GreenColo: A Novel Incentive Mechanism for Minimizing Carbon Footprint in Colocation Data Center

Conference Paper · November 2014

DOI: 10.1109/IGCC.2014.7039140

Mohammad Atiqul Islam
University of California, Riverside

GreenColo: A Novel Incentive Mechanism for Minimizing Carbon Footprint in Colocation Data Center

Mohammad A. Islam, Shaolei Ren, Xiaorui Wang
Florida International University The Ohio State University

Abstract*—As an integral part of our digital economy, data centers are growing rapidly, devouring a formidable amount of energy and leaving a huge carbon footprint. While owner-operated data centers (e.g., Google) can implement various power management techniques to reduce energy consumption, colocation data centers, which offer a "halfway" solution to users who do not want to build their own data centers or completely resort to public clouds, suffer from "split incentive" that creates a barrier for greenness: colocation operator desires greenness but does not have control over tenants' servers; tenants who own the servers may not be willing to manage their servers for greenness unless they are properly incentivized. In this paper, we aim at minimizing the carbon footprint of a colocation data center while satisfying the colocation operator's long-term budget constraint. To break the split-incentive barrier and satisfy the budget constraint, we develop a *dynamic* incentive framework, called GreenColo, in which tenants can voluntarily submit energy reduction bids along with their desired payment and, if accepted, will be financially rewarded for energy reduction. GreenColo can be implemented online based on the currently available information (e.g., tenants' bids and current carbon efficiency) and dynamically select winning bids to minimize carbon footprint. We demonstrate the effectiveness of GreenColo both analytically and empirically. Our trace-based simulation results show that GreenColo can reduce carbon footprint by 18%, while the colocation operator does not incur any additional cost (compared to the no-incentive baseline case) and tenants may save up to 25% of their colocation costs.

I. Introduction

Data centers are so critical infrastructure in our digital economy that we cannot live without them. Nonetheless, data centers may each house tens of thousands of servers and hence are very power-hungry, collectively accounting for approximately 2% of the global electricity usage and resulting in a growing trajectory of carbon footprint [19]. Consequently, numerous efforts have been made to reduce the energy consumption as well as carbon footprints of data centers (referred to as "greenness" in this paper) [13, 20, 22, 36]. While the progress towards "greenness" is undeniably encouraging, the existing efforts have primarily focused on owner-operated data centers such as Google and Amazon, leaving the "dirtiness" of another distinctly different type of data center — colocation data center — much less addressed.

Colocation data center, often simply called colocation or colo, differs from owner-operated data centers (e.g., private data center, cloud data center), where the operator owns and has complete control over the server equipment. In a colocation, multiple tenants house their own servers in one shared facility, whereas the colocation operator (i.e., facility manager) is only responsible for facility operation (e.g., reliable power distribution and cooling system) without any control over the tenants' servers. There are two major business models in colocation: retail and wholesale. Retail tenants typically house their servers in a cage or cabinet with a smaller power demand (e.g., less than 500KW), whereas wholesale tenants usually lease a dedicated data center space (e.g., a full floor or even the whole facility) and have greater power demand. While colocation pricing varies widely in different markets, a prevailing pricing model is that tenants are charged based on their peak *power* subscriptions (even though tenants do not use any *energy* during the leasing period) [10, 14, 25]. In addition, other fees, such as space costs, bandwidth costs, and actual energy costs (but typically only for wholesale tenants), may also be charged.

Why is colocation important? Colocation is an integral segment of data center industry and has been keeping a strong momentum to grow. As noted by a recent book from Google [8], "most large data centers are built to host servers from multiple companies (often called colocation, or 'colos')." By one estimate [9], there are more than 1,200 colocation data centers in the U.S., and the combined peak power demand by such facilities in New York areas are estimated to exceed 400MW (comparable to Google's global data center power demand) [5]. Moreover, the now U.S.$ 25billion global colocation market is expected to grow to U.S.$ 43billion by 2018 with a projected annual compound growth rate of 11% [3].

The fast expansion of colocation market is driven by the strong IT demand across all sectors. First, while public cloud may satisfy some of the computing demands by small/medium businesses, concerns with privacy, losing control of data, and/or lack of technical skills still pervasively exist. Hence, maintaining self-owned servers in colocations (e.g., private cloud, hybrid cloud that only partially relies on public cloud) are favored by many users. Second, although a few gigantic cloud providers (e.g., Google and Amazon) can afford constructing self-owned customized data centers, many smaller-scale cloud providers (e.g., Salesforce, Box) cannot and they house their servers in colocations for providing public cloud computing services [24]. Last but not least, some top-brand IT companies also have a considerably large footprint in colocations. For example, Akamai and Twitter are common tenants in colocations [6, 15].

In this paper, we focus on the fast-growing yet long-neglected colocation industry, with a goal of greening colocation data center — minimizing carbon footprint — while satisfying colocation operator's budget constraint. Our research is mainly motivated by the following two facts. First, according to Greenpeace's latest report released in early 2014 [15], a handful of top-brand IT companies such as Google and Apple have been doing an excellent job in making their data centers green, but the "greenness" of colocations is still lagging far behind. In fact, because of their global reach and massive sizes, colocations are identified to possess far greater potentials in driving the clean Internet, than even the current industry leaders like Google [15]. Thus, colocations imperatively need to get on board to reduce their carbon footprints. This also helps colocation operators earn green certifications (which bring tax benefits [35]) and brighten their public images. Second, for financial interests, satisfying the colocation operator's budget constraint is certainly desired when going green.

