A Contract Design Approach for Colocation Data Center Demand Response

Kishwar Ahmed
Florida International University
kahme006@fiu.edu

Mohammad A. Islam
Florida International University
misla012@fiu.edu

Shaolei Ren
University of California, Riverside
sren@ece.ucr.edu

***Abstract*—Demand response programs maintain transmission stability in power grid through reducing electricity use during peak period, making grid more efficient and robust. While numerous demand response programs are currently being deployed by utility companies, we focus on emergency demand response program, which is critical to ensure reliability during emergency situations. As a key participant in such program, we consider a critical type of data center: multi-tenant colocation data center (or colocation), where multiple tenants mange their own servers in shared space but typically lack incentives to reduce energy for demand response. To enable multi-tenant data center demand response, we propose a contract-based mechanism, called Contract-DR, which offers financial incentives to tenants to shed energy during emergency situations, reducing the usage of cost-ineffective and environmentally-unfriendly diesel generation. We conduct theoretical analysis to prove the optimality of Contract-DR and also validate it through a trace-based study.**

I. Introduction

As a national strategic plan for next-generation grid innovation, demand response programs motivate end-use customers to reduce their power consumption through financial incentives during times when wholesale electricity price is high and/or when the power system becomes unreliable due to generation shortage [1]. In recent years, participation in demand response has increased dramatically: e.g., Baltimore Gas and Electric (a utility company in Maryland, U.S.) has reported a 180% increase in its demand response capability from 2010 to 2012 [2]. Another report [3] projects a doubling of participation sites in 2020, from current 10.3 million sites to 21.9 million.

Among various programs, emergency demand response (EDR) is the most widely adopted, taking up 87% of all demand response capabilities across the U.S. [4]. During an EDR event (e.g., extremely cold/hot weather, natural disasters), energy consumers are coordinated to shed loads and collectively prevent the power grid from getting into blackouts, potentially saving billions of dollars' loss. Thus, many demand response resources, e.g., office buildings and residential customers, are emerging and sought to participate in EDR. In particular, data centers have been increasingly recognized as an appealing and critical EDR resource, because of their widespread presence and huge energy consumption. For example, on July 22, 2011, large areas of the U.S. and Canada experienced a significant shortage in electricity generation to meet the demand, and hundreds of data centers shed energy consumption and help avoid a nation-wide blackout [5]. Nonetheless, today, data centers typically participate in EDR by turning on their backup diesel generators, which are neither cost effective nor environmentally clean. In fact, diesel generation can emit as 50-60 times of toxic particles as typical power plants and has become a major pollution source in many regions, e.g., California (a major data center market) [6].

As a consequence, towards greening data center EDR, there have been many recent studies that focus on cutting server energy consumption to substitute the highly-undesired diesel generation. For example, various IT computing knobs, e.g., turning off servers, deferring delay-tolerant workloads and/or migrating workloads to other sites, have been explored to enable data center demand response [7]–[10].

The existing efforts on data center demand response, although unarguably encouraging, have been primarily focused on owner-operated/single-tenant data centers like Google and Microsoft, for which the currently available IT knobs can be easily leveraged to green the participation in EDR as demonstrated by research studies and LBNL tests [11]. By significant contrast, in this paper, we focus on demand response by another *distinct* yet even more important type of data center — multi-tenant colocation data center (also called "colocation" or simply "colo"). In a colocation, multiple tenants house and manage their own physical servers separately, whereas the data center operator is responsible for providing non-IT facility support, such as reliable power distribution, cooling, etc. Thus, colocation is managed in a decentralized/uncoordinated manner, significantly differing from an owner-operated data center where the operator fully controls everything, including both non-IT facility and IT computing.

Colocation is a widely-existing and critical segment of data center industry, with over 1,400 data centers in the U.S. alone [12], but it has been much less studied compared to Google-type data centers. It serves almost all industry sectors, including content delivery providers (e.g., Akamai), top-ranking websites (e.g., Wikipedia), medium-scale cloud providers (e.g., VMware), and even IT giants like Amazon that house some of their servers in colocations to complement their own data center infrastructure [13]. By a recent study [14], multi-tenant colocations consume as *five* times energy as those Google-type data centers combined altogether.

Most importantly, colocation is a more appealing and better EDR resource than Google-type data center. Many colocations are located in major cities where tenants operate business so that they can quickly troubleshoot their servers if any-

978-1-4673-8388-2/15/$31.00 ©2015 IEEE

TABLE I
LIST OF KEY NOTATIONS.

