Cutting Down Electricity Cost in Internet Data Centers by Using Energy Storage

Yuanxiong Guo, Zongrui Ding, Yuguang Fang, Dapeng Wu
Department of Electrical and Computer Engineering,
University of Florida, Gainesville, FL 32611, USA
Email: {guoyuanxiong@, dingzr@, fang@ece., wu@ece.}ufl.edu

***Abstract*—Electricity consumption comprises a significant fraction of total operating cost in data centers. System operators are required to reduce electricity bill as much as possible. In this paper, we consider utilizing available energy storage capability in data centers to reduce electricity bill under real-time electricity market. Laypunov optimization technique is applied to design an algorithm that achieves an explicit tradeoff between cost saving and energy storage capacity. As far as we know, our work is the first to explore the problem of electricity cost saving using energy storage in multiple data centers by considering both time-diversity and location-diversity of electricity price.**

***Index Terms*—Cloud computing, electricity cost, data center, energy storage, Laypunov optimization**

I. Introduction

With the popularity of cloud computing [1], more and more data centers are envisioned to be built in the future in order to meet the growing demand of large-scale computing resources. It is common for a cloud service provider to have multiple data centers each having hundreds of thousands of servers. Those data centers are geographically distributed for reliability as well as performance improvement. A critical issue in the operations of those data centers is the energy consumption, including both servers and air conditioning. According to the estimation from [2], large companies such as Google and Microsoft pay tens of millions of dollars for just electricity cost every year, and 30%-50% percentage of operational expenses in data centers come from electricity.

The huge energy consumption in data centers has motivated a lot of research works toward energy-efficient data centers. Power Usage Effectiveness (PUE), which measures the ratio of total building power to IT power, i.e., the power consumed by the actual computing equipment, is a popular metric to judge the energy-efficiency of a data center. Various engineering techniques, such as virtualization, advanced cooling and DC power have been developed to reduce PUE (See [3], [4] for a survey on these issues). These works focus on reducing the energy usage of the data centers.

Another stream to handle the energy consumption issue in data centers is based on the observation that electricity price is different across different time and locations under real-time electricity market. Qureshi et al. are the first to discuss the opportunity of utilizing such electricity price diversity to reduce total electricity cost by distributing more traffic to data centers with low electricity price [2]. Rao et al. investigate the problem of total electricity cost for data centers in multi-electricity-market environment subject to QoS guarantee and propose a linear programming formulation to approximately solve it [5]. Other works [6], [7] consider a similar problem with some improvements on delay and energy consumption model. These works focus on directly reducing total electricity cost using electricity price diversity. However, none of the aforementioned works consider using available energy storage capacity, typically UPS unit, in data center to further reduce the electricity cost.

Data centers have uninterrupted power supply (UPS) units to keep them powered using stored energy in case of electricity failure, before the backup diesel generation can start up and provide power. Usually, the transition to use diesel generation takes only 10-20 seconds while UPS units have enough capacity to power data center at its maximum power need between 5-30 minutes. These excess energy storage capacity can be used to save electricity cost by the simple intuition of charging when outside electricity price is low while discharging when outside electricity price is high. Our work is mainly motivated by [8], which considers the case of a single data center with energy storage under time-varying electricity price. Different from previous works, our work considers the total electricity cost minimization of an Internet-service provider having multiple data centers with energy storage under both time-varying and location-varying electricity price.

The rest of this paper is organized as follows. Section II presents the models of power consumption cost in multiple data centers, which is formulated as a stochastic programming to minimize the time-average expected electricity cost. Section III solves the optimization problem by first considering a relaxed problem and then, using the Laypunov optimization technique to design an algorithm to approximately solve the original problem. Section IV gives the performance analysis, and Section V concludes the paper.

II. Model and Formulation

In this section, we describe our models for workload, battery, delay and electricity cost. Then we present our control objective to minimize the time-average expected electricity cost. The block diagram of our system model is shown in Figure 1, which is described in detail as follows. We consider a time-slotted system.

This work was supported in part by the U.S. National Science Foundation under grants CNS-0721744.

