Energy and Network Aware Workload Management for Sustainable Data Centers with Thermal Storage

Yuanxiong Guo, *Student Member, IEEE*, Yanmin Gong, *Student Member, IEEE*, Yuguang Fang, *Fellow, IEEE*, Pramod P. Khargonekar, *Fellow, IEEE*, and Xiaojun Geng, *Member, IEEE*

Abstract—Reducing the carbon footprint of data centers is becoming a primary goal of large IT companies. Unlike traditional energy sources, renewable energy sources are usually intermittent and unpredictable. How to better utilize the green energy from these renewable sources in data centers is a challenging problem. In this paper, we exploit the opportunities offered by geographical load balancing, opportunistic scheduling of delay-tolerant workloads, and thermal storage management in data centers to facilitate green energy integration and reduce the cost of brown energy usage. Moreover, bandwidth cost variations between users and data centers are considered. Specifically, this problem is first formulated as a stochastic program, and then, an online control algorithm based on the Lyapunov optimization technique, called Stochastic Cost Minimization Algorithm (SCMA), is proposed to solve it. The algorithm can enable an explicit trade-off between cost saving and workload delay. Numerical results based on real-world traces illustrate the effectiveness of SCMA in practice.

Index Terms—Data center, energy management, thermal storage, load scheduling, Lyapunov optimization

1 Introduction

To provide Internet-scale services such as social networking and web search with low latency and high reliability, Internet-service companies usually build multiple data centers distributed across different geographical locations. These data centers consume large amounts of electricity for powering both their IT equipments and cooling infrastructures. According to [1], the electric energy consumption of data centers for Internet applications accounts for 1.3 percent of the worldwide electricity usage in 2010 and this fraction is expected to increase to 8 percent by 2020. Therefore, intensive efforts have been made by Internet-service companies to reduce the electricity cost in their data centers.

Meanwhile, Internet-service companies are increasingly interested in becoming ''sustainable'', which requires them to reduce the environmental impact (i.e., carbon footprint) besides the financial impact (i.e., electricity cost) of their data centers. As shown in [2], two thirds of the worldwide electricity drawn from the utility grid is generated by fossil-fuel generators, such as coal, or gas plants, which emit much more carbon than renewable generators such as wind turbines and solar panels. With the decreasing cost of building renewable generators, they are becoming increasingly attractive options for powering data centers, especially when the renewable energy is supported by government incentives.

However, unlike the traditional brown energy drawn from the utility grid, green energy from renewable sources, especially wind and solar, is intermittent and uncontrollable, which presents a great challenge for data centers to effectively utilize them. The challenge is, in essence, the difficulty in instantaneously balancing of energy supply and demand. Large-scale electric energy storage, mainly batteries, can resolve this difficulty, but it is still prohibitively expensive.

To help integrate green energy into data centers, geographical load balancing [3], [4] has been proposed to utilize the agility of geographically distributed data centers by directing more user requests to places where renewable energy is abundant. Although geographical load balancing is useful, there are two more opportunities that can be exploited to further facilitate renewable energy integration in data centers. One observation is that data centers usually support a wide range of IT workloads, including both delay-sensitive, interactive applications such as web browsing and searching, and delay-tolerant, batch applications such as scientific computation and massively parallel and data intensive computational jobs. The interactive workload differs from the batch workload in the following two aspects. First, the computational requirement of the interactive workload is usually small, while the batch workload requires much larger computational capability. Second, while the performance metric appropriate to the interactive workload is the response time, for the batch

- *Y. Guo, Y. Gong, Y. Fang, and P.P. Khargonekar are with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA. E-mail: {guoyuanxiong, ymgong, ppk}@ufl.edu; fang@ece.ufl.edu.*
- *X. Geng is with the Department of Electrical and Computer Engineering, University of West Florida, Pensacola, FL 32514 USA. E-mail: xgeng@uwf.edu.*

Manuscript received 19 July 2013; revised 10 Oct. 2013; accepted 18 Oct. 2013. Date of publication 3 Nov. 2013; date of current version 16 July 2014. Recommended for acceptance by V. Misic.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPDS.2013.278

workload, it is the total throughput within some time period. The delay-tolerant property of batch workloads can be exploited to increase the renewable energy utilization by delaying their services to periods when renewable sources are abundant without exceeding their execution deadlines.

Another observation is that a large portion of the power consumption in a data center comes from the cooling infrastructure. Although large-scale electric energy storage, such as batteries, is very expensive, thermal storage is much cheaper, and can be leveraged to reduce the cooling energy cost. In fact, Apple has already deployed a chilled water storage system as the thermal storage facility in its green data center in Maiden, NC [5]. With the time-varying properties of wholesale electricity price and renewable energy generation, thermal storage can store some green energy from renewable generators or cheap brown energy from the utility grid first. Later, when the electricity price is high or the green energy is unavailable, the stored energy can be released to help cool the data center, therefore, reducing the electricity bill.

With the above observations as context, we explore the problem of joint geographical load balancing, delay-tolerant workload scheduling, and thermal storage management for green energy integration in geographically distributed data centers. In additional to the brown energy cost, we also take into account the bandwidth cost between cloud users and data centers. The objective is to minimize the total operating cost of serving delay-tolerant workloads. To tackle the randomness in renewable generations, workload arrivals, and electricity prices, we formulate the problem as a stochastic program and propose an efficient online algorithm, called Stochastic Cost Minimization Algorithm (SCMA), with provable performance guarantee based on the Lyapunov optimization framework [6]. In summary, the contributions of our work are as follows:

- By taking into account the delay-tolerant workloads and thermal storage, we formulate a stochastic optimization problem to minimize the total energy plus bandwidth cost of geographically distributed data centers with renewable generation.
- We propose an online distributed control algorithm SCMA to solve the problem without the requirements of knowing the detailed statistics of underlying randomness.
- Our proposed algorithm enables an explicit tradeoff between workload delay and cost saving, which can be flexibly adjusted by a control parameter V, making it an attractive control policy for data center operators with different applications.
- Through extensive numerical evaluations using real-world traces of renewable generation, workload arrival, and wholesale electricity price, we demonstrate the effectiveness of SCMA.

The remainder of this paper is organized as follows. Section 2 reviews some related work. In Section 5, models on workloads, renewable generation, thermal storage, and total operating cost are first presented and then, a stochastic optimization problem is formulated. We propose an algorithm called SCMA to solve it in Section 4. The analytical performance results of SCMA are described in Section 5. We present the numerical evaluation results based on real-world traces in Section 6. Finally, Section 7 concludes the paper.

2 Related Work

2.1 Renewable Energy Usage in Data Centers

Renewable-powered data centers are receiving more and more attention both in industry [5], [7] and in academia [3], [4], [8], [9]. Previous studies [3], [4] explore the feasibility and benefits of using geographical load balancing for delay-sensitive interactive workloads to facilitate the integration of renewable sources into data centers. Scheduling of delay-tolerant batch workload and energy storage to help integrate renewable sources into a data center with on-site renewable generation is discussed in [8]. System implementation issues with renewable energy-aware batch workload scheduler is discussed in [9] and prototypes are built to show the effectiveness of these job schedulers. However, all the aforementioned papers either consider a single data center, a single class of application, no energy/thermal storage facility, only delay-sensitive interactive workloads, or assume perfect future information. In contrast, our work jointly manages delay-tolerant workloads with thermal storage facilities in geographically distributed data centers having on-site renewable generations without future information.

2.2 Electric/Thermal Storage in Data Centers

Use of electric energy storage devices such as uninterruptible power supply (UPS) units to help reduce electricity cost is considered in [10], [11], [12]. However, batteries such as UPS units are quite expensive and cannot be overused since frequently charging and discharging severely impacts their lifetimes. On the other hand, thermal storage is much cheaper and can be utilized to reduce the cooling cost in data centers as shown in [13]. Therefore, in our work, we utilize the thermal energy storage to help reduce the cooling cost of data centers rather than assume the electric energy storage unit as in previous work [10], [11], [12].

2.3 Energy Cost Minimization in Data Centers

Reducing the electricity cost of Internet data centers has been the focus of a lot of research work in the past decade (see the most recent ones [12], [14], [15], [16], [17] and references therein). One direction is to reduce the amount of energy usage in data centers. Two main approaches exist along this direction: achieving power-proportionality and lowering energy overhead. Power-proportionality means consuming power directly proportional to the utilization level, which can be achieved by dynamic voltage/frequency scaling (DVFS), or dynamic capacity provision (DCP). The energy overhead of a data center is measured by the power usage effectiveness (PUE) metric, which denotes the ratio of the total facility power consumption to the IT equipment power consumption. Various schemes, such as advanced cooling methods and direct current power infrastructure, have been designed to lower the PUE. Another direction is to use geographical load balancing to exploit the diversity of electricity prices in multiple data centers, where more interactive requests would be routed into data centers with

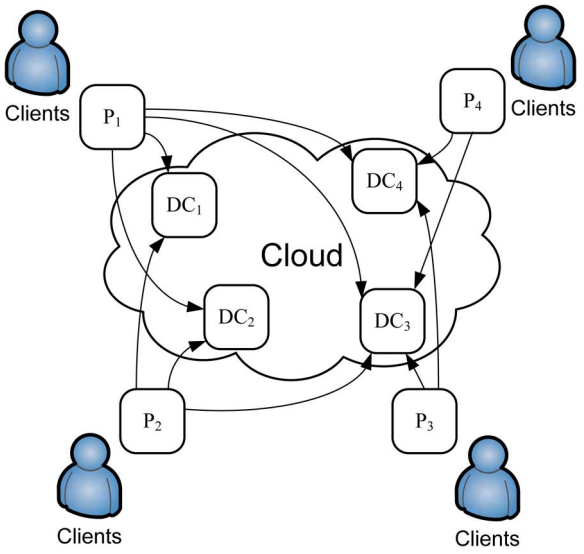


Fig. 1. Typical cloud network architecture of a CSP.