Greening colocation while satisfying budget constraint is a **challenging** research problem. *First*, due to the lack of control over tenants' servers, the existing power management techniques (e.g., turning off unused servers [22]) cannot be applied by the colocation operator. Instead, the colocation operator can only rely on facility management (e.g., upgrading cooling systems) to improve energy efficiency, but it may require substantial capital investment, especially for the existing colocations. Hence, greening colocation data center is largely at the mercy of tenants who own the servers. Nonetheless, there is a "split incentive" hurdle: although greenness is desired by colocation operator, tenants may not be willing to manage their servers for greenness unless they are *financially* incentivized. This is because tenants (especially retail tenants) pay for their peak power subscriptions and, if without any financial benefits, have little incentive to reduce actual energy usage. Moreover, even though some large wholesale tenants may be charged based on energy consumption, a flat rate is typically used and hence mis-matches time-varying carbon efficiencies and/or on-site renewables [13], (e.g., tenants' energy usage pattern may not follow carbon efficiency and/or renewable energy supply variations). *Second*, as a common business practice, colocation operator's budget is determined before a fiscal year and hence, satisfying budget constraint requires a long-term cost budgeting process. Nonetheless, the complete information (e.g., time-varying carbon efficiency, tenants' workloads) over the budgeting period is not available in practice, necessitating an efficient online approach.

To break the "split incentive" hurdle, we employ an incentive framework, called GreenColo, which rewards the participating tenants for energy reduction based on reverse auction.[1] GreenColo takes root in economics theory and is further inspired by a number of prior works that focus on different contexts (e.g., demand response in smart grid [28, 37] and wireless traffic offloading [38]). In our problem, tenants can *voluntarily* submit bids, including planned energy reduction (e.g., by turning off unused servers) and requested monetary payment, to the colocation operator. Upon receiving the bids, the colocation operator will determine wining bids, notify respective tenants for energy reduction and then financially reward them. GreenColo has the advantage of being non-intrusive, as tenants enjoy the complete freedom in participating and deciding their bids.

While rewarding tenants for energy saving and greenness, the colocation operator also needs to satisfy its long-term budget constraint. In view of the practical constraint that complete offline information is unknown, GreenColo contains an online algorithm for deciding winning bids based on the currently available information (i.e., current carbon efficiency and currently submitted bids by tenants). The key idea of the online approach is: keep track of budget deficit; if there is a deficit, then give more emphasis on reducing operational cost while minimizing carbon footprint. We also demonstrate the effectiveness of GreenColo: trace-based simulations show that carbon footprint can be reduced by 18% without any additional cost (compared to the baseline case in which no incentive is provided), while tenants may save up to 25% of their power cost by participating in GreenColo.

To sum up, our main contribution is that we address a timely yet critically important problem of reducing carbon footprint of colocation for greenness, a long-neglected driving factor for sustainable computing. We develop an efficient online reverse auction-based incentive mechanism, GreenColo, which can dynamically decide winning bids based on the currently available information to minimize carbon footprint while satisfying the colocation operator's long-term budget constraint. GreenColo creates a "**win-win**" situation: colocation becomes greener without increasing the operator's budget; tenants receive financial rewards without noticeably affecting their applications performances. To our best knowledge, our work represents the first effort in greening colocation subject to long-term budget constraint, by breaking the split-incentive hurdle between colocation operator and tenants.

The rest of this paper is organized as follows. Section II describes the incentive mechanism and model. In Section III, we present the problem formulation and develop our online algorithm GreenColo. Sections IV provides the simulation results to support our analysis. Related work is reviewed in Section V and finally, concluding remarks are offered in Section VI.

II. GreenColo Mechanism and Model

In this section, we first describe how the incentive mechanism works, and then formalize the models for tenants and colocation operator. We will consider a widely-employed discrete-time model by dividing the timescale of interest (e.g. one year) into K equal-length time slots, indexed by $t = 0, 1, \cdots, K-1$. In our study, we focus on hourly time slot, although other durations (e.g., 15 minutes) can also be considered.

A. GreenColo *mechanism*

We describe GreenColo as follows, which is executed at the beginning of each time slot.

[1] Dynamically pricing tenant's energy usage may not be desired for colocation, because: (1) it implicitly enforces all tenants to accept time-varying prices, causing business uncertainties and/or psychological concerns; and (2) it may be subject to power utility regulations [14].

TABLE I
LIST OF KEY NOTATIONS.

Notation	Description	Notation	Description
$m_i(t)$	# of servers turned off	$e_i(t)$	Server energy
$\gamma(t)$	PUE	$r(t)$	On-site renewables
$u(t)$	Electricity price	$c(t)$	Carbon footprint
$e(t)$	Electricity usage	Z	Total cost constraint
$g(t)$	Operational cost	$\bar{d}_i$	Avg. delay constraint

- Step 1: Each tenant who voluntarily participates in GreenColo decides and submits (a bundle of) bids. As will be formalized in the next subsection, the bidding information includes the number of servers to be turned off in the upcoming time slot as well as the desired incentive payment from the colocation operator as a compensation.
- Step 2: Upon receiving the bidding information from tenants, the colocation operator selects bids by solving an online optimization problem (as detailed in Section III) and then notifies the tenants of the bidding outcome.
- Step 3: If its bid is accepted, the tenant will turn off its servers as specified in the bid and also receive the corresponding payment from the colocation operator. Power metering tools in colocation can be leveraged to verify that servers are turned off.