Notation	Description
Θ	Set of tenant types, where $\Theta = \{\theta_1, \cdots, \theta_I\}$
ξ_{θ_i}	Unit cost of reducing energy for tenant of type θ_i
n_{θ_i}	$\#$ of servers turned off by tenant of type θ_i
α	Battery cost parameter
Δe_{th}	Energy reduction target set by utility
$\Delta e(\theta_i)$	Energy reduction by tenant of type θ_i
m_{θ_i}	# of tenants of type θ_i

thing goes wrong. By contrast, many megascale Google-type data centers are located in rural areas with little population. Nonetheless, it is major cities where EDR is mostly needed, because electricity demand over there is more likely to exceed or approach the grid capacity. Thus, colocations are key participants in EDR, as demonstrated by real examples [5].

While colocations are ideal participants in EDR, they typically participate by using diesel generation and greening colocation demand response faces a *unique challenge* that Google-type data centers do not have: colocation operator does not have control over tenants' servers and hence cannot directly modulate IT server energy to substitute diesel generation; and on the other hand, tenants managing the servers may not have incentives to contribute to green EDR. Thus, to motivate tenants' energy reduction for greening colocation EDR, we propose a *contract design* approach, Contract-DR, using which the colocation operator offers a set of *contracts* (i.e., energy reduction and f nancial rewards) to tenants and tenants can voluntarily select none or one of the contracts to accept. Contract-DR is designed such that tenants can have different types (in terms of cost incurred for energy reduction) and each tenant chooses contract item for its own type. We consider both discrete and continuous type of tenants in our contract design mechanism. To validate Contract-DR, we conduct a trace-based simulation involving a colocation (consisting of multiple tenants), and show that Contract-DR reduces overall energy consumption of colocation during emergency periods at a minimum cost for the colocation operator.

II. MODEL

We consider a discrete-time model by dividing the time horizon into equal-length time slots (e.g., one hour each). To simplify the notations, we exclude time notations from the model. Table I summarizes key notations.

A. Colocation model

We consider a colocation data center with a set of N tenants. Tenants are categorized to different types based on their energy reduction cost (also called "inconvenience cost"). Tenants with same inconvenience cost are grouped in the same type. The types are denoted by θ_i, where each tenant is included in one of the types, and $\Theta = \{\theta_1, \theta_2, \cdots, \theta_I\}$ denotes set consisting of all the tenant types. The number of tenants in type θ_i is denoted by m_{θ_i}. We assume that tenant of type θ_i owns w_{θ_i} homogeneous servers while our model is also applicable if tenants own multiple heterogeneous servers. We denote the service rate of a server owned by tenant of type θ_i as μ_{θ_i}. The workload arrival for tenant of type θ_i is represented by $\lambda_{\theta_i} = [0, \lambda_{max}]$.

Power consumption. We denote the total power consumption of tenant of type θ_i's servers as $p_{\theta_i} = w_{\theta_i} \cdot \left[e_{0,\theta_i} + e_{c,\theta_i} \frac{\lambda_{\theta_i}}{w_{\theta_i} \cdot \mu_{\theta_i}}\right]$, where e_{0,θ_i} denotes the static server power and e_{c,θ_i} is the computing power incurred only when a server is processing workloads for tenant of type θ_i's server. Denoting $w'_{\theta_i} = w_{\theta_i} - n_{\theta_i}$ as the amount of servers remaining after turning off n_{θ_i} servers, total power consumption of tenant of type θ_i's servers after energy reduction becomes $p'_{\theta_i} = (w_{\theta_i} - n_{\theta_i}) \cdot \left[e_{0,\theta_i} + e_{c,\theta_i} \frac{\lambda_{\theta_i}}{(w_{\theta_i} - n_{\theta_i}) \cdot \mu_{\theta_i}}\right]$. After simple calculation, it follows that, total energy reduction by tenant of type θ_i can be represented as $\Delta e(\theta_i) = (p_{\theta_i} - p'_{\theta_i}) \cdot T = n_{\theta_i} \cdot e_{0,\theta_i} \cdot T$, where T is the duration of a time slot.

Energy storage device. While participating tenants contribute to load reduction, a limitation is that tenants may not be able to collectively reduce energy and therefore colocation may fail to achieve target energy reduction. Failure to meet energy reduction target may hinder participation of colocation in EDR, potentially incurring penalty fees, and therefore we make provision for remaining required energy reduction through discharge of energy storage devices (ESD, typically, diesel generator for EDR [15]). We denote the ESD discharge amount by e_b, at a unit cost α per kWh.