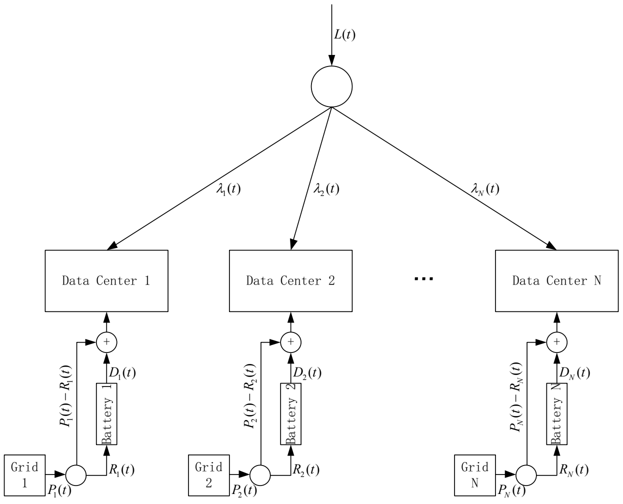

Fig. 1. Block diagram for system model

A. Workload Model

Let N denote the total number of geographically located data centers, where each data center i has a total number of M_i homogeneous servers. We assume that there is a traffic aggregator (such as DNS) which is responsible for distributing the total incoming workload to different data centers. The incoming traffic is a Poisson process with average arrival rate $L(t)$ at time slot t. Let $\lambda_i(t)$ denote the average traffic rate distributed into data center i at each slot t, $\vec{\lambda}(t) = (\lambda_1(t), \lambda_2(t), \ldots, \lambda_N(t))$, then we have

$$\sum_{i=1}^{N} \lambda_i(t) = L(t), \tag{1}$$

$$\lambda_i(t) \geq 0, \ \forall i \in [1, \ldots, N] \tag{2}$$

B. Battery Model

There are limited times of charging/discharing cycles for each battery. Besides, conversion loss occurs both in charging and discharging processes. Stored energy is also subject to dissipation with time. For notational simplicity, we ignore these factors. We assume that the battery has capacity $E_{i,max}$ in each data center i and data centers operate independently of each other.

Let $P_i(t)$ be the external energy drawn from power grid, $R_i(t)$ be the charging energy into the battery, and $D_i(t)$ be the discharging energy from the battery in each time slot t at data center i. The total amount of energy supplied to support current traffic in time slot t at data center i is given by $P_i(t) - R_i(t) + D_i(t)$. We assume $P_i(t), R_i(t)$ and $D_i(t)$ are nonnegative and upper bounded by $P_{i,max}, R_{i,max}$ and $D_{i,max}$ respectively, i.e.,

$$0 \leq R_i(t) \leq R_{i,max}, \ 0 \leq D_i(t) \leq D_{i,max}, \tag{3}$$

$$0 \leq P_i(t) \leq P_{i,max}. \tag{4}$$

The update equation of battery energy level $E_i(t)$ at data center i is given by

$$E_i(t+1) = E_i(t) + R_i(t) - D_i(t). \tag{5}$$

More complicated model of battery can be easily incorporated into our model without affecting the following analysis. As we are only interested in total charging or discharging amount during each time slot, we assume, without loss of generality, that charging and discharging can not be done simultaneously. In other words, in each time slot t, we have

$$R_i(t) > 0 \Rightarrow D_i(t) = 0, D_i(t) > 0 \Rightarrow R_i(t) = 0. \tag{6}$$

Battery energy level should be always nonnegative and can not exceed the battery capacity. So in each time slot t, we need to ensure that for each data center i,

$$0 \leq E_i(t) \leq E_{i,max}. \tag{7}$$

From constraints (5), (6), and (7), we get the following equivalent constraints in each slot t:

$$0 \leq R_i(t) \leq \min\{R_{i,max}, E_{i,max} - E_i(t)\}, \tag{8}$$

$$0 \leq D_i(t) \leq \min\{D_{i,max}, E_i(t)\}. \tag{9}$$

C. Delay Model

As in [5], we use a $M/M/n$ queuing model to analyze the average waiting time in data center i when traffic rate is $\lambda_i(t)$ and there are $m_i(t)$ active servers, each with service rate μ_i. Note that $m_i(t)$ is an integer variable and has a maximum value M_i at each data center i. Using the result from queuing theory [9], the average waiting time $W_i(t)$ is $\frac{1}{m_i(t)\mu_i - \lambda_i(t)} P_Q$ where P_Q is the queuing probability. Without loss of generality in a data center, we assume the servers are always busy if turned on. Hence, $P_Q = 1$ and $W_i(t) = \frac{1}{m_i(t)\mu_i - \lambda_i(t)}$. To meet the quality of service of customers, a maximum average waiting time $W_{i,max}$ exists for each data center i. Therefore, we have the following delay constraint:

$$\frac{1}{m_i(t)\mu_i - \lambda_i(t)} \leq W_{i,max}, \tag{10}$$

$$0 \leq m_i(t) \leq M_i, m_i \in \mathbb{N}. \tag{11}$$

D. Electricity Consumption and Cost

We are only interested in server energy consumption cost without considering cooling cost. We do not consider power-proportional issue here either, that is, we assume each server at data center i consumes either its maximum power H_i when active or zero when inactive. Hence, we have the following equation:

$$P_i(t) - R_i(t) + D_i(t) = m_i(t)H_i. \tag{12}$$

As analyzed in [2], electricity price in real-time electricity market has both time-diversity and location-diversity. At each data center i, we have time-varying electricity price $C_i(t)$ in unit of energy with maximum value $C_{i,max}$. Moreover, $C_i(t)$ and $P_i(t)$ are independent. Different data centers may have different prices at the same time due to location difference. Hence, the total electricity cost of N data centers in time slot t is $\sum_{i=1}^{N} C_i(t)P_i(t)$.

E. Problem Formulation

In this paper, we are interested in minimizing the time-average expected electricity cost. Based on the above models, our problem can be formulated as the following stochastic optimization, called **Problem One**:

$$\min_{\vec{\lambda}(t),\vec{m}(t),\vec{P}(t),\vec{R}(t),\vec{D}(t)} Q = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\mathbb{E}\{P_i(t)C_i(t)\}, \tag{13}$$

subject to constraints (1), (2), (3), (4), (6), (8), (9), (10), (11), (12) where the constraints are for each time slot t and data center $i = 1, \ldots, N$.

III. Proposed Solution

From the above formulation, there are two key control designs: (i) traffic distribution decision $\vec{\lambda}(t)$; and (ii) external energy drawn from power grid $\vec{P}(t)$ at each time slot. Once they are determined, other control variables can be easily derived. As mentioned before, the statistics of $L(t)$ and $\vec{C}(t)$ may not be known and we need to design an optimal control algorithm under uncertainty. We use the recently developed technique of Lyapunov optimization [10]. The algorithm we propose can achieve the range of $O(1/V)$ within the optimal objective value, where V is a parameter related to the battery capacity of each data center i. One salient feature of our algorithm is that it does not need any future knowledge of the system and can be implemented online.

Define the time-average value of charging and discharging at data center i under any feasible control policy of **Problem One** respectively as follows:

$$\overline{R_i} = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\{R_i(t)\}, \overline{D_i} = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\{D_i(t)\}. \tag{14}$$

Since battery energy level is evolving according to Eq. (5), summing over all $t \in 0, 1, 2, \ldots, T-1$, taking expectation of both sides, dividing both sides with T and taking $T \to \infty$ yields $\overline{R_i} = \overline{D_i}$. Hence we have the following relaxed problem, called **Problem Two**:

$$\min_{\vec{\lambda}(t),\vec{m}(t),\vec{P}(t),\vec{R}(t),\vec{D}(t)} Q = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\mathbb{E}\{P_i(t)C_i(t)\}, \tag{15}$$

subject to constraints (1), (2), (3), (4), (6), (10), (11), (12),

$$\overline{R_i} = \overline{D_i},$$

where the constraints are for each time slot t and data center $i = 1, \ldots, N$.

Denote the optimal objective value of **Problem One** as Q^{OPT} and the optimal objective value of **Problem Two** as Q^{REL}. As discussed before, any feasible solution to **Problem One** is also a feasible solution to **Problem Two**. Hence, $Q^{REL} \le Q^{OPT}$. From the framework of Laypunov optimization [10], we have the following theorem. The proof is similar to that in [8] and follows from the framework in [10], [11]. It is omitted here for brevity.