While tenants' participation in GreenColo is purely voluntary, we envision that GreenColo acts as an economic stimulus for tenants' cooperation with with the colocation operator for greenness. Our position is strengthened by increasing pressures from environmental groups [15] and recent sustainability commitment of major IT companies (e.g., Akamai and Salesforce, which house servers in worldwide colocations [6]).

B. Model

In what follows, we model each tenant's bidding decision and the operation of colocation operator. Key notations are listed in Table I. For notational convenience, we suppress the time index wherever applicable.

1) Tenant: We consider a colocation data center with N tenants, each tenant i having M_i homogeneous servers housed in the facility, for $i = 1, 2, \cdots N$. Note that our model is extensible to heterogeneous servers by viewing a tenant with heterogeneous servers as group of virtual tenants, each with homogeneous servers. In our model, we consider that tenant i turns off m_i servers to reduce energy [22, 28], although other IT control knobs, such as scaling down CPU frequencies [21], can also be considered.

Reducing energy via turning servers off may induce certain "inconveniences/costs" for tenants such as possible performance degradation, and hence, financial reward is needed as a compensation. Next, we model the tenant cost as a monotonically increasing function $h_i(m_i)$ in terms of the number of servers turned off. While the cost function can have different forms (e.g. discrete, non-linear, etc.) and each tenant has its own discretion to formulate the best suitable cost function, we provide an *example* of cost function as follows for explanation purposes. In particular, we consider two specific types of costs: delay performance cost and server unavailability cost.

- Delay performance cost: By consolidating workloads and turning off some unused servers, applications may experience delay performance degradation, causing "cost" to tenants [22, 27]. Here, we adopt the widely-used queueing-theoretic model to capture the delay cost. Specifically, by turning off m_i servers and considering an M/M/1 model at each active server, we formulate the delay cost of tenant i as

$$d_i(m_i, \lambda_i) = \beta_i \cdot \lambda_i \cdot \left(\frac{1}{\mu_i - \frac{\lambda_i}{M_i - m_i}} - d_{i,th} \right)^+, \qquad (1)$$

where μ_i is the service rate of each server (measuring the amount of workloads that can be processed in a unit time), λ_i is the arrival rate of workload equally distributed across active servers, β_i is a factor converting the experienced delay to an equivalent monetary cost, the operator $(\ \cdot\ )^+ = \max\{\ \cdot\ , 0\}$, and $d_{i,th}$ is the average delay threshold (i.e., users are indifferent of the delay performance below the threshold). Each tenant has an average delay constraint $\frac{1}{\mu_i - \frac{\lambda_i}{M_i - m_i}} \leq \bar{d}_i$, which essentially translates into an equivalent maximum server utilization constraint (a key metric for capacity autoscaling decisions on commercial cloud platforms such as Amazon EC2 [7]). Note that the considered performance model is only intended as a *guide* for capacity provisioning, and other systems approaches (e.g., resource demand modeling and prediction) can also be applied for estimating delay performance.

- Server unavailability cost: In addition to delay performance, turning off servers also results in other inconveniences (e.g., it may take a longer time in response to unexpected traffic spikes, etc.). Here, we collectively refer to these inconveniences as server unavailability cost. For illustration, we model the server unavailability cost of tenant i as a linearly increasing function $\eta_i \cdot m_i$, where $\eta_i > 0$ is a scaling factor and m_i is the number of servers turned off.[2]

By combining both delay performance and server unavailability costs, the total cost of turning off m_i servers for tenant i can be expressed as

$$h_i(m_i) = \eta_i \cdot m_i + d_i(m_i, \lambda_i), \qquad (2)$$

where the parameters are chosen at tenant i's own discretion when requesting payment from the colocation operator. Hence, we can express bidding set of tenant i

$$\mathcal{B}_i \subseteq \left\{ (m_i, h_i(m_i)) |\ m_i \in \mathbb{Z}^+ \text{ and } d_i(m_i, \lambda_i) \leq \bar{d}_i \right\}, \qquad (3)$$

where $\mathbb{Z}^+$ represents non-negative integers and the right-hand side represents the feasible bidding set from which tenant i can decide (a subset of) bids to the colocation operator.

Finally, we note our study is not restricted to the above example cost; other costs, such as possible data transfers among servers when turning off servers and migrating workloads, may also be incorporated at the tenants' own discretion.

[2] We consider that the server unavailability cost for *each* time slot is $\eta_i \cdot m_i$, when m_i servers are turned off during two consecutive time slots.

Furthermore, we will show in simulations that asking for a very high payment is against the tenants' best interest, because doing so will only reduce the chance of having their bids accepted (i.e., reducing tenants' financial rewards) without noticeably improving their application performances.

2) Colocation operator: The colocation operator is responsible for managing the facility such as power distribution and cooling system. Next, we model the colocation operator's electricity usage, operational cost, and carbon footprint.