B. Tenant utility

When shedding energy, tenants will experience some "costs", which can be interpreted as, e.g., revenue loss due to performance degradation. We refer to the cost as "inconvenience cost". We denote the cost incurred by tenant of type θ_i as function of energy reduction and def ne as following:

$$v(\theta_i, \Delta e(\theta_i)) = \xi_{\theta_i} \cdot c(\Delta e(\theta_i)), \quad (1)$$

where ξ_{θ_i} denotes cost of energy reduction, which is private information to tenants, and $c(\Delta e(\theta_i))$ represents cost as a general function of energy reduction of tenant of type θ_i. We now def ne the utility of each tenant of type θ_i as

$$u(\theta_i, \Delta e(\theta_i)) = r(\theta_i) - v(\theta_i, \Delta e(\theta_i)), \forall \theta_i \in \Theta, \quad (2)$$

where $v(\theta_i, 0) = 0$ and $r(\theta_i)$ denotes reward awarded to tenant of type θ_i.

III. FORMULATION AND ALGORITHM

In this section, we formulate the contract design problem and outline the algorithm. We f rst give two def nitions of constraints that are essential for designing *direct revelation contract*.

Def nition 1 (IR constraint): Individual Rationality (IR) constraint or participation constraint ensures that participants in contract mechanism achieve non-negative pay-off. Mathematically, for type θ_i, IR constraint can be described as following:

$$r(\theta_i) - v(\theta_i, \Delta e(\theta_i)) \geq 0. \quad (3)$$

Definition 2 (IC constraint): Incentive Compatibility (IC) constraint represents the fact that tenant chooses its own type to maximize utility. It reflects that, although tenant of type θ_i can choose another type θ_i', the tenant chooses its own type θ_i and decides to reduce $\Delta e(\theta_i)$ amount of energy for utility maximization. Mathematically, IC constraint can be specified as following:

$$r(\theta_i) - v(\theta_i, \Delta e(\theta_i)) \geq r(\theta_i') - v(\theta_i', \Delta e(\theta_i')). \tag{4}$$

The objective of colocation operator is to determine the optimal allocation of energy reduction to each customer/tenant and corresponding reward such that its total cost for EDR is minimized and also IC and IR constraints are satisfied. Total cost includes reward provided to the tenants and also the ESD discharge cost. Mathematically, the objective can be represented as following:

$$\min_{(\Delta e(\theta_i), r(\theta_i))} \sum_{\theta_i \in \Theta} m_{\theta_i} \cdot r(\theta_i) + \alpha \cdot e_b. \tag{5}$$

Besides the constraints in Eqs. in (3) and (4), the following constraint also needs to be satisfied:

$$\gamma \cdot \sum_{\theta_i \in \Theta} m_{\theta_i} \cdot \Delta e(\theta_i) + e_b = \Delta e_{th}, \tag{6}$$

where Δe_{th} denotes energy reduction constraint for EDR set by utility and γ denotes power usage effectiveness (PUE) of colocation that converts the IT energy reduction to the facility-level energy reduction.

A. Contract design with complete information

First, as a theoretical baseline case, we assume that colocation operator has complete knowledge of type of each tenant. Since the colocation operator is aware of the information of tenants' type, it can design a contract such that each tenant chooses contract element only specific for its own type. We assume that types are discrete, i.e., there are finite number of entries in the contract list. The optimization problem for colocation operator now becomes to choose contract item consisting of the following pair: $\{(\Delta e(\theta_i), r(\theta_i)), \forall \theta_i \in \Theta\}$, to optimize colocation operator's cost:

$$\begin{aligned} &\min_{(\Delta e(\theta_i), r(\theta_i))} \sum_{\theta_i \in \Theta} m_{\theta_i} \cdot r(\theta_i) + \alpha \cdot e_b, \\ &\text{s.t., Eq. (6) and IR constraint given in Eq. (3).} \end{aligned} \tag{7}$$

Now, the IR constraint is tight at the optimality, and therefore, the colocation operator will leave zero utility to the tenants. Reward received by type-θ_i user can be denoted by $r^*(\theta_i) = v(\theta_i, \Delta e(\theta_i))$. For each type θ_i, colocation operator needs to decide the optimal energy reduction based on the following optimization:

$$\begin{aligned} &\min_{\Delta e(\theta_i)} \sum_{\theta_i \in \Theta} m_{\theta_i} \cdot v(\theta_i, \Delta e(\theta_i)) + \alpha \cdot e_b, \\ &\text{s.t., Eq. (6).} \end{aligned} \tag{8}$$

The above optimization problem is convex and can be efficiently solved. Although Eq. (8) represents simplified optimization objective (without explicitly specifying IR and IC constraints), in later subsection (section III-C), we show the optimaility of the *reduced* optimization.

