Colocation Demand Response: Why Do I Turn Off My Servers?

Shaolei Ren and Mohammad A. Islam, *Florida International University*

https://www.usenix.org/conference/icac14/technical-sessions/presentation/ren

This paper is included in the Proceedings of the 11th International Conference on Autonomic Computing (ICAC '14).

June 18–20, 2014 • Philadelphia, PA

ISBN 978-1-931971-11-9

Open access to the Proceedings of the 11th International Conference on Autonomic Computing (ICAC '14) is sponsored by USENIX.

Colocation Demand Response: Why Do I Turn Off My Servers?

Shaolei Ren
Florida International University

Mohammad A. Islam
Florida International University

Abstract

Data centers are promising participants in demand response programs (i.e., reducing a large electricity demand upon utility's request), making power grid more stable and sustainable. In this paper, we focus on enabling colocation data center demand response. Colocation is an integral yet unique segment of data center industry, where multiple tenants house their servers in one shared facility. Nonetheless, differing from owner-operated data centers (e.g., Google), colocation data center suffers from "split incentive": colocation operator desires demand response for financial incentives but has no control over tenants' servers, while tenants who own the servers may not desire demand response due to lack of incentives. To break "split incentive", we propose a first-of-its-kind incentive mechanism, called iCODE (incentivizing COlocation tenants for DEmand response), based on reverse auction: tenants, who *voluntarily* submit energy reduction bids to colocation operator, will be financially rewarded if their bids are accepted. We formally model how each tenant decides its bids and how colocation operator decides winning bids. We perform a trace-based simulation to evaluate iCODE. We show that iCODE can reduce colocation energy consumption by over 50% during demand response periods, unleashing the potential of colocation demand response.

1 Introduction

Demand response program has been adopted as a national strategic plan for power grid innovation [12]. In a typical demand response program, participating customers, who reduce electricity demands upon requests by utility/load serving entity (LSE), receive financial compensation.[1] Demand response is also favorably recognized as an effective market-based mechanism for increasing the incorporation of renewables into the grid, via the provisioning of economic incentives for reshaping customers' real-time electricity demand subject to time-varying supply availability [16, 40].

Mega-scale data centers are ideal participants in demand response programs and can reduce a *large* electricity demand upon LSE's request, because of their huge yet flexible energy demand [15, 16, 28]. Nonetheless, the existing efforts [4, 5, 15, 16, 28, 29] have only focused on owner-operated data centers (e.g., Google and Amazon), while neglecting another important yet distinctly different type of data center — *colocation* data center, sometimes simply called "colocation" or "colo". In sharp contrast with owner-operated data center whose operator owns and has full control over the servers, colocation is a multi-tenant facility where multiple tenants put their own servers in one shared facility while the data center operator (i.e., facility manager) provides reliable power supply, cooling, and network access.

What makes colocation demand response challenging? A major hurdle for colocation demand response is **"split incentive"**: while colocation provider may desire demand response for incentives from LSE, its tenants may not, because tenants are typically charged based on their subscribed peak power and their bills are not subject to how much energy they consume or when they consume it [11, 35, 39].[2] LSE's incentive programs are not directly open to tenants either, since tenants only have interactions with colocation operator [39]. While colocation operator may manage the non-IT energy consumption (e.g., cooling) for demand response, such actions often have limited energy reduction (as corroborated by real-life field tests [16]) as well as possible detrimental effects on tenants' servers, which may not be as robust as state-of-the-art servers (such as Google's). Using on-site diesel generators to offset electricity usage for demand response is uneconomical for the colocation operator.

How to enable colocation demand response? This

[1] A comprehensive survey of various demand response programs can be found in [3].

[2] Energy-based pricing may also also available (especially for large wholesale tenants), but typically a flat electricity price is used, thereby making tenants "blind" to demand response opportunities.

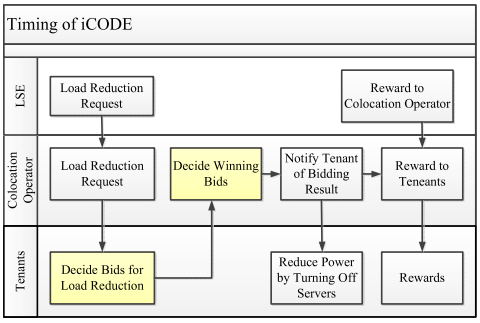


Figure 1: Timing of iCODE in colocation data center.