• Electricity usage: In colocation, IT energy is mainly consumed by tenants' equipment. For tenant i, each server has an idle/static power of $p_{i,s}$, dynamic power of $p_{i,d}$, and service rate of μ_i. With m_i servers turned off (to be optimized in the next section), the energy consumption[3] of each server can be expressed as $p_{i,s} + \frac{\lambda_i}{(M_i - m_i)\cdot\mu_i} \cdot p_{i,d}$, where $\frac{\lambda_i}{(M_i - m_i)\cdot\mu_i}$ is the server utilization and λ_i is the arrival rate of workload equally distributed across $M_i - m_i$ servers. The linear power model has been widely considered in prior work [22] and shown to be fairly accurate in production systems [12, 26]. Thus, the total server energy consumption of tenant i is $e_i = (M_i - m_i) \cdot p_{i,s} + \frac{\lambda_i}{\mu_i} \cdot p_{i,d}$. Thus, by capturing the non-IT energy consumption using power usage effectiveness (PUE, measuring the ratio of total energy to IT energy) and considering that an amount of r on-site intermittent renewable energy (e.g., solar panels) is available, we obtain the total electricity usage of the colocation as

$$e = \left\{ \gamma \cdot \sum_{i=1}^{N} \left[(M_i - m_i) \cdot p_{i,s} + \frac{\lambda_i}{\mu_i} \cdot p_{i,d} \right] - r \right\}^{+}, \quad (4)$$

where γ is the (possibly time-varying) PUE to offset a fraction of electricity usage.

• Operational cost: We focus on operational cost rather than capital cost (e.g., building the data center, installing on-site renewables). Excluding costs such as human resources that are irrelevant to our study, there are two types of operational costs in GreenColo: electricity cost and incentives provided to tenants. Given (possibly time-varying) electricity price of u, the electricity cost is $g_e = u \cdot e$, where e is the electricity usage in (4). The total incentive payment is $g_p = \sum_{i=1}^{N} h_i(m_i)$, where $h_i(m_i)$ is the incentive payment if tenant i's bid of turning off m_i servers is accepted by the colocation operator (which we will optimize in the next section). Thus, the total operational cost is

$$g = u \cdot e + \sum_{i=1}^{N} h_i(m_i). \quad (5)$$

• Carbon footprint: The grid electricity comes from various generation systems [1, 34] and, depending on the fuel type used, has different carbon emission rates. Carbon efficiencies of several common energy fuel mixes are shown in Table III (in Section IV). As it is not possible to distinguish the energy fuel type once the electricity enters the grid, we use the following formula to derive the average carbon efficiency (with a unit of g/kWh) [13]

$$\phi = \frac{\sum \phi_f \cdot b_f}{\sum b_f}, \quad (6)$$

where ϕ_f is the carbon efficiency of fuel type f and b_f is the total electricity generation from fuel type f. Thus, we express the data center carbon emission by $c = \phi \cdot e + \phi_r \cdot r$, where ϕ_r is the carbon efficiency for on-site renewables (i.e., solar in our study) and r is the amount of available renewables. Note that due to time-varying energy fuel mixes to satisfy different demands, as shown in Fig. 1(b), the resulting average carbon efficiency ϕ also varies over time.

III. Algorithm Design of GreenColo

In this section, we first present the problem formulation for GreenColo and then, in view of the lack of complete offline information, propose an online algorithm that can decide winning bids submitted by tenants without foreseeing the future information.

A. Problem formulation

The focus of our study is to optimally select tenants' bids (i.e., deciding winning bids) for minimizing the carbon footprint while ensuring that the long-term operational cost of the colocation is kept under budget. We formulate the problem as follows.

$$\textbf{P-1}: \quad \min \; \bar{c} = \frac{1}{K} \sum_{t=0}^{K-1} \left[\phi(t) e(t) + \phi_r(t) r(t) \right] \quad (7)$$

$$\text{s.t.} \quad \sum_{t=0}^{K-1} g(t) \leq Z, \quad (8)$$

$$[m_i(t), h_i(m_i(t))] \in \mathcal{B}_i(t), \; \forall \, i, t. \quad (9)$$

where K is the total number of time slots over the entire budgeting period, the objective (7) is the long-term average carbon footprint, (8) is the long-term cost constraint (including both electricity cost and incentive paid to tenants), and the last constraint (9) requires that only those bids voluntarily submitted by tenants can be chosen (i.e., colocation operator cannot *force* tenants to turn off certain number of servers against tenants' will).

It is clear that **P-1** is an offline problem formulation that involves integer programming and requires complete offline information (e.g., tenants' future bids, carbon efficiency), which, however, is not possible to obtain in advance. Moreover, the intermittent on-site renwables as well as time-varying energy fuel mixes further add to the randomness of deciding winning bids over the entire cost budgeting period. To address the lack of offline information, we propose an online algorithm, GreenColo, which only requires the currently available information and approximately solves **P-1** with a bounded deviation from optimal offline solution.

[3] Without causing ambiguity, we interchangeably use energy and power, because they are equivalent given the equal-length time slot model.

Algorithm 1 GreenColo

```
Input $(m_i(t), h_i(m_i(t)))$, $\phi(t)$, $u(t)$ and $r(t)$ at the beginning of each time slot $t = 0, 1, \cdots, K-1$ and for $i = 1, 2, \cdots N$.
Decide winning bids to minimize
    $\mathbf{P\text{-}2}: V \cdot \left\{ \phi(t) \left[ \gamma \sum_{i=1}^{N} e_i(t) - r(t) \right]^{+} + \phi_r(t) \cdot r(t) \right\} + q(t) \cdot \left\{ u(t) \left[ \gamma \sum_{i=1}^{N} e_i(t) - r(t) \right]^{+} + \sum_{i=1}^{N} h_i(m_i(t)) \right\}$
Update $q(t)$ according to (10).
```