B. Contract design with incomplete information

Next, we discuss the case where colocation operator lacks information of tenant type, which is private information to the tenant. Based on prior knowledge of tenants' type distribution, colocation operator maximizes his own expected utility. We denote tenants' numbers of all types as $\{m_{\theta_i}\}_{\theta_i \in \Theta}$, which are random variables and follow a discrete binomial distribution (as also considered in [16]). The optimization objective in such a scenario can be given as below:

$$\begin{aligned} &\min_{(\Delta e(\theta_i), r(\theta_i))} \sum_{\theta_i \in \Theta} \mathbb{E}[m_{\theta_i} \cdot r(\theta_i) + \alpha \cdot e_b | \{m_{\theta_i}\}_{\theta_i \in \Theta}], \\ &\text{s.t., Eqs. (3), (4), and (6).} \end{aligned} \tag{9}$$

Note that, IC and IR constraints bind at the optimal solution [17]. Therefore, we get the following:

$$\begin{aligned} &r^*(\theta_1) = \xi_{\theta_1} \cdot c(\Delta e(\theta_1)), \\ &r^*(\theta_i) = r^*(\theta_{i-1}) + \xi_{\theta_i}(c(\Delta e(\theta_i)) - c(\Delta e(\theta_{i-1}))). \end{aligned} \tag{10}$$

The optimization objective now becomes:

$$\begin{aligned} &\min_{\Delta e(\theta_i)} \sum_{\theta_i \in \Theta} \mathbb{E}[m_{\theta_i} \cdot r^*(\theta_i) + \alpha \cdot e_b | \{m_{\theta_i}\}_{\theta_i \in \Theta}] \\ &\text{s.t., Eq. (6),} \end{aligned} \tag{11}$$

where $r^*(\theta_i)$ is defined in (10). As in [18], we resort to exhaustive search to solve the above optimization problem. The optimality of reduced optimization problem in Eq. (11) is given in the next subsection.

C. Feasibility and optimality proofs

In Eqs. (8) and (11), we have provided a simplified formulation of contract design for complete and incomplete information cases, respectively, without explicitly accounting for IC and IR constraints. Here, we show that the simplified formulation does not lose optimaility. We first denote the set of all tenant types as $\Theta = \{\theta_1, \cdots, \theta_I\}$. The following two conditions ensure feasibility and optimality of designed contracts:

Conditions for feasibility. We assume the contract set is $C = \{(\Delta e(\theta_i)), r(\theta_i), \forall \theta_i \in \Theta\}$ with tenant costs $\xi_{\theta_1} < \xi_{\theta_2} < \cdots < \xi_{\theta_I}$. Then the contract is feasible if and only if:

$$\begin{aligned} &r(\theta_1) - \xi_{\theta_1} c(\Delta e(\theta_1)) \geq 0, \\ &r(\theta_n) - \xi_{\theta_n} c(\Delta e(\theta_n)) \geq r(\theta_{n-1}) - \xi_{\theta_{n-1}} c(\Delta e(\theta_n)). \end{aligned}$$

Conditions for optimality. For the optimal solution, individual rationality condition for the lowest type and adjacent ICs are binding, and all other conditions can be ignored. Then, we can state the following:

$$\begin{aligned} &r^*(\theta_1) = \xi_{\theta_1} \cdot c(\Delta e(\theta_1)), \\ &r^*(\theta_n) = r^*(\theta_{n-1}) + \xi_{\theta_n}(c(\Delta e(\theta_n)) - c(\Delta e(\theta_{n-1}))). \end{aligned}$$

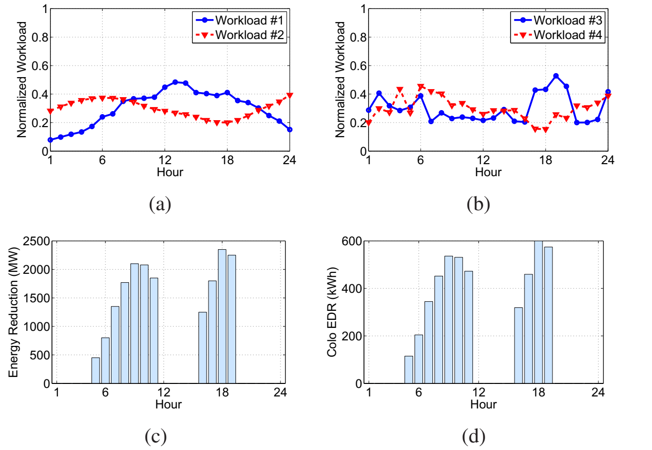


Fig. 1. Workload and energy reduction target. **(a)** & **(b)** Normalized workload trace, **(c)** Energy reduction target at PJM, **(d)** Scaled energy reduction target for colocation.