paper takes the first step to break "split incentive" hurdle for colcation demand response by properly incentivizing tenants. Specifically, we propose a first-of-its-kind market mechanism based on reverse auction that financially compensates tenants who are willing to shed their energy consumption (e.g., by turning off unused servers) for demand response. The proposed mechanism, called iCODE (incentivizing COlocation tenants for DEmand response), is fully voluntary and works in the following steps, as illustrated in Fig. 1. First, when demand response signals/requests are received by colocation operator and passed down to tenants, tenants can submit bids that include how much energy consumption they are willing to reduce and how much payment they want to receive as a compensation. Then, colocation operator selects winning bids that provide large energy reduction yet ask for reasonable payment, such that the energy reduction can be maximized while the total payment to the tenants does not exceed that received by the colocation operator from LSE. Finally, demand response is executed as planned, and payments are made accordingly. The practical implementation of iCODE is lightweight, requiring no manual efforts during execution.

We perform a trace-based simulation study to validate the effectiveness of iCODE in terms of energy reduction for demand response. Compared to the baseline case in which tenants are oblivious to demand response, iCODE can successfully incentivize tenants to reduce energy consumption by more than 50%, demonstrating a promising potential for colocation demand response.

2 Why colocation demand response?

We show two reasons that motivate our study of colocation demand response.

• Colocation is an essential and critically important business model in data center industry, offering a "halfway" solution for companies that do not want to build their own data centers or completely outsource their IT requirements to public cloud providers (e.g., for privacy concerns). Tenants in colocations include not only small and medium businesses for which building wholly-owned data centers is out of the question, but also content distribution providers (e.g., Akamai) and many of the top-branded IT companies (e.g., Amazon and Microsoft) that desire global footprints for their last-mile service latencies. Cloud computing also finds its *physical* home in colocations: e.g., medium-scale cloud providers, such as Salesforce and Box.com, provide their public cloud services in colocations, as building self-operated data centers is still uneconomical for them [34]. It is estimated that in the U.S. there are over 1,000 colocation data centers. With the explosive IT demands across all sectors, many colocation providers are also expanding their data center space [21], and recent analysis shows that colocation market is expected to grow at a compound annual growth rate of 11%, reaching US$ 43billion by 2018 [1].

• Colocations are even more suitable than owner-operated data centers for demand response. First, colocations have huge power demands, and the peak power demand of colocations in New York region exceeds 400MW (comparable to aggregate demand of Google's global data centers) [2, 9, 36]. As noted by a recent Google study [20], "most large data centers are built to host servers from multiple companies" (i.e., colocations). Second, even more importantly, many large colocations are often located in densely-populated metropolitan areas (e.g., Los Angeles [9]) where demand response is particularly desired for peak load shaving, whereas mega-scale owner-operated data centers (e.g., Google) are almost all located in rural areas with very low population densities where the need for demand response is less urgent.

3 Incentivizing colocation tenants for demand response

In this section, we first present an overview of iCODE, formalize the models for tenants and colocation operator, and then present the algorithms underlying iCODE (i.e., deciding tenants' bids and deciding winning bids).

3.1 Overview of iCODE

We begin by presenting an overview of the proposed iCODE mechanism framework to highlight its foundations and why we choose reverse auction.

Foundations of iCODE

iCODE relies the following foundations.

Technology. Turning unused servers off is one of the most extensively studied control knobs for energy saving [20, 25]. While tenants remotely house their servers in colocation, switching servers between active and sleep/off modes can be easily automated without manual efforts [20]. Thus, "turning off unused servers" without noticeably affecting tenants' business is technologically feasible.

Economics. Server energy reduction for demand response clearly requires cooperation from multiple tenants via non-technological mechanisms. Market knobs, such as pricing and incentives, have been leveraged to address various engineering issues [30, 32], and recent research has shown that owner-operated data centers are willing to shut down some servers for demand response incentives [15, 28]. Hence, we take the liberty that the proposed iCODE is worth investigating and promising for enabling colocation demand response.

Why reverse auction?

We first note that dynamically pricing energy usage for demand response, a widely-studied market mechanism (e.g., in smart grid [32]), may not be as plausible as it appears in the context of colocation. First, directly "reselling" energy and modifying energy price may be subject to strict government regulations [38]. Second, dynamically pricing tenants will implicitly enforce all tenants to face uncertain colocation costs, causing business reluctance and/or psychological concerns [30, 46]. Finally, we note that registering tenants to power utility's pricing is not feasible either, since tenants cannot plug their servers into utility's grid directly [39]; instead, tenants need colocation operator's *combined* facility support (e.g., secured access, reliable power, cooling, network), not only facility space [18].