B. GreenColo

Building upon yet extending the recently-developed Lyapunov technique [23], we propose GreenColo, which eliminates the necessity of future information to solve **P-1**. GreenColo decouples the optimization decisions by replacing the long-term cost constraint (8) with a dynamic virtual budget deficit queue. Specifically, we construct a virtual budget deficit queue that tracks the run-time deviation from the desired long-term budget target and evolves as follows

$$q(t+1) = \left[q(t) + g(t) - \frac{Z}{K} \right]^{+}, \tag{10}$$

where $g(t)$ is the electricity cost plus incentive payment, and the queue length $q(t)$ indicates the colocation's operational cost surplus over the allocated budget thus far (assuming equal budget allocation for ease of presentation [23]). Note that the "allocated" budget is not enforced as a hard constraint; instead, it is only intended as a *guidance* for GreenColo to satisfying budget constraint. A positive queue length implies that the colocation operator needs to give more weight on cost saving to meet the long-term budget constraint. Thus, leveraging this intuition and using the budget deficit queue as a weight for operational cost in the online optimization, we present the online algorithm in Algorithm 1, which will be further explained below.

1) Working principle of GreenColo*:* As shown in Algorithm 1, we construct a new optimization problem **P-2** consisting of the original objective function (carbon footprint) scaled by $V \geq 0$ (referred to as carbon parameter) plus the operational cost multiplied by the budget deficit queue. The queue acts as the weighting parameter for cost saving relative to carbon reduction. Specifically, if the colocation operator incurs a higher cost than the budgeted amount, the queue length grows and pushes the optimization problem **P-2** towards cost saving in consecutive time slots to mitigate the budget deficit. Thus, the budget deficit queue dynamically guides online winning bids decision towards satisfying the long-term budget constraint. The parameter V governs the impact of the queue length on the optimization outcome. A larger V causes the queue length to have a less impact on the optimization and, as a result, the potential deviation from long-term budget may be larger and mitigated over a greater number of time slots; and vice versa.

TABLE II
MODELLING PARAMETERS (U.S. CURRENCY).

	Tenant #1	Tenant #2	Tenant #3
Delay cost β (cent/ms/10^6jobs)	75	50	5
η (cent/server/hour)	3	3	3
Power cost ($/kW/month)	145	145	145
Service rate (jobs/hour)	360,000	180,000	30
Soft threshold on avg. delay	12 ms	24 ms	175 s
Avg. delay constraint	20 ms	40 ms	300 s

Another appealing property of **P-2** is its low computational complexity. Specifically, in both of the two possible cases (i.e., on-site renewables are sufficient or insufficient to power the colocation, respectively), the objective function in **P-2** takes an additive form and hence is decomposable across all the tenants. Thus, at each time slot, **P-2** has a complexity of $\mathcal{O}(N)$ (i.e., linearly growing with the number of tenants), which is practically favorable for large colocations with tens of tenants.

Finally, we note that one can rigorously prove based on the sample-path Lyapunov technique [23] that GreenColo is efficient: GreenColo achieves a close-to-minimum carbon footprint compared to the optimal offline algorithm with future information, while still being able to approximately satisfy the long-term budget constraint with a bounded deviation. Thus, as governed by V, there exists a tradeoff between carbon footprint minimization and budget constraint satisfaction. This observation will be further substantiated in simulations, while the proof is available in [18] but omitted here for space limitations.

IV. SIMULATION

In this section, we present a trace-based simulation study to demonstrate the effectiveness of GreenColo. We show that GreenColo can reduce colocation carbon emission by 18% and save tenants' cost by up to 25%, while incurring no additional operational cost for the colocation operator (compared to the baseline case without incentive mechanisms). We first present our setup and then our simulation results.

A. Setup

We consider a colocation data center with three large (consolidated) tenants, each of which has 10,000 servers and may represent multiple tenants in practice. The three tenants run highly delay-sensitive, moderately delay-sensitive, and delay-tolerant workloads, respectively. The modeling parameters for tenants are shown in Table II, which we explain using Tenant #1 as an example.

— The parameter β converts delay performance to monetary value and quantifies the delay cost for every 10^6 requests, if the resulting average delay exceeds the software threshold by one *millisecond*. As shown in simulations, the values of β in Table II are already high enough to ensure that application performances are not noticeably affected. Similar model is also considered in prior work [22].

— The parameter η specifies the server unavailability cost for turning off each server for one hour. While there is no public disclosure of such data, we believe that 3 cent/server/hour is reasonable: with a 150W idle power for each server (in our setting), 3 cent/server/hour is already higher than the electricity cost saving of turning off a server had the tenants run servers in their own data centers (assuming a fair 15c/kWh electricity price). In other words, if tenants would like to turn off idle servers for cost saving in their own data centers (as extensively studied [22]), they should be more willing to do so if they house their servers in colocations. As shown later, further increasing the server unavailability cost beyond the level that the colocation operator can afford will not benefit tenants, because in that case, tenants cannot receive any financial rewards or noticeably improve their application performance.

— We consider the prevailing peak power-based pricing model [10, 25], and 145 U.S.$/kW/month is a fair market value [25, 31]. Service rates indicate the average number of jobs that can be processed, the soft delay threshold indicates the desired average delay below which users are indifferent with the service quality, and average delay constraint specifies the acceptable service quality.