***Theorem** 1:* If $\vec{C}(t)$ and $L(t)$ are $i.i.d.$ over slots, then there exists a stationary, randomized policy that takes control decisions $\vec{\lambda}^{stat}(t), \vec{m}^{stat}(t), \vec{P}^{stat}(t), \vec{R}^{stat)}(t), \vec{D}^{stat}(t)$ every slot purely as a function (possibly randomized) of the current workload $L(t)$ and electricity price $\vec{C}(t)$ while satisfying the constraints of **Problem Two** and providing the following guarantees:

$$\mathbb{E}\{R_i^{stat}(t)\} = \mathbb{E}\{D_i^{stat}(t)\},\ \mathbb{E}\{\sum_{i=1}^{N} P_i^{stat}(t)C_i(t)\} = Q^{REL},$$

where the expectations above are with respect to the stationary distribution of $L(t), \vec{C}(t)$ and the randomized control decisions.

In order to derive such a policy, we need to know the statistical distribution of all combination of $\vec{C}(t)$ and $L(t)$, which usually has the problem of "Curse of Dimensionality" [12] if solved by dynamic programming. Moreover, this control policy may not be a feasible solution to **Problem One**. Instead, we use the existence of such a policy to help us design our control policy that meets all constraints of **Problem One** and derive the performance bound of our algorithm.

Before presenting our algorithm, we define a virtual queue $S_i(t)$ as a shifted version of battery energy level $E_i(t)$ in time slot t at each data center i as follows:

$$S_i(t) = E_i(t) - VC_{i,max} - D_{i,max} \tag{16}$$

where $V \ge 0$ is a constant control parameter which affects the distance to the optimal value and is related to the battery capacity. $\vec{S}(t) = (S_1(t), S_2(t), \ldots, S_N(t))$ is used to ensure that the constraint (7) is satisfied in our algorithm as illustrated later. According to Eq. (5) of $E_i(t)$, we have the same update equation for $S_i(t)$ at each data center i:

$$S_i(t+1) = S_i(t) + R_i(t) - D_i(t). \tag{17}$$

Our proposed algorithm works as follows:

```
1 At the beginning of each time slot $t$, observe $\vec{C}(t)$, $\vec{S}(t)$, and $L(t)$
2 Choose control action $\vec{\lambda}^*(t), \vec{m}^*(t)$, and $\vec{P}^*(t)$ as the solution
  to the following optimization problem, called Problem Three:

      $\min \sum_{i=1}^{N}\{-S_i(t)m_i(t)H_i + S_i(t)P_i(t) + VP_i(t)C_i(t)\}$,  (18)

  subject to constraints (1), (2), (3), (4), (6), (10), (11), (12),
  where the constraints are for each time slot $t$ and data center
  $i = 1, \ldots, N$
3 . Choose the charging and discharging values as follows:

      $R_i^*(t) = \begin{cases} P_i^*(t) - m_i^*(t)H_i & \text{if } P_i^*(t) > m_i^*(t)H_i \\ 0 & \text{else} \end{cases}$

      $D_i^*(t) = \begin{cases} m_i^*(t)H_i - P_i^*(t) & \text{if } P_i^*(t) < m_i^*(t)H_i \\ 0 & \text{else} \end{cases}$

4 Implement them and update queue variables $\vec{S}(t)$
```

For each time slot t, **Problem Three** is a mixed-integer linear programming. However, in practice, a data center usually contains thousands of servers, of which a large fraction are active. Hence, we can relax the integer constraint on $m_i(t)$, round the resulting solution without significant cost penalties and get a simple linear programming optimization problem, which can be solved efficiently in polynomial time using interior-point method [13]. The objective in **Problem Three** is derived from the proof process, which will be clear in Section IV.

The optimal solution to **Problem Three** has the following property that is useful for the future analysis of algorithmic performance:

Lemma 1: The optimal solution to **Problem Three** has the following properties:

- If $S_i(t) > 0$, the optimal solution always choose $R_i^*(t) = 0$.
- If $S_i(t) < -VC_{i,max}$, the optimal solution always choose $D_i^*(t) = 0$.

