing prices ω_{DAM} and ω_{RTM} are *not* known at the time of submitting the demand bids. Therefore, ω_{DAM} and ω_{RTM} are modeled two random variables with joint probability distribution function $f_{\omega_{DAM},\omega_{RTM}}(\cdot)$. Note that, in this study, we assume that data centers are price takers. That is, their demand bids are not large enough to have noticeable impact on the price of electricity.

2.2.2 Power Consumption

The total amount of power consumption in a data center is obtained by adding the total power consumption at the computer servers to the total power consumption at the facility, e.g., for cooling, lighting, etc. For a data center, *power usage effectiveness* (PUE), denoted by E_{usage}, is defined as the ratio of the data center's total power consumption to the power consumption at the servers [91]. The PUE values reported in the literature range from state-of-the-art 1.05 to 3.0 for the common practice [52, Section 12.3.3]. Let P_{server} denote the average power when a switched on server handles a service request. Also let $M \leq M_{max}$ denote the number of servers that are switched on at the data center. Assuming almost full CPU utilization for all switched

on servers, the total power consumption of a data center can be calculated as [95, 70]:

$$\text{Power Consumption} = E_{usage} M P_{server}, \tag{2.3}$$

Clearly, the power consumption at a data center increases as more servers are switched on to handle more service requests.

2.2.3 Quality-of-Service, SLAs, and Service Rate

Because of the limited computing capacity of data centers and the stochastic nature of workload, the service requests that are sent to a data center are first placed in a queue until they can be handled by an available computer. To satisfy quality-of-service (QoS) requirements, the waiting time / queuing delay for each incoming service request must be limited to a level that is determined by the *Service Level Agreement* (SLA). The exact SLA depends on the type of service offered which may range from cloud-based computational tasks to video streaming and web services. Two example SLAs based on the study in [28] are shown in Fig. 2.2, where each SLA is identified by three parameters D, δ, and γ. Parameter D indicates the maximum 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 *penalty* that the data center must pay to its customers every time it *cannot* handle a service request before deadline D. For the Gold SLA in Fig. 2.2, we have $D = 300$ ms, $\delta = 7 \times 10^{-5}$ dollars, and $\gamma = 3.5 \times 10^{-5}$ dollars. For the Silver SLA, we have $D = 200$ ms, $\delta = 5 \times 10^{-5}$ dollars, and $\gamma = 1.6 \times 10^{-5}$ dollars.

Let $\mu \geq 0$ denote the rate at which service requests are removed from the queue

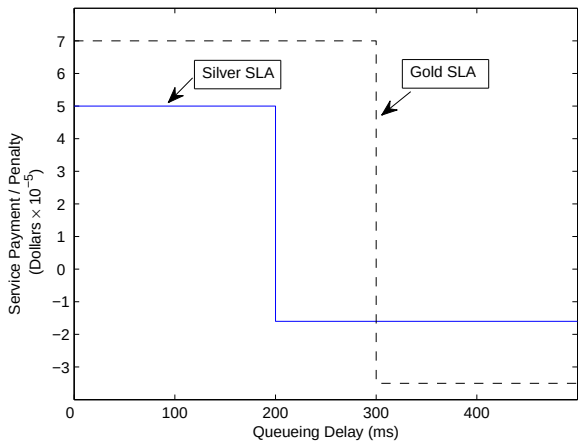

Figure 2.2: Two sample service-level agreements (SLAs) in data centers [28].

and handled by a computer server. The service rate μ depends both on the number of switched on servers M and the number of service requests κ that each server can handle per second. More specifically, we have [69]:

$$\mu = \kappa M \quad \rightarrow \quad M = \frac{\mu}{\kappa}. \tag{2.4}$$

As we switch on more servers and accordingly increase the service rate μ, more service requests can be handled *before* the SLA deadline D, which in turn increases the payments that the data center receives based on the SLAs. However, increasing μ will also increase the amount of power consumption at the data center. In fact, from (2.3) and (2.4), we have

$$\text{Power Consumption} = \phi\mu, \tag{2.5}$$

where

$$\phi = E_{usage} P_{server}/\kappa. \tag{2.6}$$

2.2.4 Revenue of Data Center

Consider a power purchase bidding period T, or any period of interest of length T for which the service rate μ is intended to be adjusted. Let $N \gg 1$ denote the number of service requests that arrive at the data center. Based on the SLA model that we discussed in Section 2.2.3, the revenue of the data center within the time period of interest can be calculated as

$$\text{Revenue} = \sum_{i=1}^{N} I_{D,i}\, \delta - (1 - I_{D,i})\, \gamma, \tag{2.7}$$

where $I_{D,i} = 1$ indicates that the i^{th} service request *was* handled before the SLA deadline D and $I_{D,i} = 0$ indicates that the i^{th} service request was *not* handled before the SLA deadline D. If T and N are large enough, then we can write

$$\text{Revenue} \approx \lambda T((1 - q(\mu))\delta - q(\mu)\gamma), \tag{2.8}$$

where λ denotes the average service arrival rate and $q(\mu)$ is the probability that a service request is *not* handled by the SLA deadline D. A model for $q(\mu)$ can be obtained through a G/D/1 queuing analysis that is already done in [68]:

$$q(\mu) = \begin{cases} q_I(\mu) & \mu_I \leq \mu \\ q'_{I+}(\mu_I)(\mu - \mu_I) + q(\mu_I) & \mu_{II} \leq \mu < \mu_I \\ q_{II}(\mu) & \mu < \mu_{II}, \end{cases} \tag{2.9}$$

where

$$q_I(\mu) = \alpha(\mu)\ \exp\left(-\frac{1}{2} \min_{n \geq 1}\ m_n(\mu)\right), \tag{2.10}$$

and

$$q_{II}(\mu) = \frac{T\lambda - T\mu}{T\lambda} = \frac{\lambda - \mu}{\lambda}. \tag{2.11}$$

Parameters μ_I and μ_{II} are obtained using [68, Algorithm 1] and the notations in (2.10) are defined as follows:

$$\alpha(\mu) = \frac{1}{\lambda\sqrt{2\pi}\sigma} e^{\frac{(\mu-\lambda)^2}{2\sigma^2}} \int_{\mu}^{\infty} (r-\mu) e^{-\frac{(r-\lambda)^2}{2\sigma^2}} dr, \tag{2.12}$$

and for each integer number $n \geq 1$ we have

$$m_n(\mu) = \frac{(D\mu + n(\mu-\lambda))^2}{n\sigma^2 + 2\sum_{l=1}^{n-1} \rho(l)(n-l)}. \tag{2.13}$$

It is worth emphasizing that the *general* service request arrival rate in (2.9)-(2.13) is modeled based on its various statistical characteristics, i.e., not only its mean λ, but also its variance σ^2 and its auto-covariance function $\rho(l)$, where lag time $l = 1, 2, \ldots$. Thus, the $q(\mu)$ model in (2.9) is significantly more elaborate and more accurate than the simplified M/M/1 queuing models that are typically used in most data center power consumption studies, e.g., in [101, 43, 62, 64].

2.3 Problem Formulation

From the results in Section 2.2, there is a *trade-off* when it comes to selecting a data center's service rate: increasing service rate increases the revenue while it also increases the cost. Addressing this trade-off is challenging due to the complexity of the queuing models and also because of the *stochastic nature* of the workload and

the DAM and RTP electricity prices. Hence, in this section, we propose a decision making process based on a stochastic optimization framework.

2.3.1 Stochastic Profit Maximization Problem

We can model the *profit* for a data center as

$$\text{Profit} = \text{Revenue} - \text{Cost}. \tag{2.14}$$

When it comes to operating a data center, it is natural to seek to maximize the data center's profit. However, due to the stochastic nature of workload and electricity price, such maximization must be in an average / statistical sense, i.e., in terms of the expected value of the profit. And of course it must be subject to the operator's risk management requirements. Therefore, we need to solve the following optimization problem to choose both the electricity purchase bidding parameters L_{DAM} and L_{RTM} as well as the service rate μ, one day in advance, i.e., at the time when the day-ahead market bid needs to be submitted:

$$\begin{aligned} \underset{\substack{L_{DAM}, L_{RTM} \\ \mu \leq \mu_{max}}}{\textbf{Maximize}} \quad & E\{\text{Profit}\} \\ \textbf{Subject to} \quad & Var\{\text{Profit}\} \leq \Gamma \\ & \text{Power Consumption} = \text{Power Purchase}, \end{aligned} \tag{2.15}$$

where $\mu_{max} = \kappa M_{max}$ and $\Gamma > 0$ is a design parameter. The choice of parameter Γ depends on whether the data center operator is *risk averse* (lower Γ) or *risk seeking* (higher Γ).

As we get closer to the actual operation time, the data center also needs to submit

its bid to the real-time market. One option is to submit the same solution of L_{RTM} that is obtained by solving (2.15) at the time of calculating and submitting the day-ahead market bid. Another option is to recalculate the profit with the *known* price of electricity in the day-ahead market, and solve (2.15) again with ω_{DAM} and L_{DAM} as constants. This will allow the data center to adjust its operation, once the exact realization of the day-ahead market price is known.

2.3.2 Mean and Variance of the Data Center Profit Function

In this section, we provide the exact mathematical models for the expected value and the variance of the profit. From (2.14) and using the definition of expected value, we have

$$E\{\text{Profit}\} = E\{\text{Revenue}\} - E\{\text{Cost}\}. \tag{2.16}$$

By substituting (2.2) and (2.8) in (2.16), we have

$$\begin{aligned} E\{\text{Profit}\} =& \lambda T((1-q(\mu))\delta - q(\mu)\gamma) - \\ & L_{DAM}E\{\omega_{DAM}\} - L_{RTM}E\{\omega_{RTM}\}. \end{aligned} \tag{2.17}$$

Next, from (2.14) and using the definition of variance, we have

$$\begin{aligned} Var\{\text{Profit}\} =& Var\{\text{Revenue}\} + Var\{\text{Cost}\} - \\ & 2Cov\{\text{Revenue}, \text{Cost}\} \end{aligned} \tag{2.18}$$

where the third term denotes the covariance of the revenue and cost. Since the randomness in the revenue function in (2.8) is solely due to the randomness of the workload and the randomness in the cost function in (2.2) is solely due to the randomness of the electricity price, and also because the price of electricity and the workload are

independent random variables, the covariance term in (2.18) is zero [11]. Therefore, in order to calculate the variance in (2.18), we need to only calculate the variance of the revenue and cost functions. However, as we will see in Section 2.4, numerical results show that the variations in cost due to the variations in the price of electricity are more significant compared to the variations in revenue due to the variations in workload. Therefore, we assume that the variance of the profit is approximated solely based on the variance of the cost. That is, we assume that

$$Var\{\text{Profit}\} \approx Var\{\text{Cost}\}. \tag{2.19}$$

From (2.2), the variance of the cost can be calculated as

$$\begin{aligned} Var\{\text{Cost}\} =& L_{DAM}^2 Var\{\omega_{DAM}\} + \\ & L_{RTM}^2 Var\{\omega_{RTM}\} + \\ & 2L_{DAM}L_{RTM}Cov\{\omega_{DAM}, \omega_{RTM}\}. \end{aligned} \tag{2.20}$$

2.3.3 A Convex Optimization Framework

We are now ready to complete the formulation of the stochastic profit maximization problem for a data center as

$$\begin{aligned} \underset{L_{DAM}, L_{RTM},\, \mu \le \mu_{max}}{\textbf{Maximize}} \quad & \lambda T\big((1-q(\mu))\delta - q(\mu)\gamma\big) - \\ & L_{DAM} E\{\omega_{DAM}\} - L_{RTM} E\{\omega_{RTM}\} \\ \textbf{Subject to} \quad & 2L_{DAM} L_{RTM} Cov\{\omega_{DAM}, \omega_{RTM}\} + \\ & L_{DAM}^2 Var\{\omega_{DAM}\} + \\ & L_{RTM}^2 Var\{\omega_{RTM}\} \le \Gamma \\ & L_{DAM} + L_{RTM} = \phi\mu, \end{aligned} \tag{2.21}$$

where $q(\mu)$ is as in (2.9). The following theorem shows that the above optimization problem is computationally tractable.

Theorem 1 *For any workload and electricity price parameters λ, σ^2, $\rho(l)$, $Cov\{\omega_{DAM}, \omega_{RTM}\}$, $Var\{\omega_{DAM}\}$, and $Var\{\omega_{RTM}\}$, the optimization problem in (2.21) is convex.*

The proof of Theorem 1 is given in the Appendix. From Theorem 1, problem (2.21) can be solved using standard convex programming techniques, c.f. [84]. Therefore, solving problem (2.21) can be considered as a practical yet optimal way to adjust the operation of the data center and to select its demand bids to the day-ahead and real-time electricity markets.