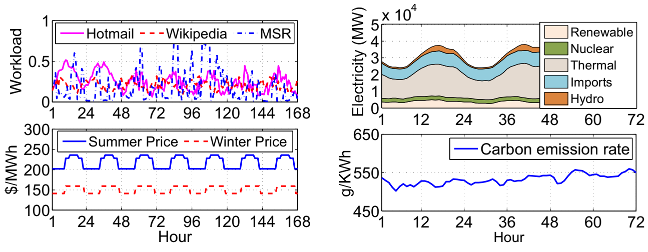
To limit free parameters, we consider that each server has an idle power of 150W and peak power of 250W. The budgeting period of our simulation is set to 1 year with each time slot of 1 hour. The default yearly budget constraint is set to 1.27 million U.S. dollars, which is the total cost of the colocation incurred when no incentive is provided and all servers are turned on as the status quo. The peak power of the colocation is 12MW and the PUE is set as 1.6, which is a fair value for colocations although some owner-operated data centers such as Google have reached a much lower PUE [11].

•**Workload.** Tenants have their own workloads: tenant #1 is running "Hotmail", tenant #2 is running "Wikipedia", and tenant #3 is running "MSR". The workload identified as "Hotmail" is taken from a 48-hour trace of 8 servers of Hotmail [32]. "Wikipedia" traces are taken from [33], which contain 10% of all user requests issued to Wikipedia from a 30-day period of September 2007, and "MSR" workload is a 1-week I/O trace of 6 RAID volumes at Microsoft Research Cambridge [32]. Due to lack of available traces for the entire budgeting period, we add up to 30% random variations and extend the available traces to get the 1-year trace. The workloads are normalized to corresponding tenant's maximum processing capacity and, by default, scaled to have an average utilization of 20% for tenants #1 and #2 and 30% for tenant #3, which is quite high even for Google [8].

•**Electricity price and on-site renewable energy.** We take the electricity price for business customers from PG&E [4], one of the largest power utilities in California. We use the "Time-of-Use" rates,[4] which have three different periods, off-peak, partial-peak and peak, and different electricity prices during different time periods. They also have different rates for Winter (November to April) and Summer (May to October). The electricity prices of 48 hours for Winter and Summer are shown in Fig. 1(a). We collect the solar power generation data from [1] for California for the year 2013 and use it as the trace for on-site renewable energy. We scale the data so that the maximum on-site renewable energy is 10% of the colocation's maximum peak energy.

TABLE III
U.S. CARBON EFFICIENCIES FOR FUEL TYPES (G/KWH) [13, 29].

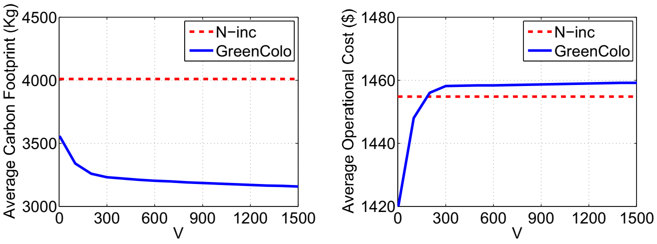
Wind	Solar	Nuclear	Coal	Gas	Imports	Hydro
22.5	46	15	968	440	562	13.5



(a)

(b)

Fig. 1. Trace data. **(a)** Workload traces [32, 33], on-site renewable energy [1], and electricity price data [4]. **(b)** Fuel mix and carbon emission rate [1, 29].



(a) Carbon footprint.

(b) Operational cost.

Fig. 2. Impact of V on carbon footprint and operational cost.

• **Carbon emission rate:** Due to lack of utility-level energy fuel mix data, we collect the fuel mix data from California ISO [1] for the year of 2013, and use carbon emission rate for energy fuel types presented in Table III to calculate carbon emission rate. The first 3-day data is shown in Fig. 1(b).

B. Results

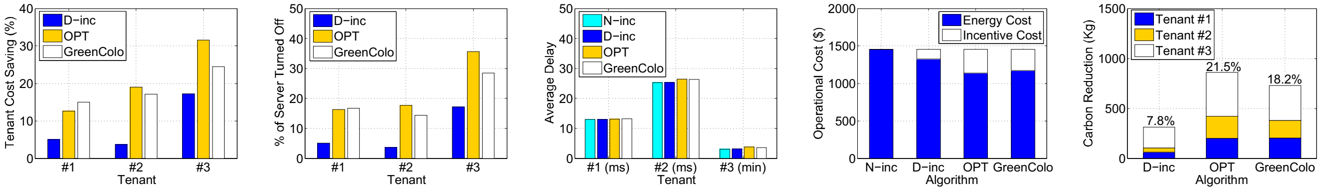
In this section, we present our simulation results. First, we introduce three benchmarks with which we compare GreenColo. Then, we examine the execution of GreenColo and show the performance comparison. Finally, we demonstrate the applicability of GreenColo in different scenarios. Unless otherwise stated, all the results are hourly values.

1) Benchmarks: We consider three benchmarks as below.

• **No Incentive (N-inc):** This is a baseline case in which no incentive is provided and the colocation is operated following the existing practice.

• **Direct Incentive (D-inc):** In D-inc, the colocation operator directly forwards the current electricity price (multiplied by an annualized PUE, reflecting the additional facility energy

[4] Charges independent of energy consumption, such as administrative fees, are excluded from our study.



(a) Tenant cost saving. (b) Servers turned off. (c) Delay performance. (d) Operational cost. (e) Carbon reduction.

Fig. 3. Performance comparison between GreenColo and benchmarks.

saving) to the tenants as an incentive for energy saving. Based on the provided direct incentive, tenants individually determine how many servers they would like to shut down to maximize their own benefits (i.e., difference between incentive received and cost incurred). D-inc does not exploit the time-varying nature of carbon emission efficiencies.