We advocate a reverse auction-based incentive mechanism, as illustrated in Fig. 1. By "reverse", we mean that in our mechanism, it is not the colocation operator who proactively offers rewards to tenants for energy reduction; instead, it is the tenants who, at their own discretion, submit bidding information (including how much energy reduction and how much payment requested) upon receiving a demand response signal. iCODE is "non-intrusive" and tenants are not enforced for demand response or entitled any penalties if they do not participate in demand response.

3.2 Model

As in [15, 16, 28], we ignore the time index and focus on one-time demand response, whose duration T is determined by LSE (e.g., 15 minutes to one hour). Next, we present the models for tenants and colocation operator.

Tenants

We consider N tenants housing their servers in one colocation. Tenant i owns M_i homogeneous servers, while a tenant having multiple heterogeneous types of servers can be viewed as multiple *virtual* tenants each having homogeneous servers. Each server belonging to tenant i has a static/idle power of $p_{i,s}$, dynamic power $p_{i,d}$, and service rate of μ_i (measured in terms of the amount of workloads that may be processed in a unit time) [25]. During the demand response period, the workload arrival rate is denoted by λ_i which can be predicted to a fairly reasonable accuracy using, e.g., regression techniques [25, 37]. Our simulation will also investigate the robustness of iCODE against inaccurate knowledge of λ_i.

Server energy reduction. The baseline case is that tenants do not participate in demand response (e.g., due to lack of incentives, or even not knowing the demand response requests). In this case, all servers are active and workloads are evenly distributed across servers for optimized performance. Thus, the average power consumption of tenant i's servers is $p_i = M_i \cdot \left[p_{i,s} + p_{i,d} \cdot \frac{\lambda_i}{M_i \cdot \mu_i} \right] = M_i \cdot p_{i,s} + p_{i,d} \cdot \frac{\lambda_i}{\mu_i}$, where $\frac{\lambda_i}{M_i \cdot \mu_i}$ is the server utilization.

If tenant i decides to participate in demand response by turning off $m_i \geq 0$ servers, then its average power will be $p_i' = (M_i - m_i) \cdot p_{i,s} + p_{i,d} \cdot \frac{\lambda_i}{\mu_i}$. Hence, energy/load reduction by tenant i will be

$$\Delta e_i(m_i) = (p_i - p_i') \cdot T = m_i \cdot p_{i,s} \cdot T, \qquad (1)$$

where $p_{i,s}$ is the static power and T is the demand response duration.

Tenant cost. Turning off some servers will result in "costs". As an *example*, we consider switching cost and delay cost [25], while other costs (e.g., management costs) can also be factored in.

Switching cost: Turning servers into sleep/off mode and bringing them back to normal operation incur switching/toggling costs, such as wear-and-tear [25]. We denote tenant i's switching cost for one server by α_i (quantified in monetary units), and thus the total switching cost for tenant i is $\alpha_i \cdot m_i$.

Delay cost: We model the workload serving process at each server as an M/M/1 queue. Thus, the average delay for tenant i's workload is $\frac{1}{\mu_i - \frac{\lambda_i}{M_i - m_i}}$. The queueing model has been widely used as an analytic vehicle to provide a reasonable approximation for the actual service process [13, 27]. Note further that delay cost is incurred only when the average delay exceeds a soft threshold $d_{i,th}$: further reducing delay below the threshold makes no difference to human perception, and hence incurs no performance penalty. A large soft delay threshold means the tenant's workloads are more delay-tolerant. Next, we can express the total delay cost as $d_i(m_i) = \lambda_i \cdot \left[\frac{1}{\mu_i - \frac{\lambda_i}{M_i - m_i}} - d_{i,th} \right]^+$, where $[\,\cdot\,]^+ = \max\{0, \cdot\}$.

Colocation operator

Colocation operator provides reliable cooling and power supplies to tenants. Here, we capture the colocation's non-IT energy reduction (e.g., cooling, power distribution, etc.) using the PUE factor γ, which typically ranges from 1.1 to 2.0 [20]. That is, with a total IT energy reduction of $\sum_i \Delta e_i$ by tenants, the facility-level energy re-

duction will be $\gamma \cdot \sum_i \Delta e_i$. To procure a load reduction from customers (including colocation), LSE announces a price denoted by q. Thus, the rewards provided by LSE to colocation operator will be $q \cdot \gamma \cdot \sum_i \Delta e_i$.