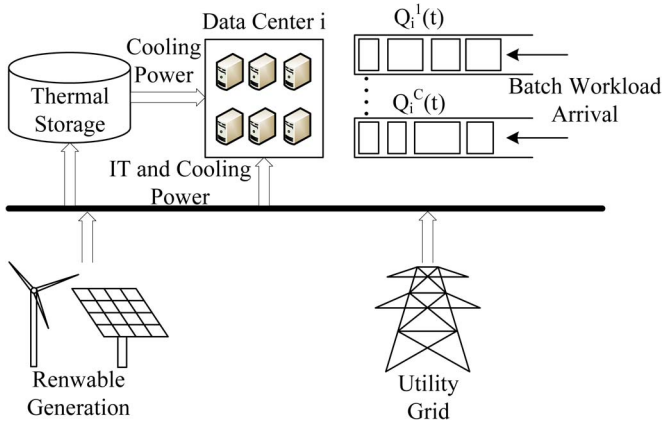


Fig. 2. Block diagram of a sustainable data center.

lower electricity prices or more renewable energy. However, most of the previous efforts mainly focus on the interactive, delay-sensitive workloads. Delay-tolerant workload scheduling is also considered in [16], [17], [18], but these papers do not consider on-site renewable energy generation or thermal storage. In [19], we only consider the energy cost minimization in a data center with renewable energy generation and thermal storage. However, neither the bandwidth cost nor the thermal storage cost is considered.

2.4 Bandwidth Cost Minimization in Data Centers

Traffic engineering in data centers for efficient network utilization has been discussed in [20], [21], [22]. These papers mainly focus on VM placement or migration in data centers to reduce bandwidth cost while ignoring many aspects of energy cost. These studies are complementary to our work in the sense that the network-aware job placement algorithms within a single data center can be exploited after our algorithm determines the jobs to be routed to each data center.

3 Modeling and Optimization

We consider a cloud service provider (CSP) having multiple geographically distributed data centers, each with on-site renewable generators and thermal storage. The typical cloud network architecture of a CSP is depicted in Fig. 1, in which there are several front-end proxies near the clients and multiple back-end remote data centers in the cloud. Assume that there are M proxies and each of them is responsible for a geographically concentrated source of requests such as a city. The proxy directs user requests to N data centers of the CSP in the cloud. Due to the spatial diversity, traffic between different pairs of proxy and data center goes through different ISPs and therefore, incurs different bandwidth costs. The sustainable data center we consider in this paper is illustrated in Fig. 2, which is explained in detail as follows. We consider a discrete-time system with time denoted by $t = 0, 1, 2, \ldots$.

3.1 The Workload Model

There are many different workloads in data centers. In general, they can be divided into the following two categories: delay-sensitive interactive workload and delay-tolerant batch workload [8]. Delay-sensitive interactive workloads such as web services usually process real-time user requests, which have to be completed within a certain time, i.e., there is a maximum response time. Some batch workloads such as scientific applications, simulations, or MapReduce jobs [23] are often delay-tolerant in the sense that they can be scheduled to run at any time as long as the jobs are completed before the deadline, i.e., there is a maximum completion time. Since interactive workloads have higher priority, they are usually provisioned first. In this paper, we focus on computation-intensive batch workload management, assuming that the management of interactive workloads has been determined by previous schemes [3], [24].

Consider C types of jobs or service requests in the delay-tolerant workloads. Each type may correspond to a specific application. Assume that all jobs are computation-intensive, and the CPU resource is the bottleneck resource. That is, a job is executed whenever the CPU resource is allocated to it. A job is represented by a tuple: (c, d_c, n_c), where c denotes the application type, d_c denotes computation demand (i.e., job length) in terms of the processor cycles, and n_c denotes the communication demand in terms of the transmitted data size between the cloud and the client. We assume that jobs of different types have different IT resource requirements (e.g., CPU, memory, storage, and network) and jobs of the same type have the same IT resource requirements.

A job or service request first arrives at the front-end proxy j. The proxy is near the clients and acts as a workload router. The proxy would decide which back-end data center the job request should be routed to for processing. We assume no data buffering at the proxy so that whenever a request arrives at the proxy, it would be routed to a data center for processing immediately. Denote the number of type c jobs arriving at proxy j in time t as $W_j^c(t)$. The job arrival rate vector at time t is denoted as $\mathbf{W}(t) = (W_j^c(t), \forall c, j)$ and the time-average rate of such an arrival vector is denoted as $\omega = \mathbb{E}\{\mathbf{W}(t)\}$. We assume that the total arrival

rate of type c jobs is bounded by a finite positive constant $W^c_{\max}$. That is,

$$\sum_{j=1}^{J} W^c_j(t) \le W^c_{\max},\ \forall c,t. \tag{1}$$

We use $\lambda^c_{ij}(t)$ to denote the number of type c jobs that is routed from proxy j to data center i in time t, and use $\boldsymbol{\lambda}^c_j(t) = (\lambda^c_{ij}(t), \forall i)$ to denote the routing vector for type c jobs at proxy j. In every time period t, $\boldsymbol{\lambda}^c_j(t)$ must draw from some feasible routing set $\Lambda^c_j(t)$, which includes, but is not limited to, the following constraints:

$$\sum_{i=1}^{N} \lambda^c_{ij}(t) = W^c_j(t), \quad \forall c,j,t \tag{2}$$

$$\lambda^c_{ij}(t) \ge 0, \quad \forall c,i,j,t. \tag{3}$$

Additional constraints can be added into the feasible set $\Lambda^c_j(t)$ to model other practical considerations. For example, if jobs of application c from proxy j can only be routed into a set of data centers I^c_j due to security concern, then we have $\lambda^c_{ij}(t) = 0, \forall i \notin I^c_j$. If a job contains several tasks which need to communicate with each other during the processing, we may need to place the whole job inside one data center to reduce the inter-DC traffic, which is much costlier than the intra-DC traffic. Then, $\lambda^c_{ij}(t)$ should be an integer value. Other practical constraints can be formulated into the set $\Lambda^c_j(t)$ similarly.

Denote the queue length of type c jobs at the back-end data center DC_i as $Q^c_i(t)$. Then, we have the following queue dynamics:

$$Q^c_i(t+1) = \max\{Q^c_i(t) - x^c_i(t), 0\} + \sum_{j=1}^{M} \lambda^c_{ij}(t), \tag{4}$$

where $x^c_i(t)$ is the number of type c jobs processed at data center i in time t. For each data center i, denote the processing speed of the server as μ_i and the total number of servers for serving delay-tolerant workloads as IT_i. Since the processed workload cannot exceed the maximum available computing resources, we have

$$\sum_{c=1}^{C} x^c_i(t) d_c \le \mathrm{IT}_i \mu_i,\ \forall i,t. \tag{5}$$

Note that in the formulation above, we implicitly assume that the jobs can be perfectly parallelized and are tolerant to interruption during running time. The jobs we consider in this work are the same as the jobs that can be supported by the Amazon EC2 spot instances [25], which are time-flexible and interruption-tolerant. We need to control the system so that the queues in the system are stabilized according to the following definition:

$$\overline{Q} \triangleq \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{i=1}^{N} \sum_{c=1}^{C} \mathbb{E}\{Q^c_i(t)\} < \infty. \tag{6}$$

3.2 The Renewable Generation Model

There are several possible approaches for Internet-service companies to utilize renewable energy in their data centers [26], where power purchasing agreement (PPA) and on-site renewable generation are two commonly used methods in industry now. In the first approach, the data center operator negotiates a long-term PPA with a renewable energy producer, and directly purchases a certain amount of the green energy generated by the producer at a negotiated price. Renewable energy certificates (RECs) are kept by the data center operator as the proof of its green energy usage. For example, Google has contracted to purchase 114 MW of wind power for 20 years from a wind project in Iowa to power its data center there [7]. The second approach is to build on-site renewable generators near data centers, which can reduce the transmission and distribution losses. For example, Apple is building the nation's largest end user-owned solar array (40 MW) and also, the largest nonutility fuel cell installation (5 MW) in the US at its new data center in Maiden, NC [5]. These on-site renewable generators will provide over 60 percent of the clean power it needs. In this paper, we focus on the second approach because it has a more direct impact on ''greening'' data centers.