Before we end this section, we would like to point out a few remarks with respect to problem (2.21). First, the risk model based on variance that we used in this problem is only one option to cope with the uncertainty in electricity prices. Another

option is to use the value-at-risk (VaR) model from [64, 86, 27] and revise it for risk management across *both* day-ahead and real-time markets. Second, it might be beneficial to also use some price prediction methods, e.g., see [26, 45, 54], specially for the case of day-ahead market prices which are less volatile as we saw in Fig. 2.1. Finally, while we included both the power procurement and data center operation variables in problem (2.21), these variables could be set at different time intervals. For example, the studies in [68, 69] have shown that it is desirable to adjust the data center's service rate(s) every 15 minutes or less. However, some real-time markets may operate at longer intervals, e.g., every one hour[1]. Nonetheless, the optimization problem in (2.21) can still be used as long as we break down the revenue and power consumption terms into multiple terms, each corresponding to a smaller interval at which the service rate is adjusted.

2.4 Case Studies

Consider a data center with a large number of $M_{max} = 50,000$ servers. The number of switched on servers M, and accordingly the service rate μ is updated periodically at the beginning of each time slot of length $T = 15$ minutes. We assume that $\kappa = 0.1$. For each switched on server, we have $P_{server} = 150$ watts. The data center's power usage effectiveness is $E_{usage} = 1.5$. An SLA parameters are set based on the Gold service model in Fig. 2.2. To simulate the total workload, we use the World Cup 98 web hits data, spanning from June 10, 1998 to July 9, 1998 [2]. The electricity price information is based on the hourly day-ahead and real-time prices that are used by Ameren and PJM during December 2012 [1, 3]. In all cases, the revenue and cost are calculated using an event-based simulation, were an event is the

[1]In that case, they may also be referred to as *hour-ahead markets* [72].

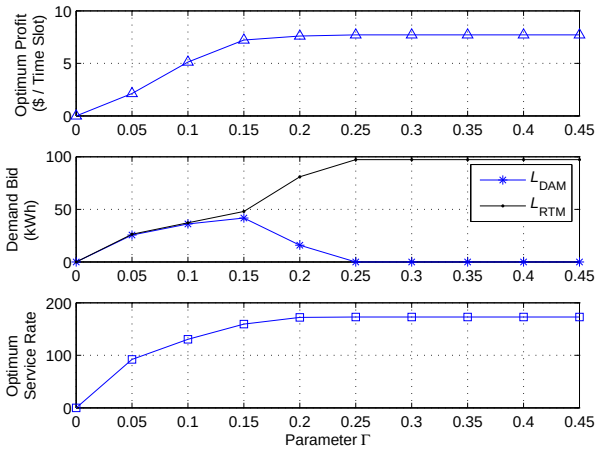

Figure 2.3: Numerical results for a single time slot for different values of Γ.

arrival of a new service request, c.f. [69].

To gain insights, we start by first looking at the detailed results for the case of solving problem (2.21) for a single time slot. Here, we use the prices from Ameren. The results are shown in Fig. 2.3 for different values of design parameter Γ. When $\Gamma = 0$, the only feasible solution is to have $\mu = L_{DAM} = L_{RAM} = 0$, i.e., shutting down the data center. As we increase Γ, looking at the scenarios where the center operator is more risk seeking, both service rate and profit increase. An interesting observation in this figure is that for the lower values of Γ, it is optimal to procure a large portion of the electricity needs from the day-ahead market as the prices in the day-ahead market are less volatile. However, as we increase Γ, such portion gradually disappears and the entire demands is eventually procured from the real-time market. This observation is along the line with our point in Section 2.1 that while real-time prices are much more volatile, they are typically lower compared to the day-ahead prices.

Next, we extend the above results to three different time slots / scenarios across

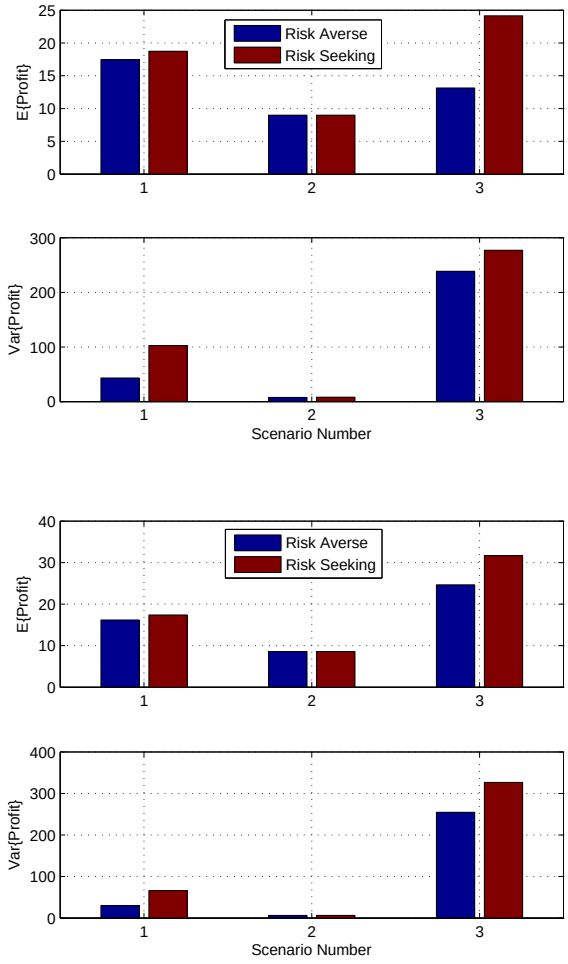

Figure 2.4: The expected value and the variance of profit for each hour that are calculated over one month: (a) Using Ameren prices. (b) Using PJM prices.

30 different days, i.e., over one month. The results are shown in Fig. 2.4(a) and (b), for the Ameren and PJM prices, respectively. Scenario 1 is for a time slot where the data center workload is at the *medium* level. Scenario 2 is for a time slot where the data center workload is at the *low* level. Scenario 3 is for a time slot where the data center workload is at the *high* level, making most servers busy. For each scenario, we compare a *risk averse* design with $\Gamma = 2$ and a *risk seeking* operation with $\Gamma = 4$. We can see that in all cases while a risk seeking design can increase

profit by more aggressively bidding in the real-time market, the higher profit comes at the cost of increasing the variance. Note that, expected values and variances are calculated across one month.

Finally, we look at the variance in the profit and compare it with the variance in cost. This is done in order to verify our assumption in (2.19). The results are shown in Fig. 2.5. Here, we are comparing $Var\{\text{Cost}\} > 0$ with $Var\{\text{Profit}\} > 0$. Note that, by definitions of variance and profit, we always have

$$0 \leq \frac{Var\{\text{Cost}\}}{Var\{\text{Profit}\}} \leq 1. \tag{2.22}$$

However, based on the results in Fig. 2.5, the above fraction is always very close to 1. This is due to the fact that if T and N are large enough, as in our case where $T = 15$ minutes and N is in the order of several thousand service requests per each time slot, the revenue in each time slot tends to have a very small variance because the summation in (2.8) tends to cancel out the randomness in each term inside the summation, as long as T and N are large. From the results in Fig. 2.5, we can conclude that the approximation in (2.19) is quite reasonable.

2.5 Conclusions

In this chapter, we took the first steps towards exploiting the diversity in the price of electricity across the day-ahead and real-time electricity markets to lower data centers' energy expenditure. Based upon our observations of some empirical service workload and electricity price data, we proposed a novel stochastic and provably convex profit maximization problem to select data centers' service rates and

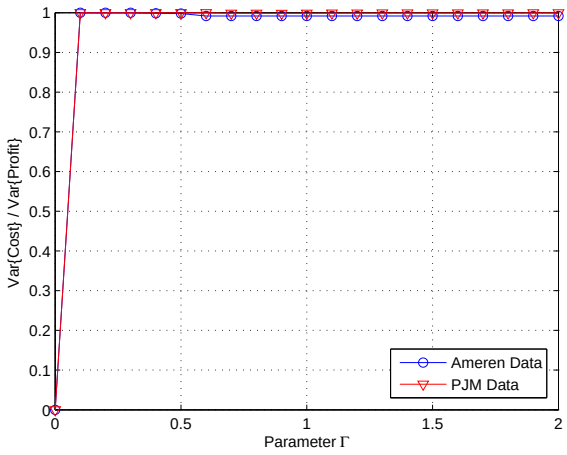

Figure 2.5: Numerical results to verify the approximation in (2.19). We can see that the variance of the profit is very close to the variance of cost.

demand bids to the day-ahead and real-time electricity markets. Our numerical results show that procuring power from the day-ahead and real-time electricity markets can significantly help data centers to lower their electricity bills; however, there is a trade-off between the expected value and variance as higher achieving profits requires more risk seeking operations.

2.5.1 Appendix

From [68, Theorems 1 and 3], for any choices of workload statistical parameters λ, σ^2, and $\rho(l)$, the probability model $q(\mu)$ is a convex and non-increasing function of service rate μ. Therefore, the objective function in (2.21) is concave with respect to μ and linear with respect to L_{DAM} and L_{RTM}. From this, and since the equality constraint in (2.21) is linear, problem (2.21) is a convex program as long as the left-hand side in the non-linear inequality constraint in (2.21) is a convex function of the optimization variables. To show this, first, we note that by definition of the

correlation coefficient, for any two random variables ω_{DAM} and ω_{RTM}, we have [11]:

$$-1 \leq \frac{Cov\{\omega_{DAM}, \omega_{RTM}\}}{Var\{\omega_{DAM}\}Var\{\omega_{RTM}\}} \leq 1. \tag{2.23}$$

From (2.23), we can further show that

$$Cov^2\{\omega_{DAM}, \omega_{RTM}\} \leq Var^2\{\omega_{DAM}\}Var^2\{\omega_{RTM}\}. \tag{2.24}$$

Next, we need to calculate the Hessian of the non-linear function in the inequality constraint in (2.21). It is obtained as

$$2\begin{bmatrix} Var\{\omega_{DAM}\} & Cov\{\omega_{DAM}, \omega_{RTM}\} & 0 \\ Cov\{\omega_{DAM}, \omega_{RTM} & Var\{\omega_{RTM}\} & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{2.25}$$

For the inequality constraint in (2.21) to be convex, the above matrix must be positive definite. Using the Schur complement and because the variance of electricity price is always positive and also because of the block diagonal structure of the matrix in (2.25), this matrix is positive definite if and only if

$$Var\{\omega_{RTM}\} - \frac{Cov^2\{\omega_{DAM}, \omega_{RTM}\}}{Var\{\omega_{DAM}\}} > 0. \tag{2.26}$$

However, the above inequality always holds due to (2.23). Therefore, the Hessian matrix in (2.25) is positive definite and accordingly the optimization problem in (2.21) is convex. ■

Chapter 3

A Convex Optimization Framework for Service Rate allocation in Finite Buffer Communications systems

3.1 Introduction

Loss Probability is commonly used as a metric to assess the performance of communication and networking systems that involve finite buffers [36, 85]. It is usually modeled as a function of service rate. A higher service rate results in a lower loss probability, which in turn improves quality-of-service (QoS). However, a higher service rate may increase the cost of service, e.g., due to using additional equipment or resources. Therefore, there is a *trade-off* between maximizing performance and minimizing cost in selecting service rate. This trade off can be systematically captured

within an optimization framework where the service rate is the optimization variable. However, a major concern is whether the problem is convex and tractable [66, 73, 25]. The convexity of loss probability models are previously studied in certain queueing systems, such as M/M/1/K queues, c.f. [79]. In this chapter, our focus is to address the unexplored problem of investigating the steady state behavior of finite G/D/1 buffers, where the arrival process has a *general* distribution and the queue is first-in first-out (FIFO), non-preemptive and non-process sharing.

3.2 Example Service Rate Allocation Problems

3.2.1 Case 1: Maximum Profit Multi-Service Scheduling

Consider a communications, networking, or computation system with $N \geq 1$ finite buffers to admit N different service types. For each service type $i = 1, \ldots, N$, the arrival rate is modeled using its mean λ_i, variance σ_i^2, and auto-covariance $\rho_i(l)$, where l is the lag-time. Let $\mu_i \geq 0$ denote the rate at which the service requests of type i are handled. For a time interval of length T, let $q_i(\mu)$ denote the loss probability at queue i. Let I_i denote the number of service requests of type i that are handled within this time interval of interest. We have

$$E\{I_i\} = T\lambda_i \left(1 - q_i(\mu_i)\right). \tag{3.1}$$

Next, let $R_i(\cdot)$ denote the revenue function that indicates the revenue the server receives as a function of the total number of handled service requests of type i, c.f. [36, 85], [15]. The *per-time interval* revenue is calculated as $R_i(T\lambda_i(1 - q_i(\mu_i)))$. Similarly, let $C_i(\mu_i)$ denote the cost that incurs to the server due to handling the

service requests of type i. The revenue and cost functions are assumed to be non-decreasing functions of their arguments. They are also concave and convex in their arguments, respectively. To maximize the total profit in the system, we need to solve the following optimization problem:

$$\underset{\mu_i \geq 0}{\textbf{Maximize}} \ \sum_{i=1}^{N} R_i(T\lambda_i(1 - q_i(\mu_i))) - \sum_{i=1}^{N} C_i(\mu_i) \tag{3.2}$$

Using the composition rules [84, p. 85], we can verify that the above problem is convex as long as the loss probability model $q_i(\mu_i)$ is convex in μ_i for every service type i.