3.3 Reverse auction in iCODE

Below, we specify these two elements of iCODE, as highlighted in Fig. 1.

Deciding tenants' bids. In order to participate in demand response, tenants need to be properly incentivized. Below, we denote tenant i's requested payment for turning off m_i servers by

$$c_i(m_i) = w_i \cdot [\alpha_i \cdot m_i + \beta_i \cdot d_i(m_i)], \tag{2}$$

where $w_i \geq 1$ is referred to as *greediness* of tenant i, and $\beta_i \geq 0$ converts delay cost to monetary values (i.e., the larger β_i, the more tenant i cares about its delay performance) [25]. Tenant i may submit multiple bids $(\Delta e_i, c_i)$, each corresponding to one value of $m_i \geq 0$ (i.e., the number of servers turned off). Moreover, tenant i may only choose to turn off up to $\bar{m}_i$ servers such that the delay performance is still tolerable. For convenience, we denote the set of tenant i's bids as $\mathbf{b}_i \subseteq \mathbf{B}_i = \{(\Delta \mathbf{e}_i, \mathbf{c}_i) \,|\, (\Delta \mathbf{e}_i(m_i), \mathbf{c}_i(m_i)),\ m_i = 0, 1, \cdots, M_i - 1\}$, such that $\mathbf{b}_i$ only contains valid bids (e.g., those bids satisfying tenant i's tolerable delay performance or equivalently, m_i is below a threshold $\bar{m}_i$).

We note that in iCODE, each tenant decides its bid purely at its own discretion. Tenants may ask for arbitrarily high payments, but doing so is not of tenants' interests because their bids will be less likely accepted and tenants will receive less payment without noticeably improving their delay performance (see Section 4.2).

Deciding winning bids. In our study, we consider the objective of maximizing the total energy reduction subject to the constraint that colocation operator does not need to compensate tenants out of its own pocket. Mathematically, the problem of deciding winning bids (DWB) can be formalized as:

$$\text{DWB}: \quad \max_{(\Delta e_i, c_i), \forall i \in I} \gamma \cdot \sum_{i \in I} \Delta e_i \tag{3}$$

$$\text{s.t.} \quad \sum_{i \in I} c_i \leq q \cdot \gamma \cdot \sum_{i \in I} \Delta e_i, \tag{4}$$

$$(\Delta e_i, c_i) \in \mathbf{b}_i \cup \{(0,0)\},\ \forall\, i \in I, \tag{5}$$

where I is the set of tenants who submit their bids to colocation operator, (3) specifies the objective of maximizing energy reduction, (4) indicates that the total compensation paid to tenants will not exceed the value received from the LSE (i.e., colocation operator will not lose profits due to active participation in demand response), and (5) specifies that colocation operator can only select "energy reduction, payment" pairs out of the bids submitted by tenants to honor their requests. We add $\{(0,0)\}$ in (5) to indicate that not necessarily all tenants' energy reduction requests will be accommodated (e.g., when they ask for very high payments).

The objective of energy reduction maximization benefits all parties involved: LSE can reduce peak power supply, tenants receive their requested monetary incentives if their bids are accepted, and colocation operator can reduce its energy bill and/or seek green certifications (e.g., LEED [41]) due to lower energy consumption. Note that iCODE can also be adapted for other purposes (e.g., maximizing colocation profit, if permitted by regulations).

While DWB is NP-hard and there exist various approximate solutions [31], we note one approach to solving DWB based on branch and bound technique that can yield a sub-optimal solution with a reasonably low complexity [6]. A sketch is provided below for brevity. Specifically, if we make a relaxation and allow e_i to take continuous values, then the requested payment in (2) is convex in e_i, and DWB becomes convex programming, for which there exists time-efficient methods [7]. The resulting energy reduction is an upper bound on the optimal value of DWB. On the other hand, if we choose a greedy-based approach (e.g., select the bids in ascending order of $\Delta e_i / c_i$), then we will obtain a lower bound on the optimal value of DWB. If the obtained upper and lower bounds are sufficiently close, then we can choose the greedy solution, because its energy reduction is close to the upper bound (and hence optimum, too). Otherwise, we can recursively solve DWB by fixing some bids to be selected and solving a smaller-scale sub-problem. To solve the sub-problem, we find lower/upper bounds via greedy/relaxation approach; if the bounds are still far apart, we further decompose the sub-problem into an even smaller-scale sub-problem. Repeat this process until the gap between the two bounds are sufficiently small or the maximum iteration number is reached. Finally, note that the computational complexity of solving DWB, although interesting by itself, is not a major bottleneck, as colocation operator receives demand response signal from LSE well beforehand and there is no need to solve DWB in real time [15, 28].