- **Optimal Offline** (OPT)**:** This is the optimal offline algorithm which, with complete future information (e.g., future bids submitted by tenants), solves the offline problem **P-1** and minimizes the carbon footprint subject to long-term budget constraint. OPT is not feasible in practice, but provides a lower bound on the carbon footprint that can be possibly achieved by GreenColo.

2) Execution of GreenColo*:* We first show the impact of control parameter V on the performance of GreenColo in Fig. 2(a) and Fig. 2(b), where N-inc is the no-incentive baseline case. We set the operational cost of N-inc as the budget constraint. It can be seen that V governs the tradeoff between carbon footprint reduction and budget constraint satisfaction: when V increases, GreenColo focuses more on reducing carbon footprint while caring less about operational cost, and vice versa. When $V \approx 150$, the desired budget constraint is satisfied, while the carbon footprint is significantly reduced compared to N-inc (by 18.2%).

3) Performance comparison: In Fig. 3, we compare the performance of GreenColo with the benchmark algorithms,

Reduce tenants' costs without noticeable performance degradation. First, we show the cost savings and delay performances of the three tenants under different algorithms. When we calculate cost saving percentages, we only consider power subscription cost based on 145 U.S.$/kW/month, assuming that tenants carefully subscribe to power based on their peak server power; other costs, such as space and network connectivity cost, vary widely by tenants and are often lower than power subscription cost [25]. Fig. 3(a) shows that using GreenColo, the tenant #3 has the highest cost saving of 25% while tenant #1 saves the least (by about 15%). This is due to differences between the tenants' delay tolerance levels: unlike tenant #1 running delay-sensitive workloads, tenant #3 runs delay-tolerant jobs and has a low delay cost, as well as a high average delay constraint. As a result, as shown in Fig. 3(b) and Fig. 3(c), tenant #3 can shut down many servers (by nearly 30% on average) without substantially affecting application performances. Tenant #1, on the other hand, turns down fewer number of servers to ensure that the resulting impact on application performance is negligible. In Fig. 3(c), we see that all the tenants' application performances when using incentive mechanism (i.e., GreenColo, OPT, D-inc) are nearly the same as those in the N-inc case. This is because tenants typically accept cost saving and green practices, only when application performance is not compromised: tenants set a sufficiently high delay performance cost parameter β to ensure that application performance is not significantly degraded.

Reduce carbon footprint without increasing operational cost. Next, we compare the operational cost and carbon reduction under different algorithms. We see from Fig. 3(d) that all the algorithms result in the same operational cost as N-inc, which we use as a reference case. Moreover, GreenColo provides a greater incentive payment to tenants than D-inc, because GreenColo is able to perform a joint optimization across all tenants by taking the advantage of tenant heterogeneity (e.g., tenant #3 voluntarily requests less payment than tenant #1 for reducing the same amount of energy). In Fig. 3(e), we show the carbon footprint reductions achieved by different tenants under different incentive mechanisms, compared to N-inc. It is observed that, although tenant #3 has the highest average utilization (i.e., 30%), it contributes the most to carbon footprint reduction, because its workloads are delay-tolerant in our simulation. We also see that GreenColo achieves a much higher carbon footprint reduction than D-inc by encouraging tenants to turn off more servers. More remarkably, the carbon footprint reduction achieved by GreenColo is fairly close to that by OPT (18.2% versus 21.5%), demonstrating the effectiveness of GreenColo even though only online information is available.

Sensitivity studies, such as robustness of GreenColo against tenants' workload prediction errors, are deferred to [18].

V. Related Work

In this section, we discuss the related work from the following perspectives.

- **Data center cost/carbon minimization:** Making data centers cost and/or carbon efficient has been studied by many prior studies [13, 16, 22]. For example, dynamically scaling server capacity provisioning to strike a balance between energy cost and performance loss has been the primary focus of several recent studies [16, 22]. Extending to a set of geo-distributed data centers, [26, 27] considers geographic load balancing to minimize the electricity cost and [13, 36] leverages spatio-temporal carbon efficiency to make data centers greener. These studies, however, focus on owner-operated data centers in which resource management can be performed at the data center operator's discretion. Thus, while the technological advances made by these studies are appealing, they cannot be directly applied to colocation data centers unless tenants are properly incentivized.

• **Incentive design:** Incentive mechanism has been successfully applied in various engineering domains, such as time-dependant pricing in wireless networks [17], rebate-based incentive in smart grid [37], and coupon-based rewarding scheme for WiFi traffic offloading [38]. Economics theory has also been applied in computer science, such as auction-based Amazon Spot Instance market [2], and auction-based scheduling in high-performance computing [30]. While these works all leverage incentive mechanisms for various purposes, none of them have considered the context of greening colocation, a unique yet fast-growing segment of data center industry. The most recent work [28] investigated colocation demand response to make power grid more sustainable/stable by reducing energy upon requests by utilities, and hence it is a *one*-step optimization problem (i.e., there is no coupling across different time slots). In sharp contrast, our work focuses on making colocation itself sustainable and addresses the following new challenges: (1) greening colocation while satisfying colocation operator's yearly budget requires long-term efforts, presenting challenges to making online decisions that cannot possibly foresee users' far future bidding information; (2) as shown in Fig. 1(b), carbon efficiency varies over time, which must be taken into account, but future carbon efficiencies may not be known in advance.