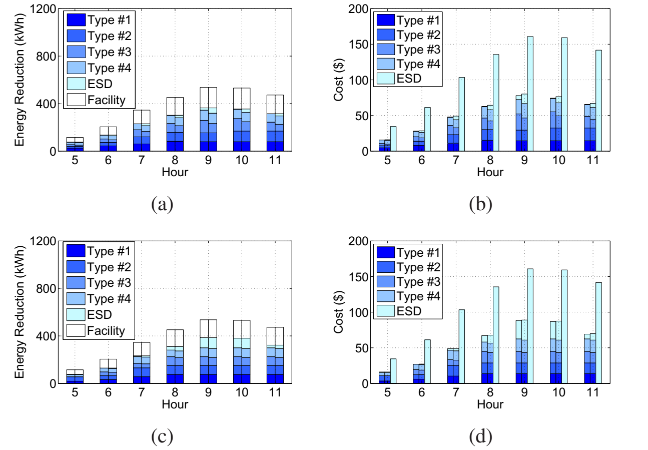


Fig. 2. Energy reduction and cost for Discrete-Complete and Discrete-Incomplete types. **(a)** & **(b)** 10 tenants, **(c)** & **(d)** 30 tenants.

Theorem 1: The simplified contract designs in Eqs. (8) and (11) minimize the colocation operator's cost while satisfying both IR and IC constraints.

Proof: The proof follows through mathematical induction and is provided in [19]. ■

IV. Simulation

In this section, we present trace-based simulation studies to validate our analysis and evaluate the performance of Contract-DR. First, we present the data sets we used for simulation purpose, and then we present the simulation results.

A. Data sets

We consider a colocation data center, with four tenants (denoted as Tenant #1, Tenant #2, Tenant #3 and Tenant #4), located at Council Bluffs - Iowa, since Iowa is a served by PJM Interconnection (a Regional Transmission Organization in the United States) [20]. We consider demand response program offered by PJM for our simulation settings. Later, we increase the number of tenants to 10 and 30 and distribute the tenants to four different types. The types we consider are denoted by θ_1, θ_2, θ_3 and θ_4. Colocation operator assumes the following distribution, where $|\theta_i|$ denotes number of tenants for type θ_i. For 10 tenants, the distribution is $|\theta_1| = 1$, $|\theta_2| = 4$, $|\theta_3| = 4$ and $|\theta_4| = 1$. For 30 tenants, the distribution is considered as: $|\theta_1| = 5$, $|\theta_2| = 10$, $|\theta_3| = 10$ and $|\theta_4| = 5$.

Workload: We evaluate four different workload traces, which we denote as "Workload #1", "Workload #2", "Workload #3" and "Workload #4". The workload traces were collected from four different sources: web service traces (e.g., Hotmail, Wikipedia), I/O log (Microsoft Research) and a university data center from May $1-7, 2012$. The workloads are normalized with respect to the maximum capacity of each tenant's servers, while maintaining average utilization of 30%. Figs. 1(a) and 1(b) show a snapshot of the traces.

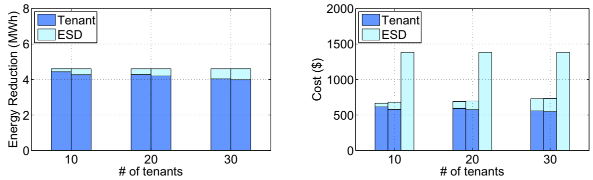
Energy reduction target: We follow the demand response requirement provided by PJM Interconnection on January 7, 2014 [20] to generate the target energy reduction throughout the time period. Fig. 1(c) shows emergency energy reduction target on January 7, 2014. We normalize emergency energy reduction target set by PJM during each time period and scale to 30% of colocation peak power to determine energy reduction target for colocation during each hour. Fig. 1(d) shows scaled-down hourly emergency reduction target for the colocation.