Remark. In this paper, we focus on how the colocation operator decides winning bids out of those submitted by tenants, while leaving the possibly strategic bidding process (e.g., tenants strategically place bids to maximize their own benefits) as a future study.

4 Performance evaluation

This section presents trace-based simulation studies to evaluate iCODE. We first present our settings, and then show the simulation results.

Table 1: Default model parameters.

Tenant	#1	#2	#3
Service rate μ (Jobs/hour)	360,000	180,000	30
Delay cost β (¢/ms/10^6jobs)	30	20	0.4
Switching cost α (¢/server)	0.5	0.5	0.5
Greediness factor w	1	1	1
Soft avg. delay threshold	12 ms	25 ms	175 s
Avg. delay constraint	20 ms	40 ms	300 s

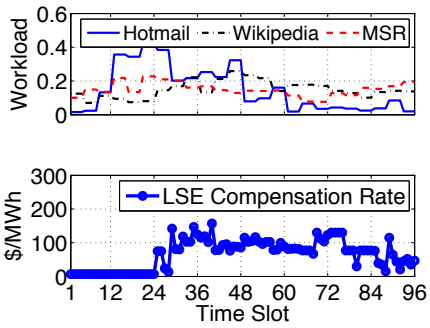


Figure 2: Traces of workload and compensation rate for energy reduction. Each time slot is 15 minutes.

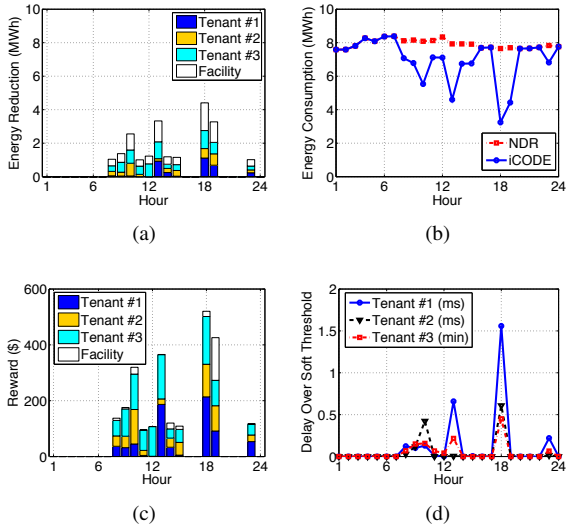


Figure 3: Comparison between iCODE and NDR. **(a)** Energy reduction by iCODE. **(b)** Energy consumption. **(c)** Incentives received. No incentives are provided in NDR. **(d)** Average delay exceeding the soft threshold in iCODE.

4.1 Settings

We consider a colocation facility located in New York, NY (a major market of colocation satisfying financial institution needs [9]), with a PUE of 1.6. The colocation participates in hour-ahead demand response program: when needed, LSE sends demand response request one hour ahead, while each demand respond period is 15 minutes. Nonetheless, due to figure space constraints, we consolidate four 15-minute time slots and show the hourly values for better presentation. Fig. 2 shows a snapshot of the LSE's one-day compensation rate for energy reduction on Feb. 24, 2014, obtained from [33].

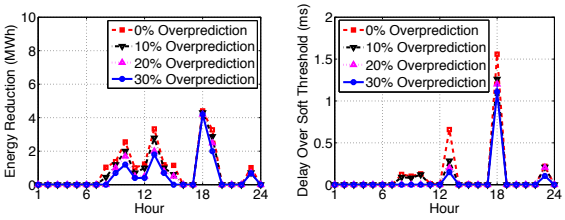
There are three tenants, each possibly representing multiple tenants in practice and having 10,000 homogeneous servers with 150W static and 100W dynamic power. Tenant #1 and #2 process delay-sensitive workload, while tenant#3 processes delay-tolerant workload. The default settings for tenant models are shown in Table 1. In particular, we note that the delay constraint (within a server) for tenant #1 is consistent with the existing interactive service requirement (e.g., web search [19]). Moreover, the switching cost of 0.5 cents for turning one server off for 15 minutes is already higher than the corresponding electricity cost saving achieved by tenants had they run servers in their own data centers (assuming a fair electricity price of 10 cents/kWh). In other words, because of higher cost savings, tenants should be more willing to turn off unused servers in colocation, than they would if they had in-house data centers (which has been extensively studied [37]). We obtain the three tenants' workload traces from [25] ("Hotmail" and "MSR") and [42] ("Wikipedia"). Fig. 2 illustrates a snapshot of the traces, where the workloads are normalized with respect to the maximum service capacity of each tenant's servers (with an average utilization of 15%).