Denote the amount of on-site renewable energy generated at data center DC_i during period t as $r_i(t)$. Since renewable energy sources, mainly solar or wind, are highly intermittent, time-varying, and uncontrollable, they may vary a lot even within one period (e.g., 10 mins) in our scenario. In practice, as observed in [10], data centers usually have excess energy storage capability in UPS units, which can provide such a ''smoothing'' function. Under this assumption, the renewable generation can be regarded as being constant during one time period.

3.3 The Thermal Storage Model

As explained in [13], there are basically two kinds of thermal storage technologies used in data centers. One is the inherent thermal masses in a data center such as the cold air and the raised metal floor. They can be over-cooled to a lower temperature by the CRAC system first, and absorb heat later as a cooling unit. The other is the dedicated thermal storage system. Thermal energy storage systems commonly use chilled liquid or ice to act as a thermal battery, enabling a data center operator to run air conditioners at night (when rates are lower) and during the day, pump the chilled liquid around the facility for cooling. While there is no extra capital cost for the first approach, its capacity is usually limited and therefore, it is only suitable for short-term storage. In this paper, we consider the second approach, where each data center has a chilled liquid/ice storage system besides the CRAC cooling system. Note that our thermal storage-based approach is orthogonal and supplementary to other approaches, such as DC power distribution and seawater cooling, used for reducing the cooling cost of data centers.

For each data center i, denote by $S^{\max}_i$ the capacity of the thermal storage, by $S_i(t)$ the energy level at period t, by $s^+_i(t)$ the energy stored (i.e., charged) into the thermal storage system in period t, and by $s^-_i(t)$ the energy released (i.e., discharged) from the thermal storage system in period t. In practice, there is conversion loss during the energy conversion process. Without loss of generality, we assume the conversion loss only happens in the charging process and denote the round-trip efficiency of the thermal storage

system at data center i as $\eta_i < 1$. Then, $S_i(t)$ would denote the usable energy in the thermal storage and has the following dynamics at data center i:

$$S_i(t+1) = S_i(t) + \eta_i s_i^+(t) - s_i^-(t). \quad (7)$$

Also, each thermal storage usually has an upper bound on the charge rate, denoted by $s_{i,\max}^+$, and an upper bound on the discharge rate, denoted by $s_{i,\max}^-$. That is,

$$0 \le s_i^+(t) \le s_{i,\max}^+, \quad (8)$$

$$0 \le s_i^-(t) \le s_{i,\max}^-. \quad (9)$$

We also define $s_{\max}^+ \triangleq \max_i s_{i,\max}^+$ as the maximum charge rate of thermal storage systems at all data centers.

Within one control period, the thermal storage can be either charged or discharged, but not both [10]. That is,

$$s_i^+(t) > 0 \Rightarrow s_i^-(t) = 0, \ s_i^-(t) > 0 \Rightarrow s_i^+(t) = 0. \quad (10)$$

However, we will temporarily ignore this constraint and decide the optimal charge/discharge control actions. Later, we will construct the control decisions that can meet that constraint without performance degradation.

For each time period, we need to ensure that the thermal energy level in data center i always satisfies the following:

$$0 \le S_i(t) \le S_i^{\max}. \quad (11)$$

Note that some thermal storage systems may have a nonzero minimum energy level requirement to protect the lifetime of their system. Without loss of generality, we assume that the minimum energy level is zero while $S_i^{\max}$ denotes the usable thermal storage capacity. The initial energy level in data center i is assumed to be $S_i(0) \in [0, S_i^{\max}]$.

Since the excessive usage of thermal storage would impact its lifetime and reliability, as with [27], the loss of the thermal storage value is modeled as a cost which is proportional to the recharged energy with a factor γ_i. That is, the operating cost of using thermal storage at data center i in period t is $\gamma_i s_i^+(t)$.

3.4 The Cost Model

Besides the cost of using thermal storage systems as described before, there are two other parts of the total operating cost: one is the energy cost used to serve the workload in data centers and the other is the bandwidth cost between the clients near the proxies and the data centers in the cloud.

To incentivize the usage of green energy from renewable generators, we assume that the marginal cost of renewable generation is zero so that the data centers should utilize it as much as possible. The cost of traditional brown energy drawn from the utility grid depends on the wholesale electricity market and is both spatially and temporally varying. Denote by $p_i(t)$ the brown energy price bought from the wholesale electricity market at data center DC_i in period t. It is both time-varying and location-dependent. We assume that $0 \le p_i(t) \le p_i^{\max}$ for all periods t and $p_i^{\max} > \gamma_i/\eta_i$.[1]

The power consumption of a server in data center i can be approximated to be linearly related to the average CPU utilization as follows [28]:

$$(1-\alpha)P_i^{idle} + \alpha P_i^{busy} \quad (12)$$

where P_i^{idle} is the power consumption when the server is in idle state, $\alpha \in [0,1]$ is the average CPU utilization level, and P_i^{busy} is the power consumption when the server is busy. Therefore, given the service rate $x_i^c(t)$ for type c jobs at period t and the maximum available active server numbers IT_i, the IT power consumption for data center i is

$$E_i(t) = \mathrm{IT}_i P_i^{idle} + \frac{\sum_c x_i^c(t) d_c}{\mu_i}\left(P_i^{busy} - P_i^{idle}\right). \quad (13)$$

Denote by $f_i(t)$ the corresponding cooling energy usage in data center i during time t. Since we focus on the thermal storage in this paper, we assume that the discharged power from the thermal storage cannot be greater than the cooling demand[2], i.e.,

$$s_i^-(t) - f_i(t) \le 0. \quad (14)$$

In practice, $f_i(t)$ may be a convex function, depending on the specific cooling infrastructures such as CRAC and air cooling systems [8]. For simplicity of analysis, we assume that $f_i(t)$ is a linear function of the total IT power consumption in the following form:

$$f_i(t) = \beta_i E_i(t), \quad (15)$$

where β_i is a factor to represent the power usage efficiency of data center. On average, β_i is around 1 for the data center industry [7]. That is, for every watt of IT power, an additional watt is consumed to cool and distribute power to the IT equipment. Although intensive research has been done to reduce the power usage efficiency of data centers, energy storage has appeared as an attractive mechanism quite recently [10], [12]. Note that our framework is quite general and can incorporate more practical cooling models such as [8]. With the above models, the energy cost plus the thermal storage operating cost of data center i in period t is as follows:

$$\mathcal{E}_i(t) = p_i(t)\left[(1+\beta_i)E_i(t) + s_i^+(t) - s_i^-(t) - r_i(t)\right]^+ + \gamma_i s_i^+(t). \quad (16)$$

Meanwhile, there is bandwidth cost involved for the communication between the jobs routed into the data center and the client near the proxy. In this paper, we use the following linear bandwidth cost model to represent the bandwidth cost between the clients and the cloud:

$$\mathcal{B}_{ij}(t) = \sum_{c=1}^{C} b_{ij} \lambda_{ij}^c(t) n_c, \quad (17)$$

where b_{ij} is the bandwidth cost coefficient between proxy j and data center i. Note that n_c is the communication

1. Note that this assumption represents that there is opportunity to utilize storage for cost reducing.

2. Note that for a electric energy storage, the discharged power can also be used to power the servers, therefore, eliminating the constraint (14). Our framework and the proposed techniques are still applicable to the case of electric energy storage systems with minor modification.

demand between the cloud and the client. Different pairs of proxy and data center have different bandwidth cost. More practical bandwidth charging model based on 95-th percentile bandwidth usage may be modeled similarly and would be our future investigation. We define $b_{\max} \triangleq \max_{ij} b_{ij}$ as the maximum bandwidth cost coefficient between any pair of proxy and data center. The total operating cost of serving delay-tolerant workloads for a CSP in time period t is $\sum_{i=1}^{N} \mathcal{E}_i(t) + \sum_{i=1}^{N}\sum_{j=1}^{M} \mathcal{B}_{ij}(t)$.