Different parameters: Values of energy reduction cost were set to 2.75¢/server, 3¢/server, 3.25¢/server and 3.5¢/server for Tenant #1, Tenant #2, Tenant #3 and Tenant #4, respectively. The value of PUE of data center was set to 1.5. Cost of using diesel generator is considered to be 300\$/MWh. Service rates of Tenant #1, Tenant #2, Tenant #3 and Tenant #4 are $360k$, $180k$, 30 and $270k$ jobs/hour, respectively. Maximum delay threshold for Tenant #1, Tenant #2, Tenant #3 and Tenant #4 are 50ms, 100ms, 600s and 60ms, respectively (where delay thresholds are calculated based on 80% utilization of the servers). Idle and peak power of each server was set to 150W and 250W, respectively.

B. Simulation results

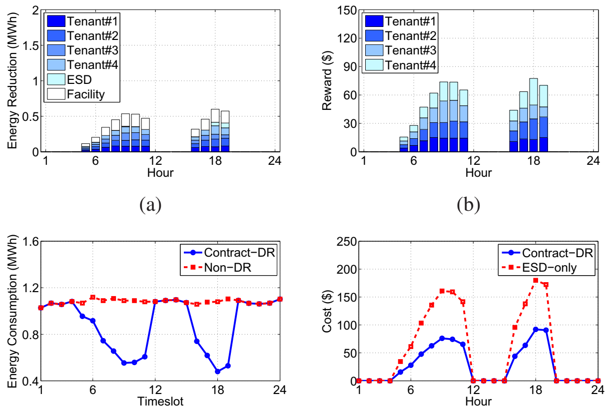
We now present simulation results we obtain through running simulation on the data set. We first compare different contract-based algorithms (i.e., Discrete-Complete: discrete types with known distribution and Discrete-Incomplete: discrete types with unknown distribution) and show the impact of increasing tenants # of these types. Next, we compare Contract-DR (Discrete-Complete type) with benchmark algorithms: Non-DR (EDR participation is not considered) and ESD-only (where target energy reduction is achieved through stored energy by energy storage device).

1) Comparison of different contract designs: Fig. 2 shows comparison of Discrete-Complete and Discrete-Incomplete. Figs. 2(a) - 2(b) demonstrate energy reduction and reward for 10 tenants during emergency hours, while Figs. 2(c) - 2(d) show energy reduction and reward for 30 tenants. In Figs. 2(a) and 2(c), two bars represent Discrete-Complete and



(a) (b)

Fig. 3. Comparison of different types. **(a)** Energy reduction, **(b)** Cost. In each group of the bars, first bar for Discrete-Complete, second bar for Discrete-Incomplete, third bar for ESD-only.



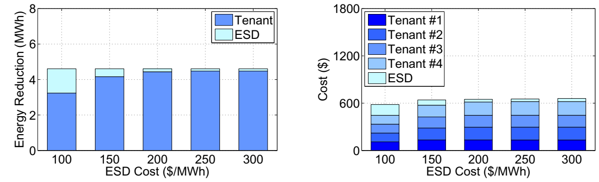
(a) (b)

(c) (d)

Fig. 4. Comparison. **(a)** Energy, **(b)** Total reward, **(c)** Energy consumption, **(d)** Colocation's cost for demand response.

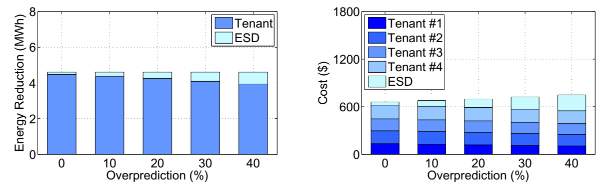
Discrete-Incomplete, respectively, while in Figs. 2(b) and 2(d), three bars represent Discrete-Complete, Discrete-Incomplete and ESD-only, respectively. As can be seen in Figs. 2(a) and 2(c), colocation achieves target energy reduction through tenants' and ESD's participation in EDR during emergency hours for both Discrete-Complete and Discrete-Incomplete. Also as evident in Figs. 2(b) and 2(d), Discrete-Complete incurs lower cost than Discrete-Incomplete, while ESD-only incurs higher cost than both Discrete-Complete and Discrete-Incomplete throughout the emergency hours.