4.2 Simulation results

In this subsection, we first compare our proposed iCODE with benchmark, called NDR. Next, investigate iCODE under various settings to demonstrate its effectiveness.

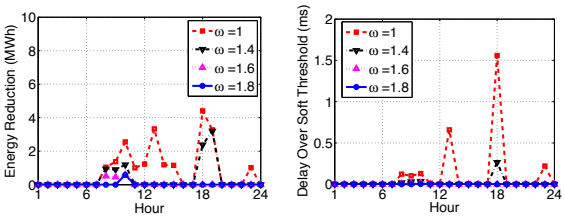
Benchmark: We choose the scenario in which no tenants participate in demand response as our benchmark, called NDR (Non-Demand Response), which is the status quo in colocation.

Comparison between iCODE and NDR. We now compare iCODE with NDR in Fig. 3. We first show the energy reduction by iCODE compared to NDR in Fig. 3(a). It can be seen that more than 4MWh energy reduction per hour can be achieved, which is a fairly significant energy reduction (equivalent to thousands of households) and demonstrates the big potential of colocation demand response. Next, we show the hourly energy consumptions by iCODE and NDR in Fig. 3(b), indicating that more than 50% energy can be slashed in some hours due to the low server utilization in colocation (e.g., 6-12% [17]). Fig. 3(c) shows the monetary incentive received by different tenants. We notice that there is some "residual" incentive paid to colocation operator by LSE, because sometimes tenants do not seek as high incentives as LSE provides. Fig. 3(d) shows the barely-



(a) (b)

Figure 4: Impact of workload overprediction. **(a)** Energy reduction. **(b)** Tenant #1's delay over soft threshold.



(a) (b)

Figure 5: Impact of greediness. **(a)** Energy reduction. **(b)** Tenant #1's delay over soft threshold.

noticeable performance degradation experienced by tenants compared to their soft delay thresholds. There is an up to around 1.5ms increase in average delay beyond the threshold for delay sensitive tenants #1 and 0.5ms for tenant #2. Tenant #3 has 30s delay, which is acceptable for delay tolerant workloads. If tenants cannot tolerate any delay exceeding their thresholds, they can easily remove those bids resulting in intolerable delay performance (see Section 3.3).

Impact of workload overprediction. To cope with unexpected possible traffic spikes, tenants can either turn on more servers as a backup or deliberately overestimate the workload arrival rate by a certain overprediction factor $\phi \geq 0$: the higher ϕ, the more overpredicts. We choose the later approach. Intuitively, when tenants are more conservative and tend more to overpredict workloads, fewer number of servers will be turned off. However, Fig. 4 shows that even when tenants overestimate the workloads by 30%, the energy reduction for demand response is not significantly compromised. We choose 30% because recent studies have shown that the workload prediction error is typically within 30% [26].

Impact of greediness. Tenants may be greedier in the sense that they desire more than their true costs for turning off servers. Here, we increase the greediness factor w_i for tenants. Equivalently, this captures the scenarios that tenants are less willing to participate in demand response unless they are provided sufficiently large incentives. Fig. 5 shows that as tenants are becoming more greedy, the performance becomes better and the energy reduction decreases. Nonetheless, we note that asking for higher payments than actual costs may not be of tenants' interests, because doing so will reduce tenants' financial rewards yet without improving their delay performances (as seen by comparing Fig. 3(d) and Fig. 5(b)).

5 Related work

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

• **Data center optimization:** Optimizing data center operation has received a surging interest recently [8,10,22]. Notably, turning on/off servers based on time-varying incoming workloads is a promising approach to enabling "power proportionality" and reducing energy consumption/cost of data centers [20, 25, 45]. By exploring geographic diversities, optimizing load balancing among multiple data centers can minimize electricity cost [36] and reduce carbon footprint [14]. These studies, however, are all intended for owner-operated data centers and hence cannot be directly applied to colocation unless tenants are properly incentivized.