3.2.2 Case 2: Stochastic Service Rate Optimization

In practice, there can be *uncertainties* even with respect to the exact statistical characteristics of the arrival process in communications systems, e.g., due to some *external factors*. For instance, consider a scenario where there is only $N = 1$ service type / service queue in the system and the statistical characteristics of the arrival process for the single service type of interest is represented by $\psi \triangleq <\lambda, \sigma, \rho(\cdot)>$ which belongs to a discrete set of outcomes Ψ with a probability mass function $f_\Psi(\cdot)$. Here, for the ease of presentation, we dropped subscript i from the mean, variance, and auto-correlation notations. The expected value of the profit is obtained as

$$E\{\text{Profit}\} = \sum_{\psi \in \Psi} [\text{Profit} \mid \psi] \, f_\Psi(\psi). \tag{3.3}$$

As an example, suppose $\Psi = \{\psi_1, \psi_2\}, f_\Psi(\psi_1) = \beta, \ f_\Psi(\psi_2) = (1 - \beta), \ \psi_1 = < \lambda_1, \sigma_1, \rho_1(\cdot)>$, and $\psi_2 = <\lambda_2, \sigma_2, \rho_2(\cdot)>$. In this case, the profit maximization problem

(3.3) becomes

$$\begin{aligned} \underset{0\leq\mu\leq\mu_{max}}{\textbf{Maximize}} \;\; & R(T\lambda_1(1-q_{\psi_1}(\mu))f_\Psi(\psi_1)+ \\ & R(T\lambda_2(1-q_{\psi_2}(\mu))f_\Psi(\psi_2)-C(\mu), \end{aligned} \tag{3.4}$$

where $q_{\psi_1}(\mu)$ and $q_{\psi_2}(\mu)$ denote the loss probability in the communications system of interest, when the arrival process carries statistical characteristics ψ_1 and ψ_2, respectively.

3.3 Loss Probability Model

In this section, we present a mathematical model for loss probability $q(\mu)$ to be used in problems similar to (3.2) and (3.4).

First, suppose $\mu \geq \lambda$. In [46], the following loss probability model of a G/D/1 queuing system was proposed for this case:

$$q_I(\mu) = \alpha(\mu)\; e^{-\frac{1}{2}\min\limits_{n\geq 1}\; m_n(\mu)}, \tag{3.5}$$

where

$$\alpha(\mu) = \frac{1}{\lambda\sqrt{2\pi}\sigma} e^{\frac{(\mu-\lambda)^2}{2\sigma^2}} \int_\mu^\infty (r-\mu) e^{-\frac{(r-\lambda)^2}{2\sigma^2}} dr \tag{3.6}$$

and for each integer number $n \geq 1$ we have

$$m_n(\mu) = \frac{(L+n(\mu-\lambda))^2}{n\sigma^2 + 2\sum_{l=1}^{n-1}\rho(l)(n-l)}. \tag{3.7}$$

Here, L denotes the size of the finite queue.

Theorem 2 *The loss probability function $q_I(\mu)$ in (3.5) is* not *convex over interval*

$\mu \in [\lambda, \lambda + 0.5\sigma]$*; however, it is* convex *and* non-increasing *over interval*

$$\mu \in [\lambda + 0.5\,\sigma, +\infty]\,. \tag{3.8}$$

The proof of Theorem 2 is given in Appendix A.

Next, suppose $\mu \leq \lambda$. This scenario may occur, e.g., in stochastic optimization, where service rate is less than the mean arrival rate under certain random scenarios. In this case, the server would always be busy. Accordingly, out of the total $T\lambda$ service requests that are received within the interval of length T, a total of $T\mu$ service requests are handled, while the rest, i.e., $T\lambda - T\mu$ service requests are dropped. Therefore, the loss probability can be approximated as

$$q_{II}(\mu) = \frac{T\lambda - T\mu}{T\lambda} = \frac{\lambda - \mu}{\lambda}. \tag{3.9}$$

As a special case, if $\mu \to 0$, then $q_{II}(\mu) \to 1$. Note that $q_{II}(\mu)$ is a linear (thus convex) and decreasing function of μ.

Using the empirical data in [2] with $T = 15$ minutes, the accuracy of the loss probability models in (3.5) and (3.9) are assessed in Fig. 3.1. We can see that $q_I(\mu)$ in (3.5) is accurate when $\mu \to \infty$ and $q_{II}(\mu)$ in (3.9) is accurate when $\mu \to 0$. However, both models lose accuracy when μ approaches λ, from the right hand side in case of $q_I(\mu)$, and from the left hand side in case of $q_{II}(\mu)$. Therefore, we propose to *adjust* and *combine* the loss probability models (3.5) and (3.9) and obtain the following

alternative loss probability model:

$$q(\mu) = \begin{cases} q_I(\mu) & \mu_I \leq \mu \\ q'_{I+}(\mu_I)(\mu - \mu_I) + q_I(\mu_I) & \mu_{II} \leq \mu \leq \mu_I \\ q_{II}(\mu) & \mu < \mu_{II}, \end{cases} \tag{3.10}$$

where $q'_{I+}(\mu_I)$ denotes the *right derivative* of function $q_I(\mu)$ at $\mu = \mu_I$. The point $\mu_I \geq \lambda + 0.5\sigma$ is chosen in a way that, the right tangent to $q_I(\mu)$ at μ_I intersects $q_{II}(\mu)$ at a point μ_{II} such that $\lambda - \sigma \leq \mu_{II} \leq \lambda$. The proof on the guaranteed existence of parameters μ_I and μ_{II} is omitted for brevity.

Theorem 3 *The loss probability function $q(\mu)$ in (3.10) is convex in service rate for its entire operation range $\mu \geq 0$.*

The above theorem directly results from Theorem 2 and the way that the loss probability function $q(\mu)$ is constructed.

To examine the accuracy of the proposed loss probability model, next we generate 100 random time series of length $T = 15$ minutes [30, Section 5.1], based on the statistical characteristics of randomly selected 15 minutes intervals of the data in [2]. Fig. 3.2 shows the mean absolute error (sorted in ascending order) of the proposed loss probability model in (3.10) and the one in (3.5) over the interval $[\lambda, \mu_I]$ for these 100 time series. From Fig. 3.2, the loss probability model in (3.10) has a lower mean absolute error than the one in (3.5).

Theorem 4 *Let us define n_{max} such that $\rho(l) = 0$ for any $l \geq n_{max}$. If $L \geq \sigma n_{max}$, there exist a parameter $\mu^\star \geq \mu_{II}$ such that the proposed loss probability model in (3.10) is a more accurate approximation of the true loss probability than the model in (3.5) over the interval $\mu^\star \leq \mu \leq \mu_I$.*

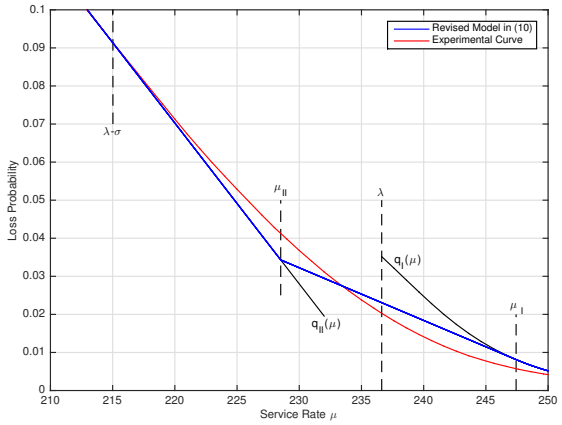

Figure 3.1: Loss probability as a function of service rate μ: empirical curve versus the three analytical curves according to (3.5), (3.9), and (3.10).

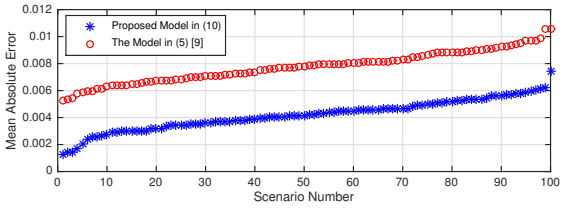

Figure 3.2: The mean absolute error of the proposed loss probability model in (10) and that of the one in (5) over 100 randomly generated time series.

The proof of Theorem 4 is given in Appendix B.

3.4 Case Studies

First, consider Case 1 in Section 3.2.1. Here, we simulate 30 time slots of length $T = 15$ minutes. We assume that $N = 3$ different types of service requests are handled by the shared server. The service request arrival rates for the first, the second, and the third service types are set based on the World Cup data on June 14th, 15th, and 16th, respectively, from 12:00 AM to 7:30 AM [2]. We set $R_i(x) = 100w_i log(1 + x)$,

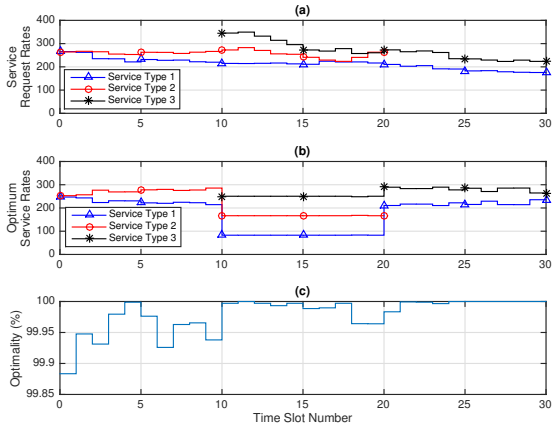

Figure 3.3: Simulation results for Case 1: (a) The mean service request arrival rates. (b) The service rates by solving (3.2). (c) The optimality in comparison with the true optimal profit obtained from an event-based simulation.

where $w_1 = 1$, $w_2 = 2$, and $w_3 = 3$ are the *revenue weighting factors*. The cost functions are fixed. Hence, problem (3.2) reduces to a multi-service queue revenue maximization problem.

Simulation results are shown in Fig. 3.3, where the operating time is divided into *three* time frames. First, during time slots 1 to 10, there are service requests for service types 1 and 2. After that, during time slots 11 to 20, service requests arrive from all three service types. Finally, during time slots 21 to 30, the server receives requests for service types 1 and 3, but not 2. From the results in Fig. 3.3, a service rate allocation based on the optimal solution of problem (2) manages the resources based on the priority of incoming service requests, giving higher service rates to service requests with higher priorities.

Next, consider Case 2 in Section 3.2.2. Think of a web-based video streaming server for a playoff soccer game. Suppose the 90 minutes normal game time is about to finish while the game is in a tie. The server administrators need to allocate resources

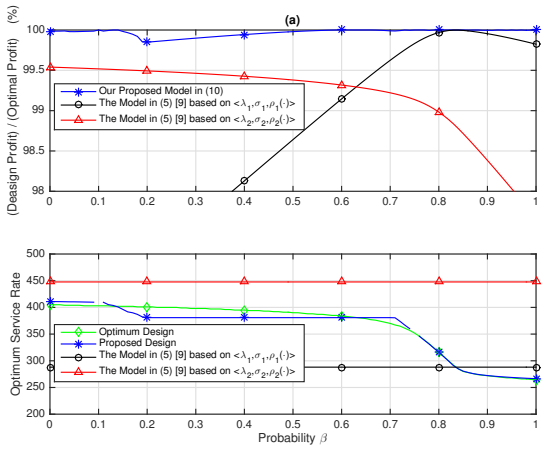

Figure 3.4: Simulation results for Case 2: (a) The optimality in maximizing the expected profit. (b) Optimal service rate based on different design approaches.

for the next $T = 15$ minutes. There are *two possibilities*. First, with probability β, one team scores and the game ends in normal time. In that case, the workload to the servers *will drop significantly* as many online viewers do not watch the post-game show. Second, with probability $1-\beta$, the game ends in tie and the *extra time* is implemented. In this case, the workload will remain high as users will continue watching the game. In this example, the outcome of the soccer game is the *external factor* in Section 3.2.2. It is natural to assume that the World Cup server administrators have a good estimate, based on historical data, about the statistical characteristics of the workload during a *post-game show streaming*, i.e., ψ_1, and during a *live game streaming*, i.e., ψ_2. However, they do *not* know which one of these two scenarios will occur. Therefore, they need to solve a stochastic optimization problem as in (3.4). Here, we use the statistical characteristics of the workload data [2] during time slot 4:15 AM to 4:30 AM on June 29th to obtain ψ_1 and the statistical characteristics of the workload data during time slot 2:00 AM to 2:15 AM on June 24th to obtain ψ_2. Note that, we have $\lambda_1 = 248.48$ and $\lambda_2 = 394.77$.

The *expected* profit for the next $T = 15$ minutes are shown in Fig. 3.4(a), where

the probability parameter β varies from 0 to 1. Here, the expected value is calculated by examining 100 random workloads that are generated based on the statistical characteristics that we obtained from the empirical data, using the *time series generation* scheme in [30, Section 5.1]. We can see that, our proposed stochastic optimization approach can accurately maximize the expected profit for all values of β. In contrast, the model in (3.5) [46] cannot be used for stochastic optimization. Instead, we should either use $<\lambda_1, \sigma_1, \rho_1(\cdot)>$ or $<\lambda_2, \sigma_2, \rho_2(\cdot)>$; either way, the results will be represented by flat curves as in Fig. 3.4(b). For certain values of β, e.g., $\beta = 0.9$, the optimal service rate μ^* is greater than λ_1 but less than λ_2. Therefore, it is necessary to use a loss probability model that works for both $\mu \geq \lambda$ and $\mu \leq \lambda$, as in (3.10).