Fig. 3 compares the three algorithms, Discrete-Complete, Discrete-Incomplete and ESD-only, in terms of total energy reduction and total cost during EDR. As can be seen from Fig. 3(a), both Discrete-Complete and Discrete-Incomplete achieve target energy reduction. However, tenants' participation in Discrete-Complete is higher than tenants' participation in Discrete-Incomplete (although the difference between tenants' participation in these two algorithms gradually decreases with increase of # of tenants, as can be seen in the figure). Fig. 3(b) shows cost comparison for different algorithms. Although Discrete-Incomplete incurs more total cost than Discrete-Complete, the cost difference gradually decreases and becomes approximately close to the Discrete-Complete algorithm with increase of number of tenants. ESD-only incurs significantly higher cost than both Discrete-Complete and Discrete-Incomplete, as evident from the figure.



(a) (b)

Fig. 5. Impact of ESD price. **(a)** Energy reduction, **(b)** Cost.



(a) (b)

Fig. 6. Workload overprediction. **(a)** Energy reduction, **(b)** Cost.

2) Comparison with non-demand response approach: Fig. 4 demonstrates comparison of Contract-DR (Discrete-Complete type) with benchmark algorithm, Non-DR, which does not consider EDR participation. Fig. 4(a) shows energy reduction achieved by different participants throughout the EDR periods. Tenants participate in energy reduction, whenever the colocation operator sets energy reduction target determined by utility. Moreover, ESD is also used to offset the remaining energy reduction shortage, whenever tenants are unable to achieve the required energy reduction objective. Fig. 4(b) shows distribution of reward amount among different tenants. It is evident that different tenants win different reward amounts from the colocation operator in proportion to the energy reduction amount achieved by the tenant. Fig. 4(c) compares Contract-DR with Non-DR in terms of energy consumption at each time period. As can be seen from the figure, Contract-DR significantly reduces energy during emergency time period, as compared to Non-DR. Fig. 4(d) shows the total cost incurred by Contract-DR and compares with ESD-only. As can be seen in the figure, Contract-DR incurs lower cost than ESD-only during each emergency hour.

3) Impact of ESD cost: We now vary ESD cost parameter and show the effect of change on Fig. 5. The parameter was varied from 150\$/MWh to 300\$/MWh. With lower value of ESD cost, Contract-DR becomes more inclined towards achieving energy reduction target through higher utilization of ESD. Fig. 5(a) shows contribution of energy reduction by tenants and ESD for different values of ESD costs. As evident from the results, with a lower ESD cost, ESD contributes more to energy reduction. Tenants, on the other hand, participate more on energy reduction when the ESD cost is comparatively higher. Fig. 5(b) shows the total cost incurred by colocation for different ESD cost values. Higher ESD value means increased participation of tenants and correspondingly higher reward being awarded to the tenants.

4) Workload overprediction: In practice, tenants may not perfectly know the amount of workloads they will receive in the upcoming time slot (e.g., hour). Thus, to avoid server overloading, tenants may intentionally leave a margin by overestimating the workloads. Fig. 6 shows effect of workload overprediction, where workload is overpredicted from 0% to 40% with an increment of 10%. Fig. 6(a) shows the percentage of energy reduction by tenants compared to energy reduction through EDR. With the increase of workload overprediction, it becomes increasingly diff cult for tenants to participate in energy reduction (since they are more conservative). Therefore, to achieve target energy reduction, ESD shares an increasing portion of energy reduction. Fig. 6(a) validates this fact: with 0% workload increase, tenants contribute to the most significant share of energy reduction, which gradually decreases with more workload overprediction, while ESD contributes more to the energy reduction. Fig. 6(b) shows cost incurred by colocation during EDR hours, where it can be seen that ESD cost increases with more workload overprediction due to increased participation of ESD in EDR.

V. Related Work

Data center demand response has received much interest in recent years due to its huge economic value [7]. For example, considering owner-operated data center, [21], [22] optimize data center resource management for ancillary services. [23] leverages both resource management techniques and ESD charging/discharging for data center demand response with the goal of shaving its peak power usage. However, all these studies are related to owner-operated data centers, and hence are not applicable for colocations where servers are managed by individual tenants rather than by a central operator. A few recent studies [24]–[26] have begun to examine multi-tenant colocation demand response, but they all assume that tenants can *proactively* propose to reduce loads by submitting their bids. This may be subject to tenants' cheating behaviors in certain scenarios [24] (which was later resolved by [25]), and also requires tenants to be "intelligent" enough to calculate their optimal bids. By contrast, this paper takes a different approach centering around the colocation operator: Contract-DR is easier for implementation in practice since it transfers the computational burden to the colocation operator such that tenants only need to say "yes or no" to the offered contracts rather than proposing their own bids. Moreover, we focus on minimizing the operator's cost, rather than the "social cost" considered in [25] that does not necessarily translate into a lower cost for colocation operator.