3.5 Problem Formulation

In this paper, we are interested in minimizing the time-average total operating cost for serving the delay-tolerant workloads over a large time horizon. Therefore, the control problem can be stated as follows: for the dynamic system defined by (4) and (7), design a control strategy which, given the past and the present random renewable supplies, workload arrivals, and electricity prices, chooses the workload routing decisions $\boldsymbol{\lambda}$, the thermal storage decisions $\mathbf{s}^+$ and $\mathbf{s}^-$, and the IT resource allocation decisions $\mathbf{x}$ such that the time-average total operating cost for serving delay-tolerant workloads is minimized. It can be formulated as the following stochastic optimization:

$$\min_{\substack{\boldsymbol{\lambda},\mathbf{x},\\ \mathbf{s}^+,\mathbf{s}^-}} : \overline{g} = \limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\{\sum_{i=1}^{N} \mathcal{E}_i(t) + \sum_{i=1}^{N}\sum_{j=1}^{M} \mathcal{B}_{ij}(t)\right\}, \quad (18a)$$

s.t.

$$x_i^c(t) \geq 0, \quad \sum_{c=1}^{C} x_i^c(t) d_c \leq \mathrm{IT}_i \mu_i, \ \forall c, i, t \quad (18b)$$

$$S_i(t+1) = S_i(t) + \eta_i s_i^+(t) - s_i^-(t), \ \forall i, t \quad (18c)$$

$$0 \leq S_i(t) \leq S_i^{\max}, \ \forall i, t \quad (18d)$$

$$0 \leq s_i^+(t) \leq s_{i,\max}^+, \ \forall i, t \quad (18e)$$

$$0 \leq s_i^-(t) \leq s_{i,\max}^-, \ \forall i, t \quad (18f)$$

$$s_i^-(t) - f_i(t) \leq 0, \ \forall i, t \quad (18g)$$

$$\left(\lambda_{ij}^c(t), \ \forall i\right) \in \Lambda_j^c(t), \ \forall c, j, t \quad (18h)$$

$$\overline{Q} < \infty. \quad (18i)$$

Here (18b) means that the total allocated IT resources cannot exceed the IT capacity. (18h) denotes that the workload admission and routing vectors should be within the feasible set, which depends on the real application. (18i) ensures that the average total queue length for buffering delay-tolerant jobs is finite so that the dynamic system is stable.

One challenge of solving the problem above is the constraint (18d), which brings the ''time-coupling'' property to the control decisions. Specifically, the current control decisions $s_i^+(t)$, $s_i^-(t)$ will have an impact on the future control decisions. In the later part, we will design a ''virtual energy queue'' to remove this ''time-coupling'' property while also ensuring the constraint (18d).

4 Algorithm Design

In this section, we design an online algorithm based on the Lyapunov optimization technique [6] to solve the stochastic optimization problem above. Because of the time-coupling constraint (18d), Lyapunov optimization technique cannot be applied directly. In the following, we first consider a relaxed problem, which fits into the framework of Lyapunov optimization. Then, we design our algorithm based on the insights provided by this relaxed problem.

4.1 Relaxed Problem

Denote the time-average expected charge and discharge rate of thermal storage i, respectively, as follows:

$$\overline{s_i^+} = \limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\{s_i^+(t)\}, \quad (19)$$

$$\overline{s_i^-} = \limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\{s_i^-(t)\}. \quad (20)$$

According to the dynamics of thermal storage energy level (7), in order to ensure $0 \leq S_i(t) \leq S_i^{\max}$ for all t, we must have the following equation:

$$\eta_i \overline{s_i^+} = \overline{s_i^-}. \quad (21)$$

Therefore, we have the following relaxed problem:

$$\min_{\substack{\boldsymbol{\lambda},\mathbf{x},\\ \mathbf{s}^+,\mathbf{s}^-}} : \overline{g} = \limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\{\sum_{i=1}^{N} \mathcal{E}_i(t) + \sum_{i=1}^{N}\sum_{j=1}^{M} \mathcal{B}_{ij}(t)\right\}, \quad (22)$$

subject to constraints (18b), (18e), (18f), (18g), (18h), (18i), and (21).

The optimal solution to the relaxed problem above is easy to characterize based on the framework of Lyapunov optimization, which is described in the following theorem. Theorem 1 (below) shows that we can achieve the minimum time average operating cost for a given workload arrival rate vector $\boldsymbol{\omega}$ using a stationary, randomized algorithm. The algorithm only chooses control decisions according to a fixed probability distribution that depends on the system state $(r_i(t), p_i(t), W_j^c(t), \forall i, j, c)$, but is independent of $(Q_i^c(t), E_i(t), \forall i, c)$. In Theorem 1, $\boldsymbol{\Omega}$ denotes the capacity region of the system, which is the closure of sets of rates $\boldsymbol{\omega}$ for which there exists a joint geographical load balancing, workload scheduling, and storage management algorithm that can ensure the queue stability (6).

Theorem 1. *If the vector $(r_i(t), p_i(t), W_j^c(t), \forall i, j, c)$ is i.i.d. over periods, then, for any arrival rate vector $\boldsymbol{\omega} \triangleq \mathbb{E}\{\mathbf{W}(\mathbf{t})\} \in \boldsymbol{\Omega}$, there exists a stationary, randomized control policy that chooses control decisions $\tilde{\lambda}_{ij}^c(t)$, $\tilde{x}_i^c(t)$, $\tilde{s}_i^+(t)$ and $\tilde{s}_i^-(t)$, based solely on the value of $(r_i(t), p_i(t), W_j^c(t), \forall i, j, c)$ irrespective of queue information while satisfying all constraints of the relaxed problem and providing the following guarantees:*

$$\mathbb{E}\left\{\sum_{j=1}^{M} \tilde{\lambda}_{ij}^c(t) - \tilde{x}_i^c(t)\right\} = 0, \quad \forall i, c, t \quad (23)$$

$$\mathbb{E}\{\eta_i \tilde{s}_i^+(t)\} = \mathbb{E}\{\tilde{s}_i^-\}, \quad \forall i, t \quad (24)$$

$$\mathbb{E}\left\{\sum_{i=1}^{N} \tilde{E}_i(t) + \sum_{i=1}^{N}\sum_{j=1}^{M} \tilde{B}_{ij}(t)\right\} = \overline{g}_{rel}^*(\boldsymbol{\omega}), \quad \forall t \quad (25)$$

where the expectations are w.r.t. the randomness in $(r_i(t), p_i(t), W_j^c(t), \forall i, j, c)$ *and possibly, randomized control decisions, and* $\overline{g}_{rel}^*(\boldsymbol{\omega})$ *is the optimal objective value of the relaxed problem (22) given an arrival rate vector* $\boldsymbol{\omega}$.

Proof. The result follows from Theorem 4.5 of [6] and is proved by using the Caratheodory's theorem. It is omitted here for brevity. □

Denote the optimal objective value of the original problem (18) as $\overline{g}^*(\boldsymbol{\omega})$ given an arrival rate vector $\boldsymbol{\omega}$. Obviously, $\overline{g}_{rel}^*(\boldsymbol{\omega}) \leq \overline{g}^*(\boldsymbol{\omega})$. Let

$$A_1 \triangleq \sum_{c=1}^{C} W_{\max}^c b_{\max} n_c + \sum_{i=1}^{N}\left\{\sum_{i=1}^{N} p_i^{\max}(1+\beta_i)\mathrm{IT}_i P_i^{busy} + (p_i^{\max}+\gamma_i)s_{i,\max}^+\right\}. \quad (26)$$

From the bounds we assumed before, we have $\overline{g} \leq A_1$ for any feasible control policy subject to constraints (18b), (18e), (18f), and (18h). Instead of solving the relaxed problem, we will use the existence of such an optimal policy to help us design our control policy that meets all constraints of the original problem (18), and derive the performance of our algorithm.

4.2 The Stochastic Cost Minimization Algorithm (SCMA)

The idea of our algorithm is to construct a Lyapunov-based control algorithm for determining the optimal workload routing, scheduling, and thermal storage management scheme.

First, we define a Lyapunov function as follows:

$$L(t) \triangleq \frac{1}{2}\sum_{i=1}^{N}\left[\sum_{c=1}^{C}\left(Q_i^c(t)\right)^2 + \left(S_i(t)-\theta_i\right)^2\right], \quad (27)$$

where θ_i is a constant to be specified later. Now define $\mathbf{K}(t) \triangleq (Q_i^c(t), S_i(t), \forall i, c)$, and define a one-period conditional Lyapunov drift as follows:

$$\nabla(t) \triangleq \mathbb{E}\{L(t+1) - L(t)|\mathbf{K}(t)\}. \quad (28)$$

Here the expectation is taken over the randomness of workload arrival, electricity price, and renewable generation, as well as the randomness in choosing the control actions. Then, following the Lyapunov optimization framework, we add a function of the expected cost over one period (i.e., the penalty function) to (28) to obtain the following *drift-plus-penalty* term:

$$\nabla_V(t) \triangleq \nabla(t) + V\mathbb{E}\left\{\sum_{i=1}^{N}\mathcal{E}_i(t) + \sum_{i=1}^{N}\sum_{j=1}^{M}\mathcal{B}_{ij}(t)|\mathbf{K}(t)\right\}, \quad (29)$$

where V is a positive control parameter to be specified later. Then, we have the following lemma regarding the *drift-plus-penalty* term:

Lemma 1. *For any feasible action under constraints (18b), (18e), (18f), (18g), and (18h) that can be implemented at period* t, *we have*

$$\nabla_V(t) \leq A_2 + \sum_{i=1}^{N}\mathbb{E}\left\{\left(S_i(t)-\theta_i\right)\left(\eta_i s_i^+(t) - s_i^-(t)\right)|\mathbf{K}(t)\right\} + \sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{E}\left\{Q_i^c(t)\left(\sum_{j=1}^{M}\lambda_{ij}^c(t) - x_i^c(t)\right)|\mathbf{K}(t)\right\} + V\sum_{i=1}^{N}\mathbb{E}\{\mathcal{E}_i(t)\} + V\sum_{i=1}^{N}\sum_{j=1}^{M}\mathbb{E}\left\{b_{ij}\sum_{c=1}^{C}\lambda_{ij}^c(t)n_c|\mathbf{K}(t)\right\}, \quad (30)$$

where A_2 *is the constant given by the following:*

$$A_2 \triangleq \sum_{i=1}^{N}\frac{\max\left\{\left(\eta_i s_{i,\max}^+\right)^2, \left(s_{i,\max}^-\right)^2\right\}}{2} + \sum_{i=1}^{N}\sum_{c=1}^{C}\frac{\left[\left(W_{\max}^c\right)^2 + \left(\frac{\mathrm{IT}_i\mu_i}{d_c}\right)^2\right]}{2}. \quad (31)$$

Proof. See the supplementary document which is available in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPDS.2013.278. □

We now present the SCMA formulation. The main design principle of our algorithm is to choose control actions that greedily minimize the R.H.S. of (30). Our algorithm can be naturally decomposed into two parts: workload routing and joint workload scheduling and storage management, as follows:

Stochastic Cost Minimization Algorithm: Initialize V and $\theta_i, \forall i$. At each period t, observe $(W_j^c(t), r_i(t), p_i(t), \forall i, j, c)$ and $\mathbf{K}(t)$, and do:

- *Workload Routing*: For each proxy j, choose the routing vector $((\lambda_{ij}^c)^*, \forall i)$ for type c jobs as the solution to the following problem:

$$\min: \quad \sum_{i=1}^{N}\left(Q_i^c(t) + Vb_{ij}n_c\right)\lambda_{ij}^c(t)$$
$$\text{s.t.} \quad \left(\lambda_{ij}^c, \forall i\right) \in \Lambda_j^c(t). \quad (32)$$

- *Workload Scheduling and Storage Management*: For each data center i, choose the workload scheduling vector $\{(x_i^c(t))^*, \forall c\}$ and thermal storage decisions $(s_i^+(t))^*$ and $(s_i^-(t))^*$ as the solution to the following linear optimization problem:

$$\textbf{Minimize}:$$
$$-\sum_{c=1}^{C}Q_i^c(t)x_i^c(t) + \left(S_i(t)-\theta_i\right)\left(\eta_i s_i^+(t) - s_i^-(t)\right) + Vy_i,$$
$$\textbf{s.t.}$$
$$y_i \geq p_i(t)(1+\beta_i)\left[\frac{\sum_{c=1}^{C}x_i^c(t)d_c}{\mu_i}\left(P_i^{busy} - P_i^{idle}\right) + \mathrm{IT}_i P_i^{idle}\right]$$

$$
\begin{aligned}
&+ (p_i(t)+\gamma_i)s_i^+(t) - p_i(t)\big(s_i^-(t)+r_i(t)\big),\\
y_i &\geq \gamma_i s_i^+(t),\\
s_i^-(t) &\leq \beta_i\left[\mathrm{IT}_i P_i^{idle} + \frac{\sum_{c=1}^{C} x_i^c(t) d_c}{\mu_i}\left(P_i^{busy} - P_i^{idle}\right)\right],\\
0 \leq s_i^+(t) &\leq s_{i,\max}^+,\\
0 \leq s_i^-(t) &\leq s_{i,\max}^-,\\
x_i^c(t) &\geq 0,\ \forall c,\\
\sum_{c=1}^{C} x_i^c(t) d_c &\leq \mathrm{IT}_i \mu_i, \qquad (33)
\end{aligned}
$$

where y_i is a slack variable used to transform the nonlinear operator $[\cdot]^+$ into linear ones.

- *Queue Update*: Update $\mathbf{K}(t)$ according to the dynamics (4) and (7).

Note that when solving the problem (33), the resulting optimal charge/discharge solution may not satisfy the constraint (10). In this case, let $H \triangleq \eta_i (s_i^+(t))^* - (s_i^-(t))^*$ and we define the actual thermal storage charge and discharge rates as follows:

$$
\big(s_i^+(t)\big)' = \begin{cases} \frac{H}{\eta_i} & \text{if } H \geq 0,\\ 0 & \text{otherwise.} \end{cases} \qquad (34)
$$

$$
\big(s_i^-(t)\big)' = \begin{cases} -H & \text{if } H < 0,\\ 0 & \text{otherwise.} \end{cases} \qquad (35)
$$

We have the following lemma regarding the optimality of the actual thermal storage charge and discharge rates:

Lemma 2. *The thermal storage charge and discharge rates $(s_i^+(t))'$ and $(s_i^-(t))'$ above is also an optimal solution to the problem (33).*

Proof. See the supplementary document which is available online. □

Under the above actual charge/discharge decisions, we present the following two properties of the structure of the optimal solution to (33) that is useful in the performance analysis.

Lemma 3. *The optimal solution to (33) with the additional constraint (10) has the following properties:*

1. If $S_i(t) - \theta_i > -V\gamma_i/\eta_i$, then $(s_i^+(t))^* = 0$.
2. If $S_i(t) - \theta_i < -Vp_i(t)$, then $(s_i^-(t))^* = 0$.

Proof. See the supplementary document which is available online. □

4.3 Interpretation of SCMA

The detailed control decisions taken by SCMA are as follows:

- The complexity of solving the workload routing problem (32) depends on the feasible set $\Lambda_j^c(t)$. Usually, it has a threshold-based solution. For example, suppose that the feasible set $\Lambda_j^c(t)$ only contains constraints (2), (3), $\lambda_{ij}^c(t) = 0, \forall i \notin I_j^c$, and $\lambda_{ij}^c(t) \in \mathbb{Z}$. The optimal solution is the following threshold-based policy: Let

$$
i^* = \arg\min_{i \in I_j^c} \big(Q_i^c(t) + Vb_{ij}n_c\big). \qquad (36)
$$

Then,

$$
\Big(\lambda_{ij}^c(t)\Big)^* = \begin{cases} W_j^c(t) & \text{if } i = i^*,\\ 0 & \text{if } i \neq i^*. \end{cases} \qquad (37)
$$

It means that all the jobs would be routed to the data center with the shortest queue length or the lowest bandwidth cost. The weights of the queue length and the bandwidth cost are adjusted by the parameter V.

- From the problem formulation (33), we can see that SCMA will always use the renewable energy $r_i(t)$ as much as possible to serve queued workloads irrespective of queue lengths and electricity prices so that the first term in the objective is minimized while the third term in the objective is unchanged. When $Vp_i(t)(1+\beta_i)d_c(P_i^{busy} - P_i^{idle})/\mu_i < Q_i^c(t)$, which means that the electricity price is low enough or the queue length for type c jobs is high enough, SCMA will also use some brown energy to serve jobs of type c if needed. For thermal storage management, when $S_i(t) - \theta_i > 0$, the stored energy in thermal storage will be used to cool the data center, since there is enough energy stored in it. Also, if the current electricity price is low enough such that $p_i(t) < (\theta_i - S_i(t))\eta_i/V - \gamma_i$, the thermal storage will store energy as much as possible for later use to leverage the opportunity of current low electricity price. The thresholds of charging or discharging depend on the current stored energy level as well as the parameter V.

Note that SCMA only requires the knowledge of the instantaneous values of system dynamics and can operate online without requiring any knowledge of the statistics of these stochastic processes. Moreover, each proxy or data center solves its own optimization problem distributively, where only the queue length information of data centers needs to be exchanged between data centers and proxies. Therefore, SCMA is easy to implement in practice.

5 Performance Analysis

In this section, we present the analytical performance results for SCMA. Detailed numerical results are described in the next section. First, we present the results when $(r_i(t), p_i(t), W_j^c(t), \forall c, i, j)$ is i.i.d. stochastic process. Note that according to the framework of Lyapunov optimization [6], our results can also be extended to the more general setting where $(r_i(t), p_i(t), W_j^c(t), \forall c, i, j)$ evolves according to some finite state irreducible and aperiodic Markov chain. Furthermore, our numerical simulation results in the next section are based on the real-world traces without any specific distribution assumption.

Theorem 2. *Suppose that $0 < V \leq V_{\max}$, where $V_{\max} \triangleq \min_i\{(S_i^{\max} - \eta_i s_{i,\max}^+ - s_{i,\max}^-)/(p_i^{\max} - \gamma_i/\eta_i)\}$. Let*

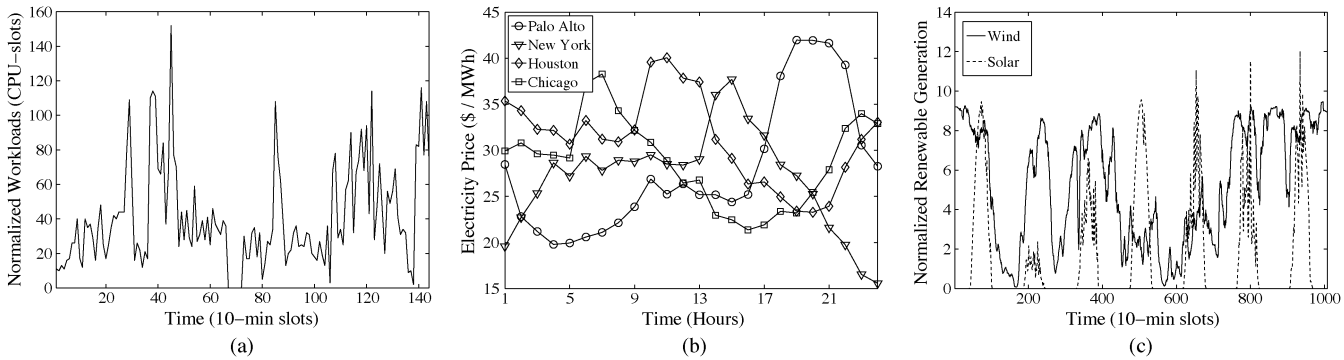


Fig. 3. Real-world traces used in evaluations. (a) 10-min average workload arrivals for one day [29]. (b) Hourly electricity prices in day-ahead markets for one day at four locations [30], [31]. (c) 10-min average solar and wind energy generation for one week [32].