3.5 Conclusions

A convex optimization framework was proposed for service rate allocation in finite communications buffers with example applications to maximum profit multi-service scheduling and stochastic service rate allocation. Using empirical data, both deterministic and stochastic case studies were investigated.

Appendix A: Proof of Theorem 2

Let us define $t \triangleq (\mu - \lambda)/\sigma$. We can reformulate (3.6) as

$$\alpha(\mu) = \frac{\sigma}{\lambda\sqrt{2\pi}} \left(1 - te^{\frac{t^2}{2}} \int_t^\infty e^{-\frac{u^2}{2}} du\right). \tag{3.11}$$

Once we take the derivative with respect to μ, we have

$$\alpha'(\mu) = \frac{1}{\lambda\sqrt{2\pi}} \left(t - (t^2+1)e^{\frac{t^2}{2}} \int_t^{\infty} e^{-\frac{u^2}{2}} du \right). \tag{3.12}$$

Also, since e^{-x} is non-increasing we have $q_I(\mu) = \max_n \ q_n(\mu)$, where for each $n \geq 1$, we define

$$q_n(\mu) \triangleq \alpha(\mu)\ e^{-\frac{1}{2}m_n(\mu)}. \tag{3.13}$$

From [84, Section 3.2.3], $q_I(\mu)$ is proven to be a convex function if we can show that for each $n \geq 1$, we have

$$\begin{aligned} {q_n}''(\mu) = e^{-\frac{1}{2}m_n(\mu)} \Bigg(& \alpha''(\mu) + \alpha(\mu){m_n'}^2(\mu)/4 \\ & - \alpha'(\mu)m_n'(\mu) - \alpha(\mu)m_n''(\mu)/2 \Bigg) \geq 0. \end{aligned} \tag{3.14}$$

We show (3.14) through the following five steps:

Step 1: Since the function under integral in (3.6) is positive within the range of integral, from (3.6) we have $\alpha(\mu) \geq 0$.

Step 2: We show that $\alpha''(\mu) \geq 0$ over interval (3.8). After taking the second derivative of α with respect to μ, we have

$$\alpha''(\mu) = \frac{(t^3+3t)e^{\frac{t^2}{2}}}{\lambda\sqrt{2\pi}\sigma} \left(\frac{t^2+2}{t^3+3t} e^{\frac{-t^2}{2}} - \int_t^{\infty} e^{-\frac{u^2}{2}} du \right). \tag{3.15}$$

Next, we note that

$$\frac{d}{dt}\left(\frac{t^2+2}{t^3+3t} e^{-\frac{t^2}{2}} - \int_t^{+\infty} e^{-\frac{u^2}{2}} du \right) = \frac{-6e^{-\frac{t^2}{2}}}{t^2(t^2+3)^2} < 0, \tag{3.16}$$

and

$$\lim_{t\to+\infty}\left(\frac{t^2+2}{t^3+3t}e^{-\frac{t^2}{2}}-\int_t^{+\infty}e^{-\frac{u^2}{2}}du\right)=0. \tag{3.17}$$

From (3.16) and (3.17), we conclude that for each $t>0$, we have

$$\left(\frac{t^2+2}{t^3+3t}e^{-\frac{t^2}{2}}-\int_t^{+\infty}e^{-\frac{u^2}{2}}du\right)\geq 0. \tag{3.18}$$

Since $t>0$ for the interval in (3.8), from (3.18), we have $\alpha''\geq 0$.

Step 3: We show that

$$-2\alpha'(\mu)/\alpha(\mu)\geq 1/(\sigma t). \tag{3.19}$$

In other words, from (3.11) and (3.12), we need to show that

$$-2\frac{\alpha'(\mu)}{\alpha(\mu)}-\frac{1}{\sigma t}=\frac{1}{\sigma}\left(\frac{2e^{\frac{t^2}{2}}\int_t^{\infty}e^{-\frac{u^2}{2}}du}{1-te^{\frac{t^2}{2}}\int_t^{\infty}e^{-\frac{u^2}{2}}du}-2t-\frac{1}{t}\right)\geq 0.$$

After reordering the terms, we can rewrite this inequality as

$$\frac{2t^2+1}{2t^3+3t}e^{-\frac{t^2}{2}}-\int_t^{\infty}e^{-\frac{u^2}{2}}du\leq 0. \tag{3.20}$$

Next, we note that

$$\frac{d}{dt}\left(\frac{2t^2+1}{2t^3+3t}e^{-\frac{t^2}{2}}-\int_t^{\infty}e^{-\frac{u^2}{2}}du\right)=\frac{(6t^2-3)e^{-\frac{t^2}{2}}}{(2t^2+3)^2\,t^2}. \tag{3.21}$$

The above derivative is negative (indicating a decreasing function) for any $0<t<\sqrt{2}/2$ and positive (indicating an increasing function) for any $t>\sqrt{2}/2$. Furthermore, we have

$$\lim_{t\to+\infty}\left(\frac{2t^2+1}{2t^3+3t}e^{-\frac{t^2}{2}}-\int_t^{\infty}e^{-\frac{u^2}{2}}du\right)=0. \tag{3.22}$$

The function on the left hand side in (3.20) has a zero at $t = 0.466$ and a minimum at $t = \sqrt{2}/2$. From these, together with (3.21) and (3.22), we conclude that (3.20) and consequently (3.19) hold as long as $t \geq 0.466$, e.g., within the interval in (3.8).

Step 4: We show that, over interval (3.8), we have

$$m_n''(\mu)/m_n'(\mu) \leq 1/(\sigma t). \tag{3.23}$$

From (3.7) and after taking the derivatives over μ, we have:

$$\frac{m_n''(\mu)}{m_n'(\mu)} = \frac{n}{L + n(\mu - \lambda)} \leq \frac{1}{(\mu - \lambda)} = \frac{1}{\sigma t}. \tag{3.24}$$

Step 5: From (3.19) and (3.23), and since $m_n'(\mu) \geq 0$ [46], we have

$$\begin{aligned} & m_n''(\mu)/m_n'(\mu) \leq -2\alpha'(\mu)/\alpha(\mu) \\ \Rightarrow \quad & -\alpha'(\mu)m_n'(\mu) - \alpha(\mu)m_n''(\mu)/2 \geq 0. \end{aligned} \tag{3.25}$$

Also, from Steps 1 and 2, over interval (3.8) we have

$$\alpha''(\mu) + \alpha(\mu)\, {m_n'}^2(\mu)/4 \geq 0. \tag{3.26}$$

From (3.25) and (3.26) and since the exponential function is non-negative, we can conclude (3.14) and the proof is complete. ■

Appendix B: Proof of Theorem 4

Since $\rho(l) \leq \sigma^2$ for any l, we have $\sum_{l=1}^{n-1} \rho(l)(n-l) \leq \sigma^2 n(n-1)/2$. From this, together with (3.7), we can show that

$$\frac{{m'_n}^2}{m''_n} = 2\,\frac{(L+n(\mu-\lambda))^2}{n\sigma^2 + 2\sum_{l=1}^{n-1}\rho(l)(n-l)} \geq 2\left(\frac{L}{n\sigma}+t\right)^2. \tag{3.27}$$

From (3.27), if $L \geq \sigma n_{max}$ then

$$\alpha(\mu){m'_n}^2/4 - \alpha(\mu)m''_n/2 \geq 0. \tag{3.28}$$

Since $\alpha''(\mu)$ and $-\alpha'(\mu)m'_n$ are non-negative, from (3.14) and (3.28), we conclude that $q_n(\mu)$ in (3.13) is a convex function of μ for any $L \geq \sigma n_{max}$. Accordingly, from [84, Section 3.2.3], $q_I(\mu)$ is convex if $L \geq \sigma n_{max}$. Next, let $\tilde{q}(\mu)$ denote the *true* loss probability at service rate μ. Since $q_I(\mu)$ is an *upper bound* for $\tilde{q}(\mu)$ [46], we have $q(\mu_I) = q_I(\mu_I) \geq \tilde{q}(\mu_I)$. Also, since (3.9) is a *lower bound* for $\tilde{q}(\mu)$ [47], we have $q(\mu_{II}) = q_{II}(\mu_{II}) \leq \tilde{q}(\mu_{II})$. From these two facts, and since $\tilde{q}(\mu)$ is a continuous function, $\tilde{q}(\mu)$ must intersect with line segment $q(\mu)$ at a point $\mu_{II} \leq \mu^\star \leq \mu_I$. Consequently, since $\tilde{q}(\mu)$ is convex [57, Theorem 3.1], we conclude that

$$q(\mu) \geq \tilde{q}(\mu) \quad \forall \mu \in [\mu^\star, \mu_I]. \tag{3.29}$$

Also, from the convexity of $q_I(\mu)$, we have

$$q_I(\mu) \geq q(\mu) \qquad \forall \mu \geq \lambda. \tag{3.30}$$

From (3.29) and (3.30), the proposed loss probability model in (3.10) is a *tighter upper bound* for $\tilde{q}(\mu)$ than the model in (3.5) for any service rate $\mu^\star \leq \mu \leq \mu_I$ when $L \geq \sigma n_{max}$. ■

Chapter 4

Energy Portfolio Optimization of Data Centers

4.1 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 [92] or a load serving entity [65]. It may also participate in wholesale electricity markets, including the day-ahead market (DAM) and real-time market (RTM) [71, 13]. Another option for data centers is to enroll in ancillary service (ANS) programs [38, 94, 23]. Data centers may also fully or partially operate by local renewable (REN) power generators such as wind turbines [70] and/or solar panels [48]. Some data centers also use on-site energy storage systems (ESS) [98]. 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 [61, 101].

The above options are summarized in Table I. The last two columns indicate

Table 4.1: Summary of Representative Related Literature

	RET	DAM	RTM	ANS	REN	ESS	GWD	SLA	RM
[71]	✗	✓	✓	✗	✗	✗	✗	✓	✓
[94]	✓	✗	✓	✓	✗	✗	✓	✗	✓
[23]	✓	✗	✗	✓	✗	✗	✗	✗	✗
[98]	✓	✗	✗	✗	✓	✓	✓	✗	✗
[64]	✗	✓	✓	✗	✗	✗	✓	✗	✓
[76]	✗	✓	✓	✗	✗	✗	✓	✗	✓
[62]	✓	✗	✓	✗	✗	✗	✓	✗	✓
[14]	✓	✗	✗	✓	✗	✓	✗	✗	✗
[81]	✗	✗	✗	✗	✗	✓	✗	✗	✗
[93]	✓	✗	✓	✗	✓	✓	✗	✗	✗
[22]	✗	✗	✗	✗	✗	✗	✗	✓	✗
[24]	✗	✗	✗	✗	✗	✗	✗	✓	✗
[21]	✗	✗	✗	✗	✓	✗	✗	✓	✗
[97]	✓	✗	✗	✗	✓	✓	✗	✗	✗
[103]	✓	✗	✗	✗	✓	✗	✓	✗	✗

whether the study takes into consideration service-level agreements (SLAs) [28] or risk management (RM) [64, 76, 62]. 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.

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 chapter, where we develop an *energy portfolio optimization framework* for data centers. The contributions in this chapter 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 chapter is comparable also with the literature on energy portfolio management in contexts *other than data centers*, e.g., see [55, 58]. 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.

4.2 Energy Management Options

4.2.1 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.

4.2.2 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 [72]. 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 [71].

4.2.3 Local Renewable Generation

Depending on their locations, data centers can use various on-site renewable generation options, such as wind turbines [70] and/or solar panels [48]. 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.

4.2.4 Offering Ancillary Services

Traditionally, ancillary services are offered by generators [72, Chapter 9]. However, large consumers, such as data centers, are also eligible to register as *load resources* to offer ancillary services [87, 38]. In this chapter, our focus is on a data center that offers *spinning reserve* [60]. 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 [19, 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 [35].

4.2.5 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 $E_{\mathrm{STR}}[t]$. We must always have

$$0 \leq E_{\mathrm{STR}}[t] \leq E_{\mathrm{STR}}^{max}, \tag{4.1}$$

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 [70].

4.2.6 Geographic Workload Distribution

As it is recently shown, e.g., in [61, 101, 70, 43, 44], 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 chapter, 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.

4.3 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 4.2. 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 4.3.9.

4.3.1 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 [51]. 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 $M[t] \leq M^{\max}$ 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

$$\mu[t] = M[t]\kappa. \tag{4.2}$$

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.

4.3.2 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 [28]. 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.

4.3.3 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 [91]. 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 [95, 70]:

$$\text{Power Consumption} = \phi\, \mu[t], \tag{4.3}$$

where $\phi = E_{\text{usage}} P_{\text{server}}/\kappa$ and the equality is due to (4.2).