Proof: Part 1: For each data center i, when $S_i(t) > 0$, suppose $R_i^*(t) > 0$, then we have $D_i^*(t) = 0$ and $P_i^*(t) > m_i^*(t)H_i$. The value of the objective is

$$\sum_{j\neq i}\{-S_i(t)m_j^*(t)H_j + (S_j(t)+VC_j(t))P_j^*(t)\}+$$
$$\{-S_i(t)m_i^*(t)H_i + (S_i(t)+VC_i(t))P_i^*(t)\} >$$
$$\sum_{j\neq i}\{-S_i(t)m_j^*(t)H_j + (S_j(t)+VC_j(t))P_j^*(t)\}+$$
$$\{-S_i(t)m_i^*(t)H_i + (S_i(t)+VC_i(t))m_i^*(t)H_i\}$$

where the last step follows from the facts that $S_i(t)+VC_i(t) > 0$ and $P_i^*(t) > m_i^*(t)H_i$. In this case, the objective with all power drawn from the gird is smaller. Hence, when $S_i(t) > 0$, $R_i^*(t)$ can not be greater than zero.

Part 2: When $S_i(t) < -VC_{i,max}$, suppose $D_i^*(t) > 0$, then we have $R_i^*(t) = 0$ and $P_i^*(t) < m_i^*(t)H_i$. The value of the objective is

$$\sum_{j\neq i}\{-S_i(t)m_j^*(t)H_j + (S_j(t)+VC_j(t))P_j^*(t)\}+$$
$$\{-S_i(t)m_i^*(t)H_i + (S_i(t)+VC_i(t))P_i^*(t)\} >$$
$$\sum_{j\neq i}\{-S_i(t)m_j^*(t)H_j + (S_j(t)+VC_j(t))P_j^*(t)\}+$$
$$\{-S_i(t)m_i^*(t)H_i + (S_i(t)+VC_i(t))m_i^*(t)H_i\}$$

where the last step follows from the facts that $S_i(t)+VC_i(t) < V(C_i(t) - C_{i,max}) < 0$ and $P_i^*(t) < m_i^*(t)H_i$. In this case, the objective with all power drawn from the grid is smaller. Hence, when $S_i(t) < -VC_{i,max}$, $D_i^*(t)$ can not be greater than zero. ■

IV. Performance Analysis

In this section, we analyze the feasibility and performance of our algorithm. First, we define an upper bound V_{max} on parameter V as follows:

$$V_{max} = \min_i \frac{E_{i,max} - R_{i,max} - D_{i,max}}{C_{i,max}}. \qquad (19)$$

Then, we have the following theorem:

Theorem 2: Suppose the initial battery energy level $E_{i,ini} \in [0, E_{i,max}]$. Implementing the above algorithm with any fixed parameter $V \in [0, V_{max}]$ for all time slots, we have the following performance guarantees:

1) The battery energy level $E_i(t)$ is always in the range $[0, E_{i,max}]$ for all time slots t.
2) All control decisions are feasible.
3) If $L(t)$ and $\vec{C}(t)$ are *i.i.d.* over slots, then the time-average cost under our algorithm is within bound B/V of the optimal value:

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\mathbb{E}\{P_i(t)C_i(t)\} \leq Q^{OPT} + B/V \quad (20)$$

where B is a constant given by

$$B \equiv \sum_{i=1}^{N} B_i \equiv \sum_{i=1}^{N} \frac{\max\{R_{i,max}^2, D_{i,max}^2\}}{2}. \qquad (21)$$

In the following, we prove **Theorem 2**.

Proof: Part 1: To show $0 \leq E_i(t) \leq E_{i,max}$, according to the definition of $S_i(t)$, it is equivalent to show that for each data center i,

$$-VC_{i,max} - D_{i,max} \leq S_i(t) \leq E_{i,max} - VC_{i,max} - D_{i,max} \quad (22)$$

As $0 \leq E_{i,ini} \leq E_{i,max}$, the above inequality holds for $t = 0$. We prove that this constraint is satisfied at all slots t by induction. Suppose inequality (22) holds for time slot t, we need to show that it also holds for time slot $t + 1$.