$\theta_i \triangleq Vp_i^{\max} + s_{i,\max}^-$ *and* $Q_i^c(0) = 0, \forall i, c$. *Then, under the SCMA algorithm, we have the following:*

1. The thermal energy queues satisfy the following for all time t under any arbitrary $(r_i(t), p_i(t), W_j^c(t), \forall c, i, j)$ process:

$$0 \le S_i(t) \le S_i^{\max}, \quad \forall i. \tag{38}$$

2. If the vector $(r_i(t), p_i(t), W_j^c(t), \forall c, i, j)$ is i.i.d. over periods, and if there exists a constant δ such that $\boldsymbol{\omega} + \delta \mathbf{1} \in \boldsymbol{\Omega}$, then the total batch workload queue length satisfies the following under any arbitrary $(r_i(t), p_i(t), W_j^c(t), \forall c, i, j)$ process:

$$\overline{Q} \le \frac{A_1 V + A_2}{\delta}. \tag{39}$$

The time-average expected total operating cost under the SCMA algorithm is within bound A_2/V of the optimal value:

$$\bar{g}^{\mathrm{SCMA}} \le \bar{g}^* + A_2/V, \tag{40}$$

where $\bar{g}^*$ is the optimal cost achieved by any feasible control policy that can stabilize the queues, and A_1, A_2 are constants given by (26) and (31), respectively.

Proof. See the supplementary document which is available online. □

6 Numerical Evaluation

In the remainder of the paper, we evaluate the performance of the SCMA under realistic traces. Our goal is threefold: (i) to illustrate the benefits by jointly considering the thermal storage, delay-tolerant workloads, and geographical load balancing in data centers to reducing the operating cost; (ii) to understand the impacts of various parameters on the control decisions made by SCMA; and (iii) to understand the trade-offs among cost reduction, workload delay, and thermal storage capacity enabled by the SCMA.

6.1 Experimental Setup

In this part, we introduce the default settings that are used throughout the evaluations unless otherwise stated. The length of a control period is 10 minutes and the time-horizon in the evaluations is 4000 periods.

6.1.1 Data Center Descriptions

We consider four data centers, one at the geographic center of each city that is known to have Google data centers: New York, Palo Alto, Chicago, and Houston. Moreover, we assume that there is a proxy located near each data center. The bandwidth cost b_{ij} between proxies and data centers is set to be proportional to the distances between cities and comparable to the energy cost. The number of available active servers in each data center is taken to be $\mathrm{IT}_i = 350$. The energy consumption of each server during one period at idle and busy state are set to be $P_i^{idle} = 100\ \mathrm{W} \times 1/6\ \mathrm{h}, \forall i$ and $P_i^{busy} = 250\ \mathrm{W} \times 1/6\ \mathrm{h}, \forall i$, respectively. Without loss of generality, the processing speed of each server is assumed to be $\mu_i = 1, \forall i$. The cooling efficiency of each data center is set to be the average value of the data center industry as $\beta_i = 1, \forall i$. Notice that here, we assume the homogenous settings of data centers in order to make the analysis of the impacts of other factors (e.g., energy prices, renewable availability, bandwidth cost) more explicitly.

6.1.2 Workload Description

As with [16], we choose MapReduce [23], which is a popular type of computation-intensive workloads in data centers, as the representative of delay-tolerant workloads. We use the historical Hadoop (an open source implementation of MapReduce) traces on a 600-machine cluster at Facebook [29] to calculate the average 10-min workload arrivals. A portion of the workload trace during one day is shown in Fig. 3a. The workload arrivals to each proxy are shifted according to the time zone. We assume there are two types of jobs, with job length $d_c = \{1, 0.5\}$ and communication demand $n_c = \{1, 0.5\}$. We assume that half of the arriving requests belong to type 1 and the other half belong to type 2. The workload traces are scaled such that the peak demand can be supported entirely by its own data center without delay.

6.1.3 Energy Price Description

We use the day-ahead hourly locational marginal prices (LMPs) in wholesale electricity markets at the above four data center locations. They are obtained from the publicly available government sources [30], [31]. A portion of the hourly electricity prices during the first 24 hours at these locations is shown in Fig. 3b.

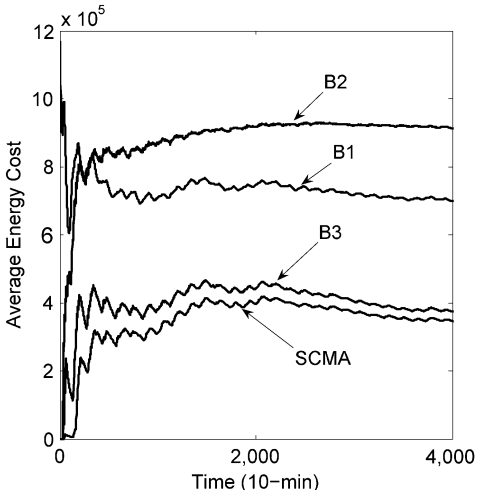


Fig. 4. Average energy cost (in unit of dollars) comparison between SCMA and baseline schemes.

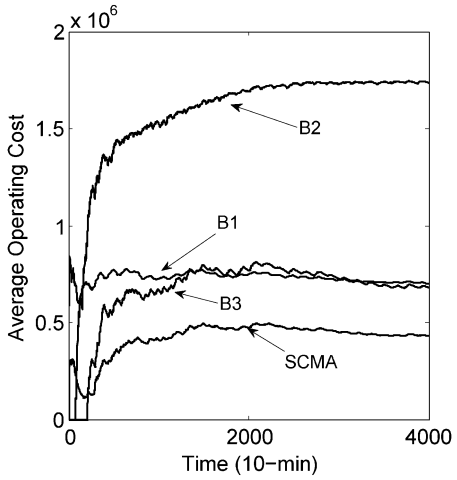


Fig. 5. Average operating cost (in unit of dollars) comparison between SCMA and baseline schemes.

6.1.4 Renewable Energy Description

We consider on-site wind generation at two locations (New York and Chicago) and on-site solar generation at the other two locations (Palo Alto and Houston). The traces of wind and solar sources are obtained from [32] that has wind speed and solar irradiance measurements every 10 minutes. The traces are scaled properly so that the average renewable production can meet half of the average power consumption at each data center. A portion of solar and wind energy at two locations during the first two days is depicted in Fig. 3c.

6.1.5 Thermal Storage Description

We assume each data center has installed a thermal storage system. The maximum charge (discharge) rate $s^{+}_{i,\max}(s^{-}_{i,\max})$ is set to be the peak cooling energy consumption during one period. The round-trip charging efficiency η_i is set to be 0.8. The storage operating cost factor $\tilde{\eta}_i$ and the storage capacity $S_i^{\max}$ are parameters of which the impact on the performance of SCMA will be investigated.

6.1.6 Algorithm Benchmarks

To provide benchmarks for the performance of SCMA, we compare it with the following three baselines that either approximate the current practice [33], or are proposed by some recent work [16], [18].

- **Baseline 1 (B1): No workload scheduling, no storage**. In this approach, the workloads are routed to the nearest data center and served immediately without any delay. This scheme is employed by many companies in practice so as to serve all the incoming workloads as soon as possible without any consideration on energy price or renewable energy availability [33].
- **Baseline 2 (B2): Renewable-oblivious workload scheduling, no storage**. This approach is very similar to that proposed in the recent work [16], which investigates jointly routing and scheduling delay-tolerant workloads in multiple data centers to leverage the opportunity of time-varying energy price. However, no renewable energy or thermal storage is taken into account in this scheme.
- **Baseline 3 (B3): Renewable-aware workload scheduling, no storage**. This approach is proposed in the recent work [18] for cost minimization in a single data center. Renewable energy availability and time-varying energy price are considered but without thermal storage. We modify its algorithm to incorporate routing decisions.

6.2 Numerical Results

The evaluation of SCMA will be organized as the following aspects.