VI. Conclusions

In this paper, we studied multi-tenant data center demand response, which is a critical resource for maintaining power grid safety during emergency events. We proposed a contract-based incentive mechanism, Contract-DR, that rewards tenants for shedding energy during EDR while achieving a target energy reduction at a minimum cost for the colocation operator. We considered cases of both complete and incomplete information, and provide a set of conditions ensuring IC and IR constraints. We also performed a trace-based simulation study and showed that Contract-DR sheds energy for EDR and also reduces the colocation operator's cost.

Acknowledgement

This work was supported in part by the U.S. National Science Foundation under grants CNS-1423137 and CNS-1453491.

References

[1] "U.S. department of energy. benef ts of demand response in electricity markets and recommendations for achieving them. a report to the united states congress pursuant to section 1252 of the energy policy act of 2005," 2006.

[2] "Demand response in the US electricity market," http://theenergycollective.com/rasika-athawale/195536/demand-response-us-electricity-market.

[3] H. Marianne and W. Eric, "Market data: Demand response. residential, commercial, and industrial demand response participation and sites, load curtailment, and spending: Global and regional market sizing and forecasts," Navigant Research. 2013.

[4] K. Managan, "Demand repsonse: A market overview," 2014, http://enaxisconsulting.com.

[5] A. Misra, "Responding before electric emergencies," http://www.afcom.com/digital-library/pub-type/communique/responding-before-electric-emergencies/.

[6] Santa Babara County, "Air Pollution Control District," http://www.ourair.org/do-you-really-need-a-diesel-generator/.

[7] Z. Liu, I. Liu, S. Low, and A. Wierman, "Pricing data center demand response," in *SIGMETRICS*, 2014.

[8] A. Wierman, Z. Liu, I. Liu, and H. Mohsenian-Rad, "Opportunities and challenges for data center demand response," in *IGCC*, 2014.

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

[10] H. Chen, M. Caramanis, and A. Coskun, in *ASP-DAC*, 2014.

[11] G. Ghatikar, V. Ganti, N. E. Matson, and M. A. Piette, "Demand response opportunities and enabling technologies for data centers: Findings from f eld studies," 2012.

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

[13] Amazon, https://aws.amazon.com/about-aws/globalinfrastructure/.

[14] NRDC, "Scaling up energy eff ciency across the data center industry: Evaluating key drivers and barriers," *Issue Paper*, Aug. 2014.

[15] EnerNOC, "Ensuring U.S. grid security and reliability: U.S. EPA's proposed emergency backup generator rule," 2013, http://www.whitehouse.gov/sites/default/f les/omb/assets/oira_2060/2060_12102012-2.pdf.

[16] L. Duan, T. Kubo, K. Sugiyama, J. Huang, T. Hasegawa, and J. Walrand, "Incentive mechanisms for smartphone collaboration in data acquisition and distributed computing," in *INFOCOM, 2012 Proceedings IEEE*.

[17] P. Bolton and M. Dewatripont, *Contract theory*. MIT press, 2005.

[18] L. Duan, L. Gao, and J. Huang, "Contract-based cooperative spectrum sharing," in *DySPAN*, 2011.

[19] K. Ahmed, M. A. Islam, and S. Ren, "Greening multi-tenant data center demand response," in *Tech. Report*, 2015, http://www.ece.ucr.edu/~sren/doc/paper/iccad_2015_full.pdf.

[20] PJM RTO, "http://www.pjm.com/."

[21] M. Ghamkhari and H. Mohsenian-Rad, "Data centers to offer ancillary services," in *SmartGridComm*, 2012.

[22] D. Aikema, R. Simmonds, and H. Zareipour, "Data centres in the ancillary services market," in *IGCC*, 2012.

[23] C. Wang, B. Urgaonkar, Q. Wang, and G. Kesidis, "A hierarchical demand response framework for data center power cost optimization under real-world electricity pricing," in *MASCOTS*, 2014.

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

[25] L. Zhang, S. Ren, C. Wu, and Z. Li, "A truthful incentive mechanism for emergency demand response in colocation data centers," in *INFOCOM*, 2015.

[26] N. Chen, X. Ren, S. Ren, and A. Wierman, "Greening multi-tenant data center demand response," in *IFIP Performance*, 2015.