- We first prove $S_i(t+1) \leq E_{i,max} - VC_{i,max} - D_{i,max}$: if $0 < S_i(t) \leq E_{i,max} - VC_{i,max} - D_{i,max}$, then from Lemma 1, we must have $R_i^*(t) = 0$. Using Eq. (17) and $0 \leq D_i(t) \leq D_{i,max}$, we have $S_i(t+1) \leq S_i(t) \leq E_{i,max} - VC_{i,max} - D_{i,max}$; if $-VC_{i,max} - D_{i,max} \leq S_i(t) \leq 0$, then from Eq. (17) and $R_i(t) \leq R_{i,max}$, we have $S_i(t+1) \leq R_{i,max}$. For any $0 \leq V \leq V_{max}$, from the definition (19) of V_{max}, we have $E_{i,max} - VC_{i,max} - D_{i,max} \geq E_{i,max} - V_{max}C_{i,max} - D_{i,max} \geq E_{i,max} - \frac{E_{i,max}-R_{i,max}-D_{i,max}}{C_{i,max}}C_{i,max} - D_{i,max} = R_{i,max} \geq S_i(t+1)$. From the above discussion, we get $S_i(t+1) \leq E_{i,max} - VC_{i,max} - D_{i,max}$.
- Then we prove $S_i(t+1) \geq -VC_{i,max} - D_{i,max}$: if $-VC_{i,max} - D_{i,max} \leq S_i(t) < -VC_{i,max}$, then from Lemma 1, we must have $D_i^*(t) = 0$. Using the Eq. (17) and $0 \leq R_i(t)$, we have $S_i(t+1) \geq S_i(t) \geq -VC_{i,max} - D_{i,max}$; if $-VC_{i,max} \leq S_i(t) \leq E_{i,max} - VC_{i,max} - D_{i,max}$, then from Eq. (17) and $D_i(t) \leq D_{i,max}$, $S_i(t+1) \geq -VC_{i,max} - D_{i,max}$. From the above discussion, we get $S_i(t+1) \geq -VC_{i,max} - D_{i,max}$.

Part 2: As the constraint on $E_i(t)$ for each data center i is satisfied as showed in Part 1 and we make our decisions to satisfy all constraints in **Problem Three**, combining them together, all constraints of **Problem One** are satisfied. Therefore, our control decisions are feasible to **Problem One**.

Part 3: We make use of Laypunov optimization [10] to derive the performance bound of our algorithm. Define Laypunov function as $L(\vec{S}(t)) = \frac{1}{2}\sum\limits_{i=1}^{N} S_i^2(t)$. Define the conditional 1-slot Laypunov drift as follows:

$$\triangle(\vec{S}(t)) = \mathbb{E}\{L(\vec{S}(t+1)) - L(\vec{S}(t)) | \vec{S}(t)\}.$$

From Eq. (17), squaring both sides, we have for each data center i,

$$\frac{S_i^2(t+1) - S_i^2(t)}{2} = \frac{(D_i(t) - R_i(t))^2}{2} - S_i(t)(D_i(t) - R_i(t)). \tag{23}$$

For any feasible solution, in each time slot t, only one of $R_i(t)$ and $D_i(t)$ can have non-zero value. Hence,

$$\frac{(D_i(t) - R_i(t))^2}{2} \leq \frac{\max\{R_{i,max}^2, D_{i,max}^2\}}{2} \equiv B_i.$$

Taking expectations of both sides of (23) given $S_i(t)$ and summing over all data centers i, we have

$$\triangle(\vec{S}(t)) \leq \sum_{i=1}^{N} B_i - \sum_{i=1}^{N} S_i(t)\mathbb{E}\{D_i(t) - R_i(t) | \vec{S}(t)\}.$$

Adding penalty term $V \sum\limits_{i=1}^{N} \mathbb{E}\{P_i(t)C_i(t)|\vec{S}(t)\}$ into both sides of the above inequality, we have the following inequality:

$$\triangle(\vec{S}(t)) + V\sum_{i=1}^{N} \mathbb{E}\{P_i(t)C_i(t)|\vec{S}(t)\} \leq \sum_{i=1}^{N} B_i - \\ \sum_{i=1}^{N} S_i(t)\mathbb{E}\{D_i(t) - R_i(t)|\vec{S}(t)\} + V\sum_{i=1}^{N} \mathbb{E}\{P_i(t)C_i(t)|\vec{S}(t)\}$$

From the Eq. (12), we have $D_i(t) - R_i(t) = m_i(t)H_i - P_i(t)$. Plugging into the above inequality, we get

$$\triangle(\vec{S}(t)) + V\sum_{i=1}^{N} \mathbb{E}\{P_i(t)C_i(t)|\vec{S}(t)\} \leq \sum_{i=1}^{N} B_i + \\ \sum_{i=1}^{N} \mathbb{E}\{-S_i(t)m_i(t)H_i + S_i(t)P_i(t) + VP_i(t)C_i(t)|\vec{S}(t)\}$$