6.2.1 Cost Savings

Note that prior studies mainly focus on reducing energy cost without considering the bandwidth cost for workload routing. To evaluate the energy cost saving due to our algorithm SCMA by leveraging delay-tolerant workloads, thermal storage, and geographical load balancing, we first assume that the bandwidth cost $b_{ij} = 0, \forall i, j$ so that we can focus on the energy cost. Since the performances of SCMA, **B2**, and **B3** all depend on the parameter V, for fair comparison, we choose the parameter V in different schemes such that the average delay of queued workloads in these schemes are equal. Note that **B1** has no delay. Moroever, the SCMA is under the following parameter settings: the storage operating cost factor $\gamma_i = 10, \forall i$ and the storage capacity $S_i^{\max}$ is assumed to be able to support the average cooling demand of a data center for 10 hours. The result is shown in Fig. 4. From the figure, we can observe that SCMA outperforms all benchmark schemes. Specifically, by comparing SCMA with **B3**, we can observe that thermal storage can indeed help reduce the total electricity cost. Moreover, although **B2** considers the time-varying electricity price, it is renewable-oblivious and tries to serve workloads only when the electricity price is low enough. Therefore, it wastes a lot of renewable energy and performs the worst. This shows the importance of renewable-aware workload management in data centers with on-site renewable generation. Finally, by comparing **B3** with **B1**, we can see the advantage of delay-tolerant workloads in improving renewable energy utilization and reducing electricity cost.

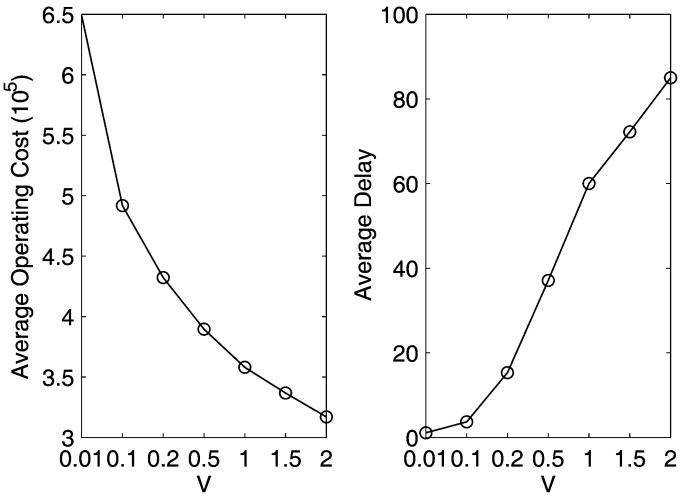


Fig. 6. Average total operating cost (in unit of dollars) and delay performance (in unit of control periods) of SCMA with different V.

Then, we compare our algorithm SCMA with the baseline schemes above while considering the bandwidth cost. Obviously, the bandwidth cost of **B1** is zero by our assumption because it always routes all workloads to the nearest data centers. The result of the average operating cost for the other three algorithms is shown in Fig. 5. By taking into account the different bandwidth costs between proxies and data centers, our algorithm can achieve the largest total operating cost saving. The importance of network-awareness is clear from the figure above. Note that **B1** and **B3** have similar operating cost since the bandwidth cost of **B1** is minimum although its energy cost is higher than that of **B3**. **B2** has the worst performance of operating cost since both the energy cost and the bandwidth cost are the highest among all schemes.

6.2.2 Trade-Off Between Cost and Delay

In this part, we focus on the trade-offs among delay, total operating cost, and thermal storage capacity in SCMA. We choose different V and observe the corresponding total operating cost and average workload delay in SCMA. The result is shown in Fig. 6. As we can observe from the figure, with the increase of the parameter V, SCMA can get lower total operating cost with trade-offs in the workload delay, which validates the analytical performance results in Theorem 2. Note that by selecting a larger V, SCMA would be more aggressively minimizing the operating cost, which may delay more jobs to be served later when enough renewable energy is available or energy price is low, causing larger queuing delay.

6.2.3 Impact of Storage Cost

To evaluate the impact of thermal storage cost on the operating cost saving, we fix parameters V and $S_i^{\max}, \forall i$, and evaluate SCMA under different $\gamma_i = [0, 5, 10, 15, 20, 25, 30], \forall i$. The result is shown in Fig. 7. We can observe that with the increase of thermal storage cost factor γ_i, the operating cost saving is smaller. When γ_i is very large, SCMA does not use the thermal storage at all. However, even in this case, there is still cost saving compared with **B1** because of the delay-tolerant workload scheduling and geographical load balancing.

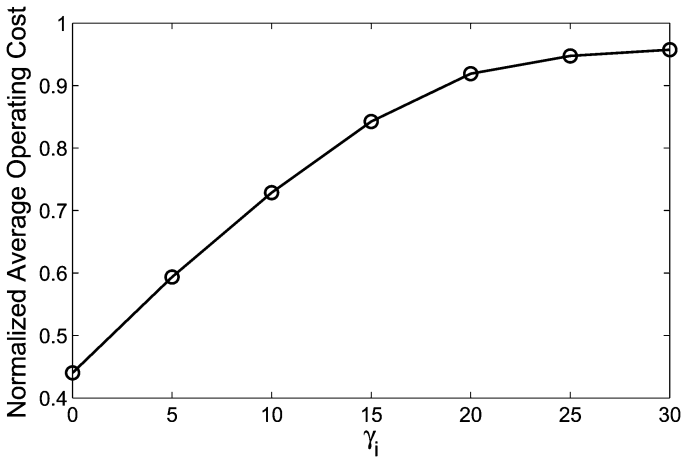


Fig. 7. Impact of thermal storage cost factor γ_i on the average cost (normalized to the average cost of **B1**).

7 Concluding Remarks

In this paper, we studied the problem of joint network-aware workload routing, delay-tolerant workload scheduling, and thermal storage management to improve the renewable energy utilization and reduce the time-average total operating cost in data centers. We design an online control algorithm called SCMA and demonstrate its effectiveness through both analytical analysis and numerical evaluations. Moreover, SCMA provides an explicit trade-off between cost saving and workload delay.

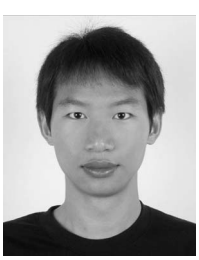
Acknowledgment

This work was supported in part by the U.S. National Science Foundation under Grants ECCS-1129061, ECCS-1129062, CNS-1239274, CNS-1343356, and Eckis Professor endowment at the University of Florida.

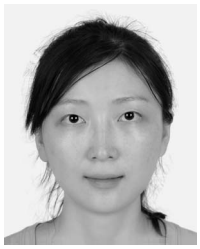
References

[1] J. Koomey, *Growth in Data Center Electricity Use 2005 to 2010*. Burlingame, CA, USA: Analytics Press, 2011.
[2] P.X. Gao, A.R. Curtis, B. Wong, and S. Keshav, ''It's Not Easy Being Green,'' in *Proc. ACM SIGCOMM*, Aug. 2012, pp. 211-222.
[3] Z. Liu, M. Lin, A. Wierman, S. Low, and L. Andrew, ''Greening Geographical Load Balancing,'' in *Proc. ACM SIGMETRICS*, 2011, pp. 233-244.
[4] Z. Liu, M. Lin, A. Wierman, S. Low, and L. Andrew, ''Geographical Load Balancing with Renewables,'' in *Proc. GreenMetrics*, 2011, pp. 1-5.
[5] Apple and the Environment. [Online]. Available: http://www.apple.com/environment/renewable-energy/
[6] M.J. Neely, *Stochastic Network Optimization With Application to Communication and Queueing Systems*. San Rafael, CA, USA: Morgan & Claypool Publishers, 2010.
[7] Google's PPAs: What, How, and Why. [Online]. Available: http://www.google.com/about/datacenters/energy.html
[8] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang, M. Marwah, and C. Hyser, ''Renewable and Cooling Aware Workload Management for Sustainable Data Centers,'' in *Proc. ACM SIGMETRICS/PERFORMANCE*, 2012, pp. 175-186.
[9] I. Goiri, K. Le, T.D. Nguyen, J. Guitart, J. Torres, and R. Bianchini, ''Greenhadoop: Leveraging Green Energy in Dat-Processing Framework,'' in *Proc. EuroSys*, 2012, pp. 57-70.
[10] R. Urgaonkar, B. Urgaonkary, M.J. Neely, and A. Sivasubramaniam, ''Optimal Power Cost Management Using Stored Energy in Data Centers,'' in *Proc. ACM SIGMETRICS*, San Jose, CA, USA, June 2011, pp. 221-232.