4.3.4 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 $\omega_{\mathrm{DAM}}[t]$, $\omega_{\mathrm{RTM}}[t]$, $G_{\mathrm{REN}}[t]$, $\omega_{\mathrm{ANS}}[t]$, $Y_{\mathrm{ANS}}[t]$ and $\omega_{\mathrm{CAL}}[t]$ 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 [82]. For each scenario $k = 1, \ldots, K$, we denote the realizations of the random variables as $\omega^k_{\mathrm{DAM}}[t]$, $\omega^k_{\mathrm{RTM}}[t]$, $G^k_{\mathrm{REN}}[t]$, $\omega^k_{\mathrm{ANS}}[t]$, $Y^k_{\mathrm{ANS}}[t]$ and $\omega^k_{\mathrm{CAL}}[t]$. Recall that the retail electricity price $\omega_{\mathrm{RET}}[t]$ 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 $L^k_{\mathrm{RTM}}[t]$.

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

$$L_{\mathrm{RET}}[t] + L_{\mathrm{DAM}}[t] + (1 - I_{\mathrm{ANS}}[t] Y^k_{\mathrm{ANS}}[t]) L^k_{\mathrm{RTM}}[t], \tag{4.4}$$

where

$$I_{\mathrm{ANS}}[t] = \mathbb{I}\left(L_{\mathrm{ANS}}[t] > 0\right). \tag{4.5}$$

Here, $\mathbb{I}(\cdot)$ is a 0-1 indicator function. If $L_{\mathrm{ANS}}[t] = 0$, then $I_{\mathrm{ANS}}[t] = 0$. If $L_{\mathrm{ANS}}[t] > 0$, then $I_{\mathrm{ANS}}[t] = 1$. To understand the last term in (4.4), recall from Section 4.2.4 that if the data center offers reserve service, i.e., $I_{\mathrm{ANS}}[t] = 1$, and it receives a reserve call signal under scenario k, i.e., $Y^k_{\mathrm{ANS}}[t] = 1$, 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

$$\begin{aligned} & L_{\mathrm{RET}}[t]\omega_{\mathrm{RET}}[t] + L_{\mathrm{DAM}}[t]\omega^k_{\mathrm{DAM}}[t] \\ & + (1 - I_{\mathrm{ANS}}[t]Y^k_{\mathrm{ANS}}[t])L^k_{\mathrm{RTM}}[t]\omega^k_{\mathrm{RTM}}[t]. \end{aligned} \tag{4.6}$$

4.3.5 Service Rate Allocation

From (4.3) and (4.4), at each random scenario k and each time slot t, the following *power balance* equation must hold:

$$\begin{aligned} & L_{\mathrm{RET}}[t] + L_{\mathrm{DAM}}[t] + (1 - I_{\mathrm{ANS}}[t]Y^k_{\mathrm{ANS}}[t])L^k_{\mathrm{RTM}}[t] \\ & \quad = \phi\, \mu^k[t] - G^k_{\mathrm{REN}}[t] + (E_{\mathrm{STR}}[t] - E_{\mathrm{STR}}[t-1])/\tau, \end{aligned} \tag{4.7}$$

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 (4.7) incorporate the impact of local renewable generator and energy storage unit, respectively. We can rewrite (4.7) as

$$\begin{aligned} \mu^k[t] = \frac{1}{\phi}\Big[& L_{\mathrm{RET}}[t] + L_{\mathrm{DAM}}[t] \\ & + (1 - I_{\mathrm{ANS}}[t]Y^k_{\mathrm{ANS}}[t])L^k_{\mathrm{RTM}}[t] \\ & + G^k_{\mathrm{REN}}[t] - (E_{\mathrm{STR}}[t] - E_{\mathrm{STR}}[t-1])/\tau\Big]. \end{aligned} \tag{4.8}$$

4.3.6 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

$$\tau\lambda[t](\delta - (\delta+\gamma)q(\mu^k[t])) \tag{4.9}$$

and

$$L_{\text{ANS}}[t]\omega^k_{\text{ANS}}[t] + L_{\text{ANS}}[t]Y^k_{\text{ANS}}[t]\omega^k_{\text{CAL}}, \tag{4.10}$$

respectively. First, we explain (4.9). 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 [69], we have

$$q(\mu) = \alpha(\mu)\ \exp\left(-\frac{1}{2}\min_{n\geq 1}\ m_n(\mu)\right), \tag{4.11}$$

where

$$\alpha(\mu) = \frac{1}{\lambda\sqrt{2\pi}\sigma}e^{\frac{(\mu-\lambda)^2}{2\sigma^2}}\int_{\mu}^{\infty}(r-\mu)e^{-\frac{(r-\lambda)^2}{2\sigma^2}}dr, \tag{4.12}$$

$$m_n(\mu) = \frac{(D\mu + n(\mu-\lambda))^2}{n\sigma^2 + 2\sum_{l=1}^{n-1}\rho_l[t](n-l)}, \quad \forall n \geq 1. \tag{4.13}$$

The above model is based on the assumption that service rate is higher than average service request arrival rate. An extension of (4.11) when this assumption is relaxed is given in [68]. As for the model in (4.10), the first term is the reserve capacity payment and the second term is the reserve call payment.

4.3.7 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:

$$\Psi = \left\{ \sum_{t=1}^{T} \text{Profit}^k[t] \;\middle|\; 1 \leq k \leq K \right\} \tag{4.14}$$

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 4.3.8. Note that, from the discussions in Section 4.3.4, 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:

$$\begin{gathered} \text{Average of } \beta \text{ Fractile Lowest Total Profit Values} \geq \Gamma \\ \Longleftrightarrow \\ \text{Average of } \beta \text{ Fractile Lowest Elements in } \Psi \geq \Gamma, \end{gathered} \tag{4.15}$$

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 [67, 50, 99, 31]. 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 (4.15), we first sort the elements in set Ψ in an ascending order to obtain the following set:

$$\overline{\Psi} = \text{Sort } (\Psi), \qquad \overline{\Psi}^1 \leq \cdots \leq \overline{\Psi}^K. \tag{4.16}$$

From (4.16), the constraint in (4.15) is equivalent to

$$\sum_{k=1}^{\beta K} \frac{\overline{\Psi}^k}{\beta K} \geq \Gamma. \tag{4.17}$$

Next, we note that, from [83, Definition 3], we have

$$\begin{aligned} \text{CVaR}_{1-\beta}\left(-\sum_{t=1}^{T} \text{Profit}[t]\right) &= \sum_{k=\beta K}^{1} \frac{-\overline{\Psi}^k}{\beta K} \\ &= \sum_{k=1}^{\beta K} \frac{-\overline{\Psi}^k}{\beta K} \\ &= \sum_{k=1}^{K} \frac{-\overline{\Psi}^k}{\beta K} + \sum_{k=\beta K}^{K} \frac{\overline{\Psi}^k}{\beta K} \\ &= -\sum_{k=1}^{\beta K} \frac{\overline{\Psi}^k}{\beta K}, \end{aligned} \tag{4.18}$$

where CVaR denotes the standard operator for *conditional value at risk* [83, 80].

From (4.17) and (4.18), we can express the risk restrain constraint in (4.15) as

$$-\text{CVaR}_{1-\beta}\left(-\sum_{t=1}^{T}\text{Profit}[t]\right) \geq \Gamma. \tag{4.19}$$

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 (4.19) *do not* cancel out each other.

4.3.8 Risk-aware Profit Maximization Problem

From the expressions in (4.6), (4.9), and (4.10), the data center's profit at time slot t under scenario k is calculated as

$$\begin{aligned}\text{Profit}^k[t] = \tau\lambda[t](\delta - (\delta+\gamma)q(\mu^k[t])) \\ + L_{\text{ANS}}[t]\omega^k_{\text{ANS}}[t] \\ + L_{\text{ANS}}[t]Y^k_{\text{ANS}}[t]\omega^k_{\text{CAL}} \\ - L_{\text{RET}}[t]\omega_{\text{RET}}[t] - L_{\text{DAM}}[t]\omega^k_{\text{DAM}}[t] \\ - (1 - I_{\text{ANS}}[t]Y^k_{\text{ANS}}[t])L^k_{\text{RTM}}[t]\omega^k_{\text{RTM}}[t].\end{aligned} \tag{4.20}$$

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

$$\begin{aligned}\textbf{max} \quad & \sum_{t=1}^{T}\frac{1}{K}\sum_{k=1}^{K}\text{Profit}^k[t] \\ \textbf{s.t.} \quad & \text{Eqs. } (4.1), (4.5), (4.8), (4.11), (4.20), t = 1, \ldots, T \\ & -\text{CVaR}_{1-\beta}\left(-\sum_{t=1}^{T}\text{Profit}[t]\right) \geq \Gamma,\end{aligned} \tag{4.21}$$

From [83, Theorem 16], the last constraint in (4.21) can be reformulated and equivalently expressed as

$$\begin{aligned}
&\sum_{t=1}^{T} \text{Profit}^k[t] + \zeta + \eta_k \geq 0, \qquad && k = 1, \ldots, K, \\
&\eta_k \geq 0, && k = 1, \ldots, K, \\
&\zeta + \frac{1}{\beta}\frac{1}{K}\sum_{k=1}^{K} \eta_k \leq -\Gamma,
\end{aligned} \tag{4.22}$$

where ζ and η_k for all $k = 1, \ldots, K$ are auxiliary variables. By replacing the last constraint in optimization problem (4.21) with the set of inequalities in (4.22), the optimization problem (4.21) can be equivalently expressed as

$$\begin{aligned}
\textbf{max} \quad & \sum_{t=1}^{T} \frac{1}{K} \sum_{k=1}^{K} \text{Profit}^k[t] \\
\textbf{s.t.} \quad & \text{Eqs. } (4.1), (4.5), (4.8), (4.11), (4.20), \quad t = 1, \ldots, T \\
& \sum_{t=1}^{T} \text{Profit}^k[t] + \zeta + \eta_k \geq 0, && k = 1, \ldots, K, \\
& \eta_k \geq 0, && k = 1, \ldots, K, \\
& \zeta + \frac{1}{\beta}\frac{1}{K}\sum_{k=1}^{K} \eta_k \leq -\Gamma.
\end{aligned} \tag{4.23}$$

By solving optimization problem (4.23), 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 4.2.

4.3.9 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 $L_{i,\mathrm{RET}}$, $L_{i,\mathrm{DAM}}$, $L_{i,\mathrm{RTM}}$, $G_{i,\mathrm{REN}}$, $I_{i,\mathrm{ANS}}$, $L_{i,\mathrm{ANS}}$, $Y_{i,\mathrm{ANS}}$ and $E_{i,\mathrm{STR}}$ 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 $\omega_{i,\mathrm{RET}}[t]$, $\omega_{i,\mathrm{DAM}}[t]$ and $\omega_{i,\mathrm{RTM}}[t]$ respectively. Also, the amount of available renewable generation at the location of ith data center within time slot t is denoted by $G_{i,\mathrm{REN}}[t]$. Moreover, $Y_{i,\mathrm{ANS}}[t] \in \{0,1\}$ indicates whether a reserve capacity call signal is received at ith data center within time slot t. For the time slot t, $Y_{i,\mathrm{ANS}}[t] = 1$ means a reserve capacity call signal is received by the ith data center, while $Y_{i,\mathrm{ANS}}[t] = 0$ 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 $\omega_{i,\mathrm{ANS}}$ and $\omega_{i,\mathrm{CAL}}$ respectively.

Let $I_{i,\mathrm{ANS}}[t] \in \{0,1\}$ denote whether data center i participates in the reserve market at time slot t. Specifically, for each time slot t, $I_{i,\mathrm{ANS}}[t] = 1$ means that data center i does participate in the reserve market, while $I_{i,\mathrm{ANS}}[t] = 0$ means that data center i does not participate in the reserve market. Similar to the discussion in Section 4.3.4 we have

$$I_{i,\mathrm{ANS}}[t] = \mathbb{I}\left(L_{i,\mathrm{ANS}}[t] > 0\right) \tag{4.24}$$

where, $\mathbb{I}(\cdot)$ is defined in Section 4.3.4.