• **Data center demand response:** Data centers are promising participants in demand response programs. For example, [16] conducts field tests, showing that data centers can reduce energy consumption by 10-25% upon receiving demand response signals. Focusing on owner-operated data centers, [4, 15, 23] study resource management optimization for demand response and frequency regulation in power grid, and [24,28,44] consider the interactions between data centers and utilities and study pricing strategies by utilities. [5] addresses frequency regulation by controlling facility energy consumption via battery charging/discharging, but this technique is difficult to scale due to limited battery size in practice [43].

To our best knowledge, our study makes the first step towards unifying interests of colocation operator and tenants to unleash the promising potential of colcation demand response.

6 Conclusions

In this paper, we studied colocation demand response and proposed an reverse auction-based incentive mechanism, iCODE, which offers tenants with financial rewards for energy reduction. iCODE just requires a lightweight and "non-intrusive" control module that can be automated during run time. We performed a trace-based simulation study to show that iCODE can reduce the hourly energy consumption by over 50%, which is a fairly large amount of energy reduction for demand response programs. iCODE is a first-of-its-kind mechanism to break "split incentive" between colocation operator and tenants, and can also be extended to address other issues in colocation (e.g., energy inefficiency).

References

[1] Colocation market - worldwide market forecast and analysis (2013 - 2018), http://www.marketsandmarkets.com/Market-Reports/colocation-market-1252.html.

[2] Telegeography colocation database, http://www.telegeography.com/research-services/colocation-database/.

[3] U.S. Federal Leadership in Environmental, Energy and Economic Performance - EXECUTIVE ORDER 13514, http://www.whitehouse.gov/administration/eop/ceq/sustainability.

[4] AIKEMA, D., SIMMONDS, R., AND ZAREIPOUR, H. Data centres in the ancillary services market. In *IGCC* (2012).

[5] AKSANLI, B., AND ROSING, T. S. Providing regulation services and managing data center peak power budgets. In *DATE* (2014).

[6] BOYD, S., GHOSH, A., AND MAGNANI, A. Branch and bound methods, http://www.stanford.edu/class/ee392o/bb.pdf, 2003.

[7] BOYD, S., AND VANDENBERGHE, L. *Convex Optimization*. Cambridge University Press, 2004.

[8] CHENG, D., JIANG, C., AND ZHOU, X. Heterogeneity-aware workload placement and migration in distributed sustainable datacenters. In *IPDPS* (2014).

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

[10] DENG, N., STEWART, C., GMACH, D., AND ARLITT, M. F. Policy and mechanism for carbon-aware cloud applications. In *NOMS* (2012).

[11] ENAXIS CONSULTING. Pricing data center co-location services, 2009, http://enaxisconsulting.com/downloads/2/67f7fb873eaf29526a11a9b7ac33bfac/1317636458_data_center_pricing.pdf.

[12] FEDERAL ENERGY REGULATORY COMMISSION. Assessment of demand response and advanced metering, 2012.

[13] GANDHI, A., HARCHOL-BALTER, M., DAS, R., AND LEFURGY, C. Optimal power allocation in server farms. In *SIGMETRICS* (2009).

[14] GAO, P. X., CURTIS, A. R., WONG, B., AND KESHAV, S. It's not easy being green. *SIGCOMM Comput. Commun. Rev. 42*, 4 (Aug. 2012), 211–222.

[15] GHAMKHARI, M., AND MOHSENIAN-RAD, H. Energy and performance management of green data centers: a profit maximization approach. In *SmartGridCom* (2012).

[16] GHATIKAR, G., GANTI, V., MATSON, N. E., AND PIETTE, M. A. Demand response opportunities and enabling technologies for data centers: Findings from field studies, 2012.

[17] GLANZ, J. Power, pollution and the internet. In *The New York Times* (Sep. 22, 2012).

[18] HARBOR RIDGE CAPITAL. Colocation data centers: Overview, trends & m&a, http://www.harborridgecap.com.

[19] HE, Y., ELNIKETY, S., LARUS, J., AND YAN, C. Zeta: Scheduling interactive services with partial execution. In *SOCC* (2012).

[20] HOELZLE, U., AND BARROSO, L. A. *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines*. Morgan and Claypool Publishers, 2009.

[21] KERRIGAN, J., AND HOROWITZ, D. Avison young data center practice newsletter, Jan. 2014.

[22] LI, S., ABDELZAHER, T., AND YUAN, M. Tapa: Temperature aware power allocation in data center with map-reduce. In *IGCC* (2011).

[23] LI, S., BROCANELLI, M., ZHANG, W., AND WANG, X. Data center power control for frequency regulation. In *PES* (2013).