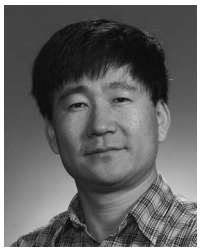
[11] S. Govindan, A. Sivasubramaniam, and B. Urgaonkar, ''Benefits and Limitations of Tapping Into Stored Energy for Datacenters,'' in *Proc. ISCA*, 2011, pp. 341-352.
[12] Y. Guo, Z. Ding, Y. Fang, and D. Wu, ''Cutting Down Electricity Cost in Internet Data Centers by Using Energy Storage,'' in *Proc. IEEE GLOBECOM*, 2011, pp. 1-5.
[13] Y. Wang, X. Wang, and Y. Zhang, ''Leveraging Thermal Storage to Cut the Electricity Bill for Datacenter Cooling,'' in *HotPower*, 2011, pp. 1-5.
[14] M. Lin, A. Wierman, L.L.H. Andrew, and E. Thereska, ''Dynamic Right-Sizing for Power-Proportional Data Centers,'' in *Proc. IEEE INFOCOM*, 2011, pp. 1098-1106.
[15] M. Lin, Z. Liu, A. Wierman, and L.L.H. Andrew, ''Online Algorithms for Geographical Load Balancing,'' in *Proc. IGCC*, 2012, pp. 1-10.
[16] Y. Yao, L. Huang, A. Sharma, L. Golubchik, and M. Neely, ''Data Centers Power Reduction: A Two Time Scale Approach for Delay Tolerant Workload,'' in *Proc. IEEE INFOCOM*, 2012, pp. 1431-1439.
[17] D. Xu and X. Liu, ''Geographical Trough Filling for Internet Datacenters,'' in *Proc. IEEE INFOCOM Mini-Conf.*, 2012, pp. 2881-2885.
[18] S. Ren, Y. He, and F. Xu, ''Provably-Efficient Job Scheduling for Energy and Fairness in Geographically Distributed Data Centers,'' in *Proc. IEEE ICDCS*, 2012, pp. 22-31.
[19] Y. Guo, Y. Gong, Y. Fang, P.P. Khargonekar, and X. Geng, ''Optimal Power and Workload Management for Green Data Centers with Thermal Storage, IEEE GLOBECOM, Atlanta, GA, USA, 2013.
[20] X. Meng, V. Pappas, and L. Zhang, ''Improving the Scalability of Data Center Networks with Traffic-Aware Virtual Machine Placement,'' in *Proc. IEEE INFOCOM*, San Diego, CA, USA, Mar. 2010, pp. 1-9.
[21] N. Buchbinder, N. Jain, and I. Menache, ''Online Job-Migration for Reducing the Electricity Bill in the Cloud,'' in *Proc. NETWORKING*, 2011, pp. 172-185.
[22] M. Alicherry and T.V. Lakshaman, ''Network Aware Resource Allocation in Distributed Clouds,'' in *Proc. IEEE INFOCOM*, Orlando, FL, USA, Mar. 2012, pp. 963-971.
[23] J. Dean and S. Ghemawat, ''Mapreduce: Simplified Data Processing on Large Clusters,'' in *Proc. OSDI*, 2004, pp. 137-149.
[24] 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. IEEE INFOCOM*, 2010, pp. 1-9.
[25] Amazon EC2 Spot Instances. [Online]. Available: http://aws.amazon.com/ec2/spot-instances/
[26] C. Ren, D. Wang, B. Urgaonkar, and A. Sivasubramaniam, ''Carbon-Aware Energy Capacity Planning for Datacenters,'' in *Proc. IEEE MASCOTS*, 2012, pp. 391-400.
[27] S.B. Peterson, J.F. Whitacre, and J. Apt, ''The Economics of Using Plug-In Hybrid Electric Vehicle Battery Packs for Grid Storage,'' *J. Power Sources*, vol. 195, no. 8, pp. 2377-2384, Apr. 2010.
[28] A. Gandhi, M. Harchol-Balter, R. Das, and C. Lefurgy, ''Optimal Power Allocation in Server Farms,'' in *Proc. 11th ACM SIGMETRICS*, Seattle, WA, USA, Aug. 2009, pp. 157-168.
[29] Y. Chen, A. Ganapathi, R. Griffith, and R. Katz, ''The Case for Evaluating Mapreduce Performance Using Workload Suites,'' in *Proc. IEEE MASCOTS*, 2011, pp. 390-399.
[30] Federal Energy Regulatory Commission. [Online]. Available: http://www.ferc.gov/
[31] United States Energy Information Administration. [Online]. Available: http://www.eia.gov/
[32] NREL: Measurement and Instrumentation Data Center. [Online]. Available: http://www.nrel.gov/midc/
[33] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, ''Cutting the Electric Bill for Internet-Scale Systems,'' in *Proc. ACM SIGCOMM*, 2009, pp. 123-134.



Yuanxiong Guo received his BEng degree from the Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2009. He has been working towards the PhD degree at the Department of Electrical and Computer Engineering at University of Florida, Gainesville, USA since August 2010. His current research interests are in the area of cyber-physical systems including smart grids, sustainable data centers, and cloud computing. He is a recipient of the Best Paper Award from IEEE GLOBECOM 2011, Houston, TX, USA. He is a Student Member of the IEEE.



Yanmin Gong received her BEng degree in electrical engineering from Huazhong University of Science and Technology, Wuhan, China, in 2009, and MSc degree in electrical engineering from Tsinghua University, Beijing, China, in 2012. She has been working towards the PhD degree at the Department of Electrical and Computer Engineering at University of Florida, Gainesville, USA, since August 2012. Her current research interests are in the area of optimization, security and privacy in cyber-physical systems and mobile computing. She is a Student Member of the IEEE.



Yuguang Fang received the BS/MS degree in Mathematics from Qufu Normal University, China, in 1987, a PhD degree in Systems Engineering from Case Western Reserve University, USA, in 1994 and a PhD degree in Electrical Engineering from Boston University, USA, in 1997. He is currently a professor in the Department of Electrical and Computer Engineering at University of Florida, USA. He held a University of Florida Research Foundation (UFRF) Professorship from 2006 to 2009, a Changjiang Scholar Chair Professorship with Xidian University, China, from 2008 to 2011, and a Guest Chair Professorship with Tsinghua University, China, from 2009 to 2012. He has published over 350 papers in refereed professional journals and conferences. He received the National Science Foundation Faculty Early Career Award in 2001 and the Office of Naval Research Young Investigator Award in 2002. He has also received a 2010-2011 UF Doctoral Dissertation Advisor/Mentoring Award, 2011 Florida Blue Key/UF Homecoming Distinguished Faculty Award and the 2009 UF College of Engineering Faculty Mentoring Award. Dr. Fang is a Fellow of IEEE. He is currently the Editor-in-Chief of *IEEE Transactions on Vehicular Technology*. He was the Editor-in-Chief of IEEE Wireless Communications (2009-2012) and serves/served on several editorial boards of technical journals including *IEEE Transactions on Mobile Computing* (2003-2008, 2011-present), *IEEE Network* (2012-present), *IEEE Transactions on Communications* (2000-2011), *IEEE Transactions on Wireless Communications* (2002-2009), *IEEE Journal on Selected Areas in Communications* (1999-2001), *IEEE Wireless Communications Magazine* (2003-2009), and *ACM Wireless Networks* (2001-2013). He is currently serving as the Technical Program Committee Co-Chair for IEEE INFOCOM'2014.

Pramod P. Khargonekar received BTech degree in electrical engineering from the Indian Institute of Technology, Bombay, India and MS degree in mathematics and PhD degree in electrical engineering from the University of Florida, USA. He has held faculty positions at the University of Minnesota, USA and The University of Michigan, USA. He was Chairman of the Department of Electrical Engineering and Computer Science at Michigan from 1997 to 2001, and also held the title Claude E. Shannon Professor of Engineering Science there. From 2001 to 2009, he was Dean of the College of Engineering and is now Eckis Professor Electrical and Computer Engineering at the University of Florida. He served as Deputy Director for Technology at the U.S. Department of Energy's Advanced Research Projects Agency C Energy (ARPA-E). He is currently serving the U.S. National Science Foundation as Assistant Director for Engineering. His current research interests are focused on renewable energy and electric grid, neural engineering, and systems and control theory. He is a recipient of the NSF Presidential Young Investigator Award, the American Automatic Control Council's Donald Eckman Award, the Japan Society for Promotion of Science Fellowships, and a Distinguished Alumnus Award and Distinguished Service Award from the Indian Institute of Technology, Bombay. He is a co-recipient of the IEEE W.R.G. Baker Prize Award, the IEEE CSS George S. Axelby Best Paper Award, and the AACC Hugo Schuck Best Paper Award. He was a Springer Professor at the University of California, Berkeley, USA in 2010. He is a Fellow of IEEE and is on the list of Web of Science Highly Cited Researchers.

Xiaojun Geng received the BSc and MSc degrees in astronautical engineering from Northwestern Polytechnic University, Xi'an, China, in 1993 and 1996, respectively, and received two PhD degrees in electrical and computer engineering, one from Shanghai Jiao Tong University, Shanghai, China, in 1999, and the other from the University of Florida, Gainesville, USA, in 2003. In 2003, Dr. Geng joined the Department of Electrical and Computer Engineering as an Assistant Professor at California State University, Northridge, USA where in 2009, she became an Associate Professor. In 2013, she joined the Department of Electrical and Computer Engineering as an Assistant Professor in the University of West Florida, USA. Her research interests include control of discrete-event systems, coordination of multi-agent systems, and power management of smart grid. She is a member of the IEEE.