At each time slot t, we assume that the statistical characteristics of random vari-

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

$$\sum_{i=1}^{N} \lambda^k_i[t] = \lambda[t] \qquad k = 1, \ldots, K. \tag{4.25}$$

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 [39, 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

$$\sigma^k_i[t]^2 = \left(\frac{\lambda^k_i[t]}{\lambda[t]}\right)^2 \sigma^2[t], \qquad \rho^k_{i,l}[t] = \left(\frac{\lambda^k_i[t]}{\lambda[t]}\right)^2 \rho_l[t]. \tag{4.26}$$

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 4.3.5, we have

$$\begin{aligned}\mu_i^k[t] = \frac{1}{\phi}\Big[& L_{i,\mathrm{RET}}[t] + L_{i,\mathrm{DAM}}[t] \\ & + (1 - I_{i,\mathrm{ANS}}[t] Y_{i,\mathrm{ANS}}^k[t]) L_{i,\mathrm{RTM}}^k[t] \\ & + G_{i,\mathrm{REN}}^k[t] - (E_{i,\mathrm{STR}}[t] - E_{i,\mathrm{STR}}[t-1])/\tau \Big].\end{aligned} \tag{4.27}$$

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 4.3.6 and 4.3.8, the total profit of the data centers under scenario k is obtained as

$$\sum_{i=1}^{N} \sum_{t=1}^{T} \mathrm{Profit}_i^k[t], \tag{4.28}$$

where

$$\begin{aligned}\mathrm{Profit}_i^k[t] = & \tau \lambda_i^k[t] (\delta - (\delta + \gamma) q_i(\mu_i^k[t]), \lambda_i^k[t]) - \xi_i \lambda_i^k[t] \\ & + L_{i,\mathrm{ANS}}[t] \omega_{i,\mathrm{ANS}}^k[t] \\ & + L_{i,\mathrm{ANS}}[t] Y_{i,\mathrm{ANS}}^k[t] \omega_{i,\mathrm{CAL}}^k \\ & - L_{i,\mathrm{RET}}[t] \omega_{i,\mathrm{RET}}[t] - L_{i,\mathrm{DAM}}[t] \omega_{i,\mathrm{DAM}}^k[t] \\ & - (1 - I_{i,\mathrm{ANS}}[t] Y_{i,\mathrm{ANS}}^k[t]) L_{i,\mathrm{RTM}}^k[t] \omega_{i,\mathrm{RTM}}^k[t].\end{aligned} \tag{4.29}$$

and

$$q_i(\mu_i^k[t], \lambda_i^k[t]) = \alpha(\mu_i^k[t], \lambda_i^k[t]) \exp\left(-\frac{1}{2} \min_{n \geq 1} \; m_n(\mu_i^k[t]) \right). \tag{4.30}$$

As in (4.12) and (4.13), we have

$$\alpha(\mu_i^k[t], \lambda_i^k[t]) = \left(exp\left(\frac{(\mu_i^k[t] - \lambda_i^k[t])^2}{2\sigma_i^k[t]^2} \right) \Big/ \lambda_i^k[t]\sqrt{2\pi}\sigma_i^k[t] \right) \int_{\mu_i^k[t]}^{\infty} (r - \mu_i^k[t]) e^{-\frac{(r-\lambda_i^k[t])^2}{2\sigma_i^k[t]^2}} dr, \tag{4.31}$$

and

$$m_n(\mu_i^k[t]) = \frac{(D\mu_i^k[t] + n(\mu_i^k[t] - \lambda_i^k[t]))^2}{n\sigma_i^k[t]^2 + 2\sum_{l=1}^{n-1} \rho_{i,l}^k[t](n-l)}, \quad \forall n \geq 1. \tag{4.32}$$

Similar to the discussion in Section 4.3.7, the average total profit over β fractile lowest profit random scenarios is kept above a threshould Γ, if the following inequality holds:

$$-\text{CVaR}_{1-\beta}\left(-\sum_{i=1}^{N}\sum_{t=1}^{T} \text{Profit}_i^k[t] \right) \geq \Gamma. \tag{4.33}$$

We seek to maximize the *aggregated* expected profit of *all* data centers. Different from the single data center case in Section 4.3.8, here, $\lambda_i^k[t]$ for $i = 1, \ldots, N$ is an optimization variable. The following risk-aware energy portfolio optimization problem gives the optimum operation variables of data centers:

$$\begin{aligned} \textbf{max} \quad & \sum_{t=1}^{T} \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{N} \text{Profit}_i^k[t] \\ \textbf{s.t.} \quad & \text{Eqs. } (4.24), (4.25), (4.27), (4.29), (4.30) \qquad \forall t, i, k, \\ & -\text{CVaR}_{1-\beta}\left(-\sum_{i=1}^{N}\sum_{t=1}^{T} \text{Profit}_i^k[t] \right) \geq \Gamma. \end{aligned} \tag{4.34}$$

Similar to the discussion in Section 4.3.8, from [83, 89], the optimization problem (4.34) can be equivalently expressed as

$$
\begin{aligned}
\textbf{max} \quad & \sum_{t=1}^{T} \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{N} \text{Profit}_i^k[t] \\
\textbf{s.t.} \quad & \text{Eqs. (4.24), (4.25), (4.27), (4.29), (4.30)} && \forall t, i, k, \\
& \sum_{t=1}^{T} \sum_{i=1}^{N} \text{Profit}_i^k[t] + \zeta + \eta_k \geq 0, && \forall k, \\
& \eta_k \geq 0, && \forall k, \\
& \zeta + \frac{1}{\beta} \frac{1}{K} \sum_{k=1}^{K} \eta_k \leq -\Gamma.
\end{aligned}
\tag{4.35}
$$

If $N = 1$, then problem (4.35) reduces to problem (4.23).

4.4 Solution Method

Problem (4.35) is a mixed-integer *nonlinear* program, which is hard to solve. Notice that, even if we relax the binary constraints in (4.24), problem (4.35) 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 (4.27) and (4.29). In this Section, we first propose a solution approach based on combining convex programming with the branch-and-bound method [41]. This approach is *guaranteed* to give the optimal solution of the problem in (4.35). After that, we will also propose an approximate solution for the problem in (4.35) which is based on mixed integer linear programming (MILP) and can be solved efficiently, e.g., using CPLEX [4].

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 (4.27) and (4.29) by introducing the following inequality as a new

constraint to the problem in (4.35):

$$0 \leq L^k_{i,\text{RTM}}[t] \leq (1 - I_{i,\text{ANS}}[t]Y^k_{i,\text{ANS}}[t])(\kappa M^{\max}\phi), \tag{4.36}$$

where $\kappa M^{\max}\phi$ is the maximum value that $L^k_{i,\text{RTM}}[t]$ can take. To see this, we note that from (4.36), if $I_{i,\text{ANS}}[t]Y_{i,\text{ANS}} = 1$, then $L^k_{i,\text{RTM}}[t]$ is forced to zero. Hence, the value of $L^k_{i,\text{RTM}}[t]$ is the same as that of $(1 - I_{i,\text{ANS}}[t]Y^k_{i,\text{ANS}}[t])L^k_{i,\text{RTM}}[t]$, 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 $1 - I_{i,\text{ANS}}[t]Y^k_{i,\text{ANS}}[t] = 1$ and the value of $L^k_{i,\text{RTM}}[t]$ is again the same as that of $(1 - I_{i,\text{ANS}}[t]Y^k_{i,\text{ANS}}[t])L^k_{i,\text{RTM}}[t]$. After replacing $(1 - I_{i,\text{ANS}}[t]Y^k_{i,\text{ANS}}[t])L^k_{i,\text{RTM}}[t]$ in (4.27) and (4.29) with $L^k_{i,\text{RTM}}[t]$, and adding (4.36) as a new constraint to (4.35), the following optimization problem is obtained which is equivalent to (4.35):

$$\begin{aligned}
\textbf{max} \quad & \sum_{t=1}^{T} \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{N} \text{Profit}^k_i[t] \\
\textbf{s.t.} \quad & \text{Eqs. (4.25), (4.30)} && \forall t, i, k, \\
& \sum_{t=1}^{T} \sum_{i=1}^{N} \text{Profit}^k_i[t] + \zeta + \eta_k \geq 0, && \forall k, \\
& \eta_k \geq 0, && \forall k, \\
& \zeta + \frac{1}{\beta}\frac{1}{K}\sum_{k=1}^{K} \eta_k \leq -\Gamma \\
& 0 \leq L^k_{i,\text{RTM}}[t] \leq (1 - I_{i,\text{ANS}}[t]Y^k_{i,\text{ANS}}[t])\kappa M^{\max}\phi,
\end{aligned} \tag{4.37}$$

where

$$\begin{aligned}\text{Profit}_i^k[t] = \tau\lambda_i^k[t](\delta - (\delta+\gamma)q_i(\mu_i^k[t]), \lambda_i^k[t]) - \xi_i\lambda_i^k[t] \\ + L_{i,\text{ANS}}[t]\omega_{i,\text{ANS}}^k[t] + L_{i,\text{ANS}}[t]Y_{i,\text{ANS}}^k[t]\omega_{i,\text{CAL}}^k \\ - L_{i,\text{RET}}[t]\omega_{i,\text{RET}}[t] - L_{i,\text{DAM}}[t]\omega_{i,\text{DAM}}^k[t] \\ - L_{i,\text{RTM}}^k[t]\omega_{i,\text{RTM}}^k[t],\end{aligned} \tag{4.38}$$

and

$$\begin{aligned}\mu_i^k[t] = \frac{1}{\phi}\Big[L_{i,\text{RET}}[t] + L_{i,\text{DAM}}[t] + L_{i,\text{RTM}}^k[t] \\ + G_{i,\text{REN}}^k[t] \\ - (E_{i,\text{STR}}[t] - E_{i,\text{STR}}[t-1])/\tau\Big].\end{aligned} \tag{4.39}$$

Note that, from [68, Theorem 2], $q_i(\mu_i^k[t], \lambda_i^k[t])$ in (4.30) is a convex function of $\mu_i^k[t]$. Also, from (4.26), $q_i(\mu_i^k[t], \lambda_i^k[t])$ in (4.30) 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 [10, 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 (4.38) is convex and therefore the optimization problem (4.37) is a mixed-integer convex program. It can be solved with *guaranteed optimality* using convex programming and branch-and-bound method [41].

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. [74, Section 10.4], [88, Section 13.5], and [16, 104, 59, 63, 40, 18]. Similarly, in this chapter, we replace $\lambda_i^k[t]q_i(\mu_i^k[t], \lambda_i^k[t])$

in (4.38) with its *two-dimensional piece-wise outer-linearized approximation* [20]:

$$z_i^k[t] = \max_p \{A_{p,i}[t]\mu_i^k[t] + B_{p,i}[t]\lambda_i^k[t] + C_{p,i}[t]\}. \tag{4.40}$$

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 [29, Proposition 6.4.1], we have:

$$\lambda_i^k[t]q_i(\mu_i^k[t], \lambda_i^k[t]) = \lim_{P\to\infty} z_i^k[t]. \tag{4.41}$$

From (4.38), minimizing the objective function in (4.37) involves minimizing $z_i^k[t]$; accordingly, we can replace (4.40) with

$$z_i^k[t] \geq \{A_{p,i}[t]\mu_i^k[t] + B_{p,i}[t]\lambda_i^k[t] + C_{p,i}[t]\} \quad \forall p. \tag{4.42}$$

After substituting the term $\lambda_i^k[t]q_i(\mu_i^k[t], \lambda_i^k[t])$ in (4.38) with $z_i^k[t]$ and adding the constraint in (4.42) to the problem (4.37), the problem (4.37) becomes a mixed integer linear program and can be solved with existing software such as CPLEX and MOSEK.

4.5 Case Studies

4.5.1 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 [69], 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 [3]. The data for the reserve capacity call signal is from PJM, based on its historical synchronized reserve events [5]. The data for reserve capacity price is from PJM [6]. We set $L_{\mathrm{CAL}} = L_{\mathrm{RTM}}$ [37]. The PJM datasets are from January 1, 2004 to January 30, 2004. The data for wind speed is from [7], and the wind turbine power-versus-wind-speed curve is from [8]. The statistical data of the workload is from the web hits of Wikipedia on 9/19/2007 [9]. 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 [68], which is an extension of the model in (4.11) to the entire range of service rate.

4.5.2 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. 4.1. 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 $L_{\mathrm{ANS}}\omega_{\mathrm{RTM}}$ 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.

Next, we compare a risk averse design with $\Gamma = 50$ and a risk seeking design with

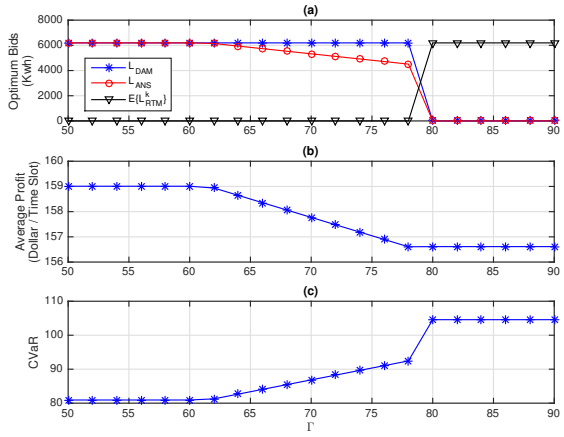

Figure 4.1: The impact of risk management parameter Γ: (a) optimal day-ahead energy and reserve market bids, (b) expected profit, (c) CVaR of profit.

$\Gamma = 80$. The results are shown in Fig. 4.2, 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.

4.5.3 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. 4.3. The profit increases as we increase the 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

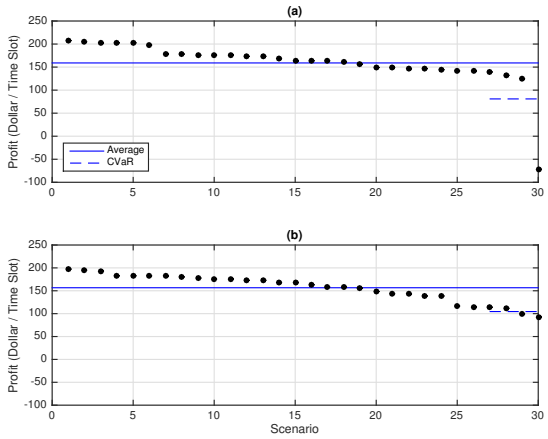

Figure 4.2: The profit values over 30 scenarios for a design that is: (a) risk seeking, (b) risk averse. The profits are sorted in a descending order.