VI. Conclusion

In this paper, we recognized that the split-incentive hurdle between colocation operator and tenants is limiting the carbon efficiency of colocation. To address this issue, we investigated a reverse auction-based incentive mechanism, using which tenants who voluntarily reduce energy consumption can receive financial rewards. We developed an online algorithm, GreenColo, which, based on tenants' bidding information, dynamically selects the number of servers to turn off and rewards tenants while satisfying the long-term budget constraint. In our trace-based simulation study, we showed that GreenColo can achieve 18% carbon reduction for the colocation and save tenant's cost by up to 25%, while the colocation operator does not incur any additional cost compared to the baseline case in which no incentive is provided.

Acknowledgement

This work was supported in part by the National Science Foundation under Grant No. IIP-1338922, CNS-1143607, and CNS-1423137.

References

[1] California ISO, http://www.caiso.com/.

[2] Amazon EC2 Spot Instances, http://aws.amazon.com/ec2/spot-instances/.

[3] Colocation market - worldwide market forecast and analysis (2013 - 2018). http://www.marketsandmarkets.com.

[4] http://www.pge.com/.

[5] Telegeography colocation database, http://www.telegeography.com/.

[6] Akamai. Environmental sustainability policy, http://www.akamai.com/html/sustainability/our_commitment.html.

[7] Amazon EC2, https://aws.amazon.com/ec2/.

[8] L. A. Barroso, J. Clidaras, and U. Hoelzle. *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines*. Morgan & Claypool, 2013.

[9] DatacenterMap. Colocation USA, http://www.datacentermap.com/usa/.

[10] Enaxis Consulting. Pricing data center co-location services, 2009, http://enaxisconsulting.com.

[11] Equinix, www.equinix.com.

[12] X. Fan, W.-D. Weber, and L. A. Barroso. Power provisioning for a warehouse-sized computer. In *ISCA*.

[13] P. X. Gao, A. R. Curtis, B. Wong, and S. Keshav. It's not easy being green. *SIGCOMM Comput. Commun. Rev.*, 42(4):211–222, Aug. 2012.

[14] J. Glanz. Landlords double as energy brokers. In *The New York Times*, May 23, 2013.

[15] Greenpeace. Clicking clean: How companies are creating the green internet, 2014.

[16] B. Guenter, N. Jain, and C. Williams. Managing cost, performance and reliability tradeoffs for energy-aware server provisioning. In *IEEE Infocom*, 2011.

[17] S. Ha, S. Sen, C. Joe-Wong, Y. Im, and M. Chiang. Tube: time-dependent pricing for mobile data. In *SIGCOMM*, 2012.

[18] M. A. Islam, S. Ren, and X. Wang. Greening colocation data center, 2014, http://www.cs.fiu.edu/~sren/doc/tech/greencolo_full.pdf.

[19] J. G. Koomey. Growth in data center electricity use 2005 to 2010, 2011.

[20] K. Le, R. Bianchini, J. Zhang, Y. Jaluria, J. Meng, and T. D. Nguyen. Reducing electricity cost through virtual machine placement in high performance computing clouds. SuperComputing, 2011.

[21] S. Li, M. Brocanelli, W. Zhang, and X. Wang. Data center power control for frequency regulation. In *PES*, 2013.

[22] M. Lin, A. Wierman, L. L. H. Andrew, and E. Thereska. Dynamic right-sizing for power-proportional data centers. In *IEEE Infocom*, 2011.

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

[24] J. Novet. Colocation providers, customers trade tips on energy savings, Nov. 2013.

[25] D. S. Palasamudram, R. K. Sitaraman, B. Urgaonkar, and R. Urgaonkar. Using batteries to reduce the power costs of internet-scale distributed networks. In *SoCC*, 2012.

[26] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs. Cutting the electric bill for internet-scale systems. In *SIGCOMM*, 2009.

[27] L. Rao, X. Liu, L. Xie, and W. Liu. Reducing electricity cost: Optimization of distributed internet data centers in a multi-electricity-market environment. In *IEEE Infocom*, 2010.

[28] S. Ren and M. A. Islam. Colocation demand response: Why do I turn off my servers? In *ICAC*, 2014.

[29] J. V. Spadaro, L. Langlois, and B. Hamilton. Greenhouse gas emissions of electricity generation chains: Assessing the difference. *IAEA bulletin*, 42(2):19–28, 2000.

[30] M. Taifi. Banking on decoupling: Budget-driven sustainability for hpc applications on auction-based clouds. *SIGOPS Oper. Syst. Rev.*, 47(2):41–50, July 2013.

[31] Terremark, www.terremark.com.

[32] E. Thereska, A. Donnelly, and D. Narayanan. Sierra: a power-proportional, distributed storage system. *Tech. Rep. MSR-TR-2009-153*, 2009.

[33] G. Urdaneta, G. Pierre, and M. Van Steen. Wikipedia workload analysis for decentralized hosting. *Computer Networks*, 2009.

[34] U.S. DoE. http://energy.gov/.

[35] U.S. Green Building Council. Leadership in energy & environmental design, http://www.usgbc.org/leed.

[36] Y. Zhang, Y. Wang, and X. Wang. Greenware: Greening cloud-scale data centers to maximize the use of renewable energy. In *Middleware*, 2011.

[37] H. Zhong, L. Xie, and Q. Xia. Coupon incentive-based demand response: Theory and case study. *IEEE Trans. Power Systems*, 28(2):1266–1276, May 2013.

[38] X. Zhuo, W. Gao, G. Cao, and Y. Dai. Win-coupon: An incentive framework for 3g traffic offloading. In *ICNP*, 2011.