Comparing with the objective of **Problem Three**, it is obvious that our algorithm is always trying to greedily minimize the R.H.S. of the above inequality at each time slot t over all possible feasible control policies including the optimal, stationary policy given in Theorem 1. Plugging this policy into R.H.S. of the above inequality, we obtain the following:

$$\triangle(\vec{S}(t)) + V\sum_{i=1}^{N} \mathbb{E}\{P_i(t)C_i(t)|\vec{S}(t)\} \leq \sum_{i=1}^{N} B_i + \\ V\sum_{i=1}^{N} \mathbb{E}\{P_i^{stat}(t)C_i(t)|\vec{S}(t)\} = B + VQ^{REL} \leq B + VQ^{OPT}$$

where $B \equiv \sum_{i=1}^{N} B_i$. Taking the expectation of both sides, using the law of iterative expectation, and summing over $t \in \{0, 1, 2, \ldots, T-1\}$, we have

$$V\sum_{i=1}^{N}\sum_{t=0}^{T-1} \mathbb{E}\{P_i(t)C_i(t)|\vec{S}(t)\} \leq BT + VTQ^{OPT} \\ - \mathbb{E}\{L(\vec{S}(T))\} + \mathbb{E}\{L(\vec{S}(0))\}.$$

Diving both side by T, let $T \to \infty$ and using the facts that $E\{L(\vec{S}(0))\}$ are finite and $E\{L(\vec{S}(t))\}$ are nonnegative, we arrive at the following performance guarantee:

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N} \mathbb{E}\{P_i(t)C_i(t)\} \leq Q^{OPT} + B/V.$$

where Q^{OPT} is the optimal objective value, B is a constant, and V is a control parameter which has a maximum value given by Eq. (19). ■

V. Conclusion

In this paper, we apply the Laypunov optimization technique to solve the problem of optimal traffic distribution and battery charging/discharging management in Internet data centers under location-varying and time-varying electricity price. The algorithm we propose matches the intuition of distributing more traffic into data centers with low electricity price and charging when electricity price is low while discharging when electricity price is high. Moreover, it is easy to implement online and can give analytic bound on the performance. With the increase of battery capacity, our algorithm can get arbitrarily close to optimal value. However, our algorithm is centralized and we plan to design decentralized control algorithm in the future.

References

[1] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4, pp. 50–58, April 2010.

[2] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the electric bill for internet-scale systems," in *ACM SIGCOMM*, Barcelona, Spain, August 2009.

[3] S. Albers, "Energy-efficient algorithms," *Communications of the ACM*, vol. 53, pp. 86–96, May 2010.

[4] A. Beloglazov, R. Buyya, Y. C. Lee, and A. Y. Zomaya, "A taxonomy and survey of energy-efficient data centers and cloud computing systems," 2010. [Online]. Available: http://arxiv.org/abs/1007.0066

[5] 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 Conference on Computer Communications, INFOCOM 2010*, San Diego, March 2010.

[6] J. Li, Z. Li, K. Ren, X. Liu, and H. Su, "Towards optimal electric demand management for internet data centers," 2010. [Online]. Available: http://www.ece.iit.edu/~kren/Electricity_Datacenter_10.pdf

[7] A.-H. Mohsenian-Rad and A. Leon-Garci, "Energy-information transmission tradeoff in green cloud computing," in *Proc. of IEEE Global telecommunications conference, Globecom'10*, Miami, March 2010.

[8] R. Urgaonkar, B. Urgaonkary, M. J. Neely, and A. Sivasubramaniam, "Optimal power cost management using stored energy in data centers," in *ACM International Conference on Measurement and Modeling of Computer Systems, SIGMETRICS 2011*, SAN JOSE, June 2011.

[9] D. P. Bertsekas and R. G. Gallager, *Data networks*, 2nd ed. Prentice-hall New York, 1992.

[10] M. J. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Morgan Claypool, 2010.

[11] L. Georgiadis, M. J. Neely, and L. Tassiulas, *Resource allocation and cross-layer control in wireless networks*. Hanover, MA, USA: Now Publishers Inc., April 2006, vol. 1.

[12] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 2nd ed. Athena Scientific, 2000.

[13] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.