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.

4.5.4 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. 4.4 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.

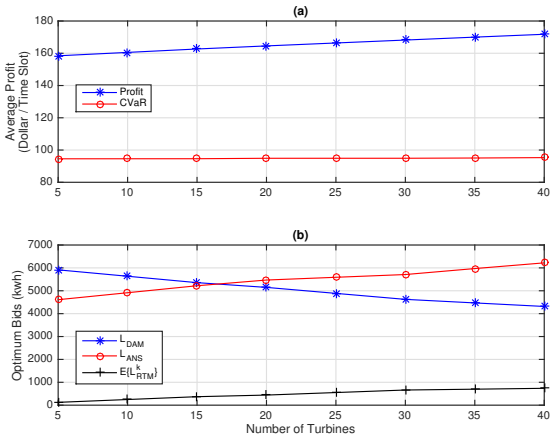

Figure 4.3: Operation with renewable generators based on the number of wind turbines: (a) average profit, (b) optimal day-ahead energy and reserve bids.

4.5.5 Impact of SLA Parameters

The optimum bids and the resulted expected profit for different ratios of SLA parameters δ/γ are shown in Fig. 4.5, where $\Gamma = 60$, γ is fixed and δ changes. From the results in Fig. 4.5(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. 4.5(b), when $\delta/\gamma \geq 3$, the optimum reserve bid is zero and

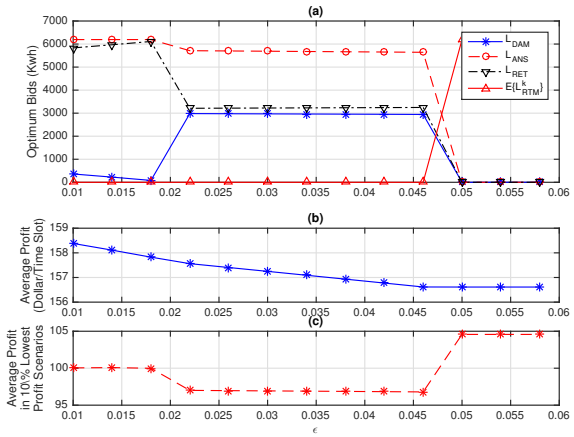

Figure 4.4: The impact of power purchase from Retail Market (a) optimum bids, (b) average profit and (c) Average of profit in 10% lowest profit scenarios

power is purchased from the real-time market, not the day-ahead market. From the results in Fig. 4.5(c), the average profit increases as the SLA parameter δ increases. Finally, from the results in Fig. 4.5(d), the CVaR is always kept above Γ.

4.5.6 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. 4.6. In single-time-slot energy portfolio management, problem (4.23) is solved T times for T time slots, where $\Gamma = \Gamma_{ST} = 80$. In contrast, under multiple-time-slots energy portfolio management, problem (4.23) 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 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-

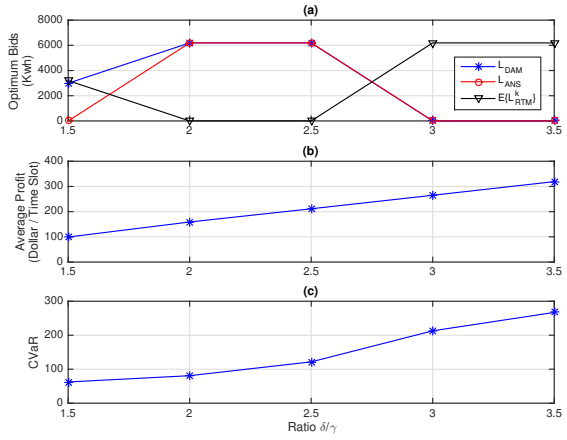

Figure 4.5: The impact of changing SLA parameters on the data center operation: (a) optimal bids, (b) average L^k_{RTM} and (c) Average of profit and (d) CVaR are shown for different ratios of δ/γ, where γ is fixed and δ is changing.

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.

4.5.7 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 [8]. 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.

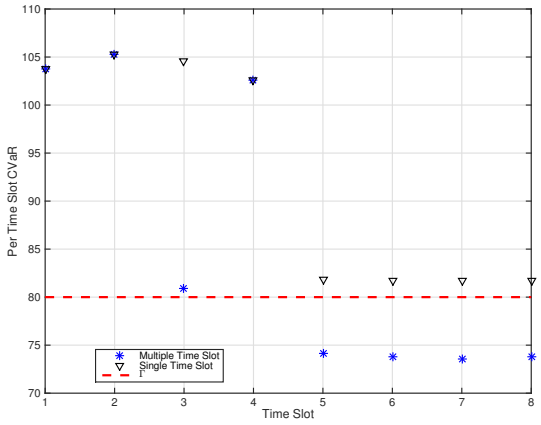

Figure 4.6: Per-time-slot CVaR for profit based on two designs: risk management on a single time slot; and joint risk management across multiple time slots.

Here, the data center is allowed to submit reserve bids at time slots 28, 29, 30, 31, 38, 50, 63, 72, 73, 75 [5].

Fig. 4.7 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. 4.8 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. 4.7(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 4.2.4. 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.

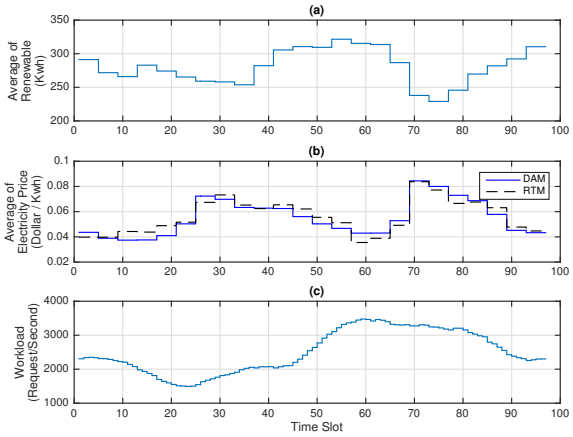

Figure 4.7: The system parameters for the case study in Sections 4.5.7 and 4.5.12: (a) the average local renewable power generation, (b) the average day-ahead and real-time market prices, (c) Internet workload.

4.5.8 Geographical Workload Distribution

Consider two geographically distributed data centers. The first one is equipped with 50 wind turbines of the type in [8]. It can also purchase electricity from a retailer at price $(1 + .01)E\{\omega_{\text{DAM}}\}$, where the data for ω_{DAM} is from [3]. The second data center can purchase electricity directly from the day-ahead and real-time wholesale markets [3]. It can also bid in the reserve market. Here, we have $\Gamma = 75$ and $\xi_i = 0$, for all $i = 1, \ldots, N$. Fig. 4.9 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. 4.9(c) shows the optimum bids and Fig. 4.9(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.

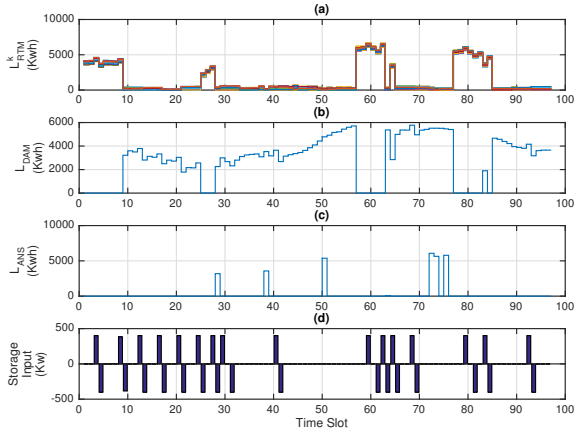

Figure 4.8: The optimal operation results for the case study in Section 4.5.7: (a) optimum bids to real-time market, (b) the optimal bids to the day-ahead market, (c) the optimum bid to reserve market, (d) the optimal charge and discharge schedule of the energy storage unit.

4.5.9 Impact of Communication Cost

Fig. 4.10 shows the optimum total profit of data centers over a time slot of length $T = 15$ minutes, and the optimum 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. 4.10, 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.

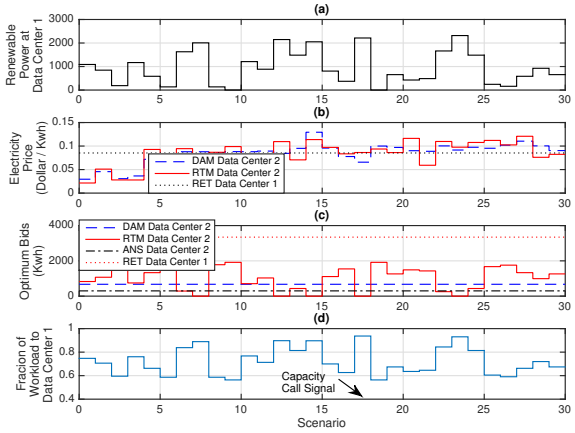

Figure 4.9: Two coordinated data centers: (a) renewable generation at data center 1; (b) electricity prices; (c) optimum bids; (d) optimum workload distribution.

4.5.10 Comparison to other Profit Maximization Models

In this Section, we compare the performance of the proposed profit maximization models with the ones in [61] and [101] for the case of a data center that purchases electricity from both day-ahead and real-time markets. Fig. 4.11 shows that the proposed model in (4.23) gives a higher profit for all time slots during a 24 hours time interval. Specially, in time slots 25 to 28, the model in (4.23) significantly outperforms the models in [61] and [101], as the model in (4.23) 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 [61] and [101] 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 (4.23), [61], and [101] procure electricity from the day-ahead market. However, our approach in (4.23) still outperforms the models in [61] and [101] in time slots 29 to 32, as the model in (4.23) takes into account the SLA, while the other two models do not.

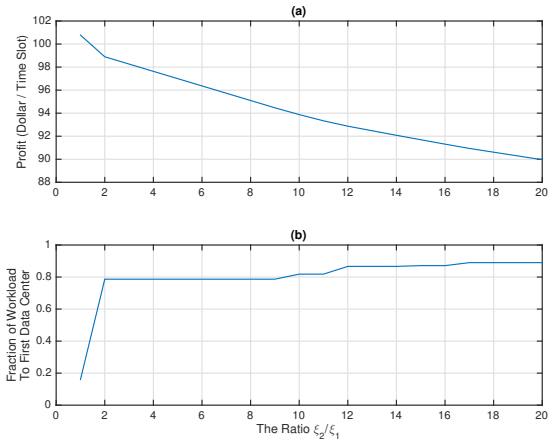

Figure 4.10: The impact of communication cost on geographical workload distribution with two coordinated data centers: (a) the optimum profit; (b) optimum fraction of workload that is forwarded to the first data center

4.5.11 Computational Time and Optimality of Proposed Solution

Again consider the simulation setup in Section 4.5.7. In this section, we examine the impact of changing the number of linearization segments P on the computation time and the optimization accuracy. Fig. 4.12(a) shows that the profit that is obtained from (4.23) increases as the number of segments increases, but it is saturated when the number of segments reaches 25. Also, Fig. 4.12(b) shows that the computation time in solving problem (4.23) has an increasing trend when the number of segments increases. In overall, from Figs. (4.12)(a) and (b), one can achieve reasonable optimality and computational time by using the optimization formulation in (4.23). 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.

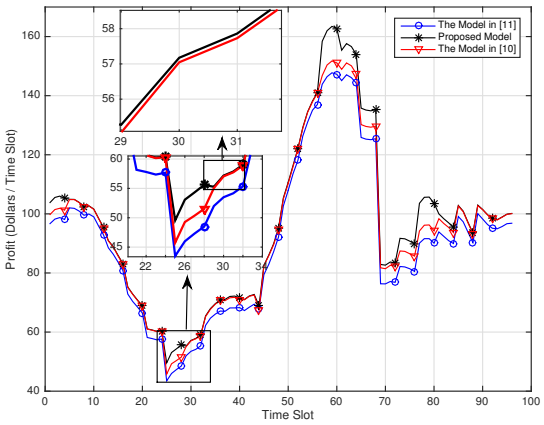

Figure 4.11: The profit of data center over one single time slot for our proposed profit maximization design as well as for the designs in [101] and [61].

4.5.12 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 (4.23) gives the optimum operating variables for this setup, if we add the following constraint to the problem in (4.23):

$$\mu^k[1] = \cdots = \mu^k[96], \qquad \forall k \leq K. \tag{4.43}$$

Notice that a whole day constitutes of 96 time slots of length $T = 15$ minutes, and therefore the constraint in (4.43) indicates that the service rate is fixed over one whole day and under the kth scenario. Fig. 4.13 shows the optimal operating variables, for the simulation setup in Fig. 4.7, when the constraint (4.43) is added to (4.23). Here, the optimum reserve market bid is obtained as zero over all time slots. Also, from Fig. 4.13(c), the optimum service rates at the realization of each random scenario is the same over all time slots.