[24] LI, Y., CHIU, D., LIU, C., PHAN, L. T., GILL, T., AGGARWAL, S., ZHANG, Z., LOO, B. T., MAIER, D., AND MCMANUS, B. Towards dynamic pricing-based collaborative optimizations for green data centers. In *ICDEW* (2013).

[25] LIN, M., WIERMAN, A., ANDREW, L. L. H., AND THERESKA, E. Dynamic right-sizing for power-proportional data centers. In *IEEE Infocom* (2011).

[26] LIU, Z., CHEN, Y., BASH, C., WIERMAN, A., GMACH, D., WANG, Z., MARWAH, M., AND HYSER, C. Renewable and cooling aware workload management for sustainable data centers. In *SIGMETRICS* (2012).

[27] LIU, Z., LIN, M., WIERMAN, A., LOW, S. H., AND ANDREW, L. L. Greening geographical load balancing. In *SIGMETRICS* (2011).

[28] LIU, Z., LIU, I., LOW, S., AND WIERMAN, A. Pricing data center demand response. In *Sigmetrics* (2014).

[29] LIU, Z., WIERMAN, A., CHEN, Y., RAZON, B., AND CHEN, N. Data center demand response: avoiding the coincident peak via workload shifting and local generation. In *SIGMETRICS* (2013).

[30] MA, J., DENG, J., SONG, L., AND HAN, Z. Incentive mechanism for demand side management in smart grid using auction. *IEEE Trans. Smart Grid* (2014 (PrePrint)).

[31] MARTELLO, S., AND TOTH, P. *Knapsack Problems: Algorithms and Computer Implementations*. John Wiley & Sons, Inc., New York, NY, USA, 1990.

[32] MOHSENIAN-RAD, H., WONG, V. W. S., JATSKEVICH, J., SCHOBER, R., AND LEON-GARCIA, A. Autonomous demand side management based on game-theoretic energy consumption scheduling for the future smart grid. *IEEE Trans. Smart Grid 1*, 3 (Dec. 2010), 320–331.

[33] NEW YORK ISO. http://www.nyiso.com/.

[34] NOVET, J. Colocation providers, customers trade tips on energy savings, Nov. 2013.

[35] PALASAMUDRAM, D. S., SITARAMAN, R. K., URGAONKAR, B., AND URGAONKAR, R. Using batteries to reduce the power costs of internet-scale distributed networks. In *SoCC* (2012).

[36] QURESHI, A., WEBER, R., BALAKRISHNAN, H., GUTTAG, J., AND MAGGS, B. Cutting the electric bill for internet-scale systems. In *SIGCOMM* (2009).

[37] RAO, L., LIU, X., XIE, L., AND LIU, W. Reducing electricity cost: Optimization of distributed internet data centers in a multi-electricity-market environment. In *IEEE Infocom* (2010).

[38] REGULATORY ASSISTANCE PROJECT. Electricity regulation in the US: A guide, Mar. 2011, `www.raponline.org`.

[39] SILICON VALLEY POWER. Data center program, `http://siliconvalleypower.com/index.aspx?page=2088`.

[40] U.S. DOE. `http://energy.gov/`.

[41] U.S. GREEN BUILDING COUNCIL. Leadership in energy & environmental design, `http://www.usgbc.org/leed`.

[42] VASUDEVAN, V., PHANISHAYEE, A., SHAH, H., KREVAT, E., ANDERSEN, D. G., GANGER, G. R., GIBSON, G. A., AND MUELLER, B. Safe and effective fine-grained tcp retransmissions for datacenter communication. *SIGCOMM Comput. Commun. Rev. 39*, 4 (Aug. 2009), 303–314.

[43] WANG, D., REN, C., SIVASUBRAMANIAM, A., URGAONKAR, B., AND FATHY, H. Energy storage in datacenters: what, where, and how much? In *SIGMETRICS* (2012).

[44] WANG, H., HUANG, J., LIN, X., AND MOHSENIAN-RAD, H. Exploring smart grid and data center interactions for electric power load balancing. *SIGMETRICS Perform. Eval. Rev. 41*, 3 (Jan. 2014), 89–94.

[45] YAO, Y., HUANG, L., SHARMA, A., GOLUBCHIK, L., AND NEELY, M. J. Data centers power reduction: A two time scale approach for delay tolerant workloads. In *Infocom* (2012).

[46] ZHONG, H., XIE, L., AND XIA, Q. Coupon incentive-based demand response: Theory and case study. *IEEE Trans. Power Systems 28*, 2 (May 2013), 1266–1276.