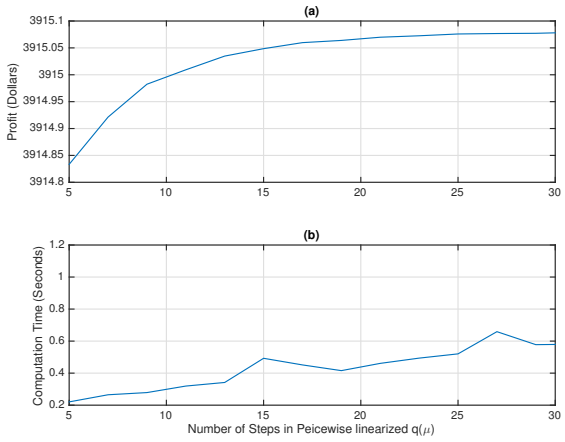

Figure 4.12: The impact of number of line segments on the results on the case study in Section 4.5.7: (a) the optimum profit; and (b) computational time.

4.6 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.

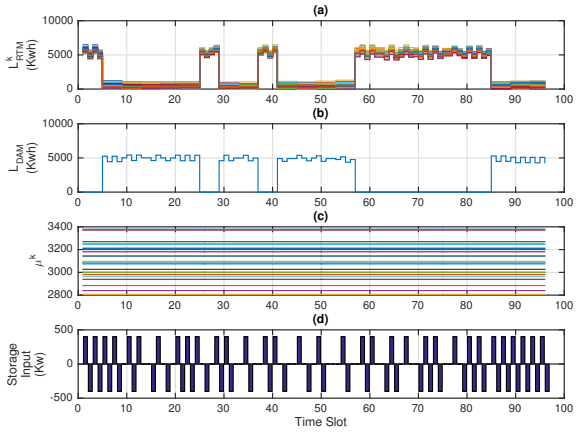

Figure 4.13: The optimal operation results for the case study in Section 4.5.12: (a) the optimal bids to real-time market, (b) the optimal bids to the day-ahead market, (c) The optimum service rates, (d) the optimal charge and discharge schedule of the energy storage unit.

Bibliography

[1] https://www2.ameren.com/RetailEnergy/realtimeprices.aspx.

[2] Http://ita.ee.lbl.gov/html/contrib/WorldCup.html.

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

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

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

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

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

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

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

[10] A. Harel, "Convexity Properties of the Erlang Loss Formula", *Operation Research*, vol. 38, no. 3, pp. 499–505, May 1990.

[11] A. Leon-Garcia, *Probability, Statistics, and Random Processes For Electrical Engineering*, 3rd, Prentice Hall, Jan. 2011.

[12] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the Electric Bill for Internet-Scale Systems", in: *Proc. of the ACM SIGCOMM*, Barcelona, Spain, 2009.

[13] 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, Nov. 2014.

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

[15] L. Ast, T. Cinkler, G. Fodor, S. Racz, and S. Blaabjerg, "Blocking probability approximations and revenue optimization in multi-rate loss networks", *Modelling and Simulation of Computer Systems and Networks*, vol. 68, no. 1, pp. 56–65, Jan. 1997.

[16] 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.

[17] B. Aksanli, T. Rosing, and E. Pettis, "Distributed battery control for peak power shaving in datacenters", in: *Proc. of IEEE International Green Computing Conference*, Arlington, VA, June 2013.

[18] B. Feijoo and R. R. Meyer, "Piecewise-Linear Approximation Methods for Nonseparable Convex Optimization", *Management Science*, vol. 34, no. 3, pp. 411–419, Mar. 1988.

[19] B. Kirby, *Ancillary services: technical and commercial insights*, prepared for Wartsila, July 2007.

[20] 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.

[21] 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.

[22] C. Wang and M. Groot, "Enabling Demand Response in a Computer Cluster", in: *Proc. of IEEE Smart Grid Comm*, Vancouver, BC, Oct. 2013.

[23] H. Chen, M.C. Caramanis, and A. K. Coskun, "The data center as a grid load stabilizer", in: *Proc. of ASPDAC*, Tokyo, Japan, Jan. 2014.

[24] Y. Choi and Y. Lim, "A Cost-efficient Mechanism for Dynamic VM Provisioning in Cloud Computing", in: *Proc. of ACM RACS*, Towson, MD, Oct. 2014.

[25] F. R. B. Cruz, "Optimizing the throughput, service rate, and buffer allocation in finite queueing networks", *Journal of Electronic Notes in Discrete Mathematics*, vol. 35, no. 1, pp. 163–168, Dec. 2009.

[26] D. Bunn, "Forecasting loads and prices in competitive power markets", *Proceedings of the IEEE*, vol. 88, no. 2, pp. 163–169, Feb. 2002.

[27] D. Das and B. F. Wollenberg, "Risk assessment of generators bidding in day-ahead market", *IEEE Trans. on Power Systems*, vol. 20, no. 1, pp. 416–424, Feb. 2005.

[28] 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.

[29] D. P. Bertsekas, *Convex Optimization Theory*, Nashua, NH: Athena Scientific, 2009.

[30] D. P. Kroese, T. Taimre, and Z. I. Botev, *Handbook of monte carlo methods*, Hoboken, NJ, USA: John Wiley & Sons Inc., 2011.

[31] 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.

[32] E. Pinheiro, R. Bianchini, E. V. Carrera, and T. Heath, "Dynamic Cluster Reconfiguration for Power and Performance", in: *Compilers and Operating Systems for Low Power*, ed. by M. Kandemir L. Benini and J. Ramanujam, Kluwer Academic Publishers, 2003.

[33] ERCOT, *Controllable Load Resource (CLR) Participation in the ERCOT Market*, www.ercot.com.

[34] ERCOT, *Load Participation in the ERCOT Nodal Market*, June 2007.

[35] ERCOT, *Load Participation in the ERCOT Nodal Market*, June 2007.

[36] F. Papier, *Optimization of Rental Systems: Queuing Loss Theory for the Optimization of Cargo Vehicle Rental Systems*, cologne germany: Kolner Wissenschaftsverlag, 2007.

[37] *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, Feb. 2012.

[38] M. Ghamkhari and H. Mohsenian-Rad, "Data centers to offer ancillary services", in: *Proc. of IEEE Smart Grid Comm*, Tainan City, Taiwan, Nov. 2012.

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

[40] 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, Oct. 1994.

[41] 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.

[42] S. Gurumurthi, A. Sivasubramaniam, M. Kandemir, and H. Franke, "Reducing disk power consumption in servers with DRPM", *Computer*, vol. 36, no. 12, pp. 59–66, Dec. 2003.

[43] 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, Oct. 2010.

[44] H. Mohsenian-Rad and A. Leon-Garcia, "Energy-Information Transmission Tradeoff in Green Cloud Computing", in: *Proc. of IEEE Conference on Global Communications (Globecom)*, Miami, FL, Dec. 2010.

[45] H. Mohsenian-Rad and A. Leon-Garcia, "Optimal Residential Load Control with Price Prediction in Real-Time Electricity Pricing Environments", *IEEE Trans. on Smart Grid*, vol. 1, no. 2, pp. 120–133, 2010.

[46] H. S. Kim and N. B. Shroff, "Loss probability calculations and asymptotic analysis for finite buffer multiplexers", *IEEE/ACM Trans. on Networking*, vol. 9, no. 6, pp. 755–768, Dec. 2001.

[47] D. P. Heyman, "Comments on a Queueing Inequality", *Management Science*, vol. 26, no. 9, pp. 956–959, Sept. 1980.

[48] 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, Apr. 2012.

[49] J. D. Kueck, A. F. Snyder, F. Li, and I. B. Snyder, "Use of Responsive Load to Supply Ancillary Services in the Smart Grid: Challenges and Approach", in: *Proc. of IEEE Smart Grid Comm*, Oct. 2010.

[50] 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.

[51] J. Dilley, "Web Server Workload Characterization", *Hewlett-Packard Laboratories* 1996.

[52] J. H. Kim and M. J. Lee, *Green IT: Technologies and Applications*, Springer, 2011.

[53] J. Heo, D. Henriksson, X. Liu, and T. Abdelzaher, "Integrating Adaptive Components: An Emerging Challenge in Performance-Adaptive Systems and a Server Farm Case-Study", in: *Proc. of the IEEE International Real-Time Systems Symposium*, Tucson, AZ, Dec. 2007.

[54] J. Nair, S. Adlakha, and A. Wierman, *Energy procurement strategies in the presence of intermittent sources*, (Under Submission) [Online] http://users.cms.caltech.edu/~ preprint.pdf, Nov. 2013.

[55] 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, Nov. 2006.

[56] K. K. Nguyen, M. Cheriet, M. Lemay, M. Savoie, and B. Ho, "Powering a Data Center Network via Renewable Energy: A Green Testbed", *IEEE Internet Computing*, vol. 17, no. 1, pp. 40–49, Jan. 2013.

[57] K. Kumaran, M. Mandjes, and A. Stolyar, "Convexity properties of loss and overflow functions", *Operations Research Letters*, vol. 31, no. 2, pp. 95–100, Mar. 2003.

[58] 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 Nov. 2011.

[59] C. Y. Kao and R. R. Meyer, "Secant Approximation Methods for Convex Optimization", *Computer Sciences Technical Report 352* Apr. 1979.

[60] B. Kirby, *Spinning Reserve From Responsive Loads*, Oak Ridge National Laboratory, 2003.

[61] 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*, Orlando, FL, 2010.

[62] 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.

[63] 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.

[64] 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, Oct. 2013.

[65] Zhenhua Liu, Iris Liu, Steven Low, and Adam Wierman, "Pricing Data Center Demand Response", in: *Proc. of ACM Sigmertics*, Austin, TX, June 2014.

[66] M. Chiang, A. Sutivong, and S. Boyd, "Efficient nonlinear optimization of queuing systems", in: *Proc. of IEEE Globecom*, Taiwan, Nov. 2002.

[67] 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, Aug. 2009.

[68] M. Ghamkhari and H. Mohsenian-Rad, "A Convex Optimization Framework for Service Rate Allocation in Finite Communications Buffers", *Accepted for publication in IEEE Communications Letters* 2015.

[69] 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.

[70] 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.

[71] 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.

[72] M. Shahidehpour, H. Yamin, and Z. Li, *Market Operations in Electric Power Systems*, New York, NY: IEEE Press, 2002.

[73] Smith. J. Macgregor, "Multi-server, Finite Waiting Room, M/G/c/K Optimization Models", *Asia-Pacific Journal of Operational Research*, vol. 25, no. 04, pp. 531–561, 2008.

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

[75] P. Wattles, "Load resources providing Ancillary Services in Electric Reliability Council of Texas (ERCOT)", in: *Proc. of IEEE Conference on Innovative Smart Grid Technologies*, Washington, DC, Jan. 2012.

[76] 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.

[77] R. Bianchini, "Leveraging renewable energy in data centers: present and future", in: *Proc. of ACM International symposium on High-Performance Parallel and Distributed Computing*, Delft, Netherlands, June 2012.

[78] R. Katz, "Tech Titans Building Boom", *IEEE Spectrum*, vol. 46, no. 2, pp. 40–54, Feb. 2009.

[79] R. Nagarajan and D. Towsley, "A Note on the Convexity of the Probability of a Full Buffer in the M/M/1/K Queue", *IEEE Journal on Selected Areas in Communication*, pp. 92–85, Aug. 1992.

[80] R. T. Rockafellar and S. Uryasev, "Optimization of Conditional Value-at-Risk", *Journal of Risk*, vol. 2, pp. 21–41, 2000.

[81] 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.

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

[83] 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.

[84] S. Boyd and L. Vandenberghe, *Convex Optimization*, New York, NY, USA: Cambridge University Press, 2004.

[85] S. Bu, F. R. Yu, Y. Cai, and X. P. Liu, "When the Smart Grid Meets Energy-Efficient Communications: Green Wireless Cellular Networks Powered by the Smart Grid", *IEEE Transactions on Wireless Communications*, vol. 11, no. 8, pp. 3014–3024, Aug. 2012.

[86] S. J. Deng and S. S. Oren, "Electricity derivatives and risk management", *Energy*, vol. 31, no. 6, pp. 940–953, May 2006.

[87] 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 Sept. 2014.

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

[89] 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.

[90] S. Steinke, N. Grunwald, L. Wehmeyer, R. Banakar, M. Balakrishnan, and P. Marwedel, "Reducing energy consumption by dynamic copying of instructions onto onchip memory", in: *Proc. of the International Symposium on System Synthesis*, Kyoto, Japan, Oct. 2002.

[91] U.S. Environmental Protection Agency, *EPA Report on Server and Data Center Energy Efficiency*, Final Report to Congress, Aug. 2007.

[92] 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.

[93] 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.

[94] 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.

[95] 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.

[96] X. P. Zhang, *Restructured Electric Power Systems: Analysis of Electricity Markets with Equilibrium Models*, Wiley IEEE Press, July 2010.

[97] 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*, Dec. 2013.

[98] 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, Dec. 2011.

[99] 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, Oct. 2013.

[100] Z. Liu, A. Wierman, Y. Chen, B. Razon, and N. Chen, "Data center demand response: Avoiding the coincident peak via workload shifting and local generation", in: *Proc. of ACM Sigmetrics*, Pittsburgh, PA, 2013.

[101] 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.

[102] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. H. Andrew, "Greening geographical load balancing", in: *Proc. of ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, New York, NY, June 2011.

[103] 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, Apr. 2015.

[104] 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, Apr. 2008.