A First Look at Colocation Demand Response*

Shaolei Ren
Florida International University

Mohammad A. Islam
Florida International University

ABSTRACT

Large data centers can participate in demand response programs and receive financial benefits by reducing energy consumption upon utility's request. However, the existing research has only considered demand response by owner-operated data centers (e.g., Google), leaving out another distinctly different yet integral part of the data center industry — multi-tenant colocation data centers (a.k.a., colocation or "colo"), where the space is shared by multiple tenants for housing self-owned servers. A major hurdle hindering colocation demand response is "split incentive": the colocation operator may desire demand response, but lacks control over the tenants' servers; the tenants, on the other hand, can reduce server energy consumption but may not desire demand response unless they are properly incentivized. In this paper, we present a first-of-its-kind study on colocation demand response and propose an incentive mechanism, called iCODE (incentivizing COlocation tenants for DEmand response), which breaks the split-incentive barrier for colocation demand response. iCODE allows the tenants to *voluntarily* bid for energy reduction when demand response is needed and receive monetary rewards if their bids are accepted. We formally model tenants' bids and how the colocation operator decides the winning bids to maximize total energy reduction without profit loss. We demonstrate the potential of colocation demand response by using a trace-based simulation to show that iCODE can significantly reduce energy consumption (e.g., up to over 50%) during demand response periods.

1. Introduction

Demand response programs, in which customers reduce energy usage upon requests by utilities, have been successfully adopted in electricity markets to aid reliable grid operation and peak demand shaving. The formidable yet flexible electricity demand makes large data centers ideal for participating in demand response programs. Nonetheless, the research to date has been dominantly focused on owner-operated data center (e.g., Google) where the operator has complete control over the servers [2, 4]. Another important type of data center — multi-tenant *colocation* data center (a.k.a. colocation or "colo") — has been overlooked by the literature. Fundamentally differing from owner-operated data centers, a colocation houses multiple tenants that manage their own servers in shared space, while the colocation operator is mainly responsible for facility management (e.g., power, cooling).

Why colocation demand response? Our study on colocation demand response is motivated by the following two facts. **First**, colocation is a critical and integral part of the global data center industry and caters for the increasing IT demands across all industry sectors. With an estimated annual growth rate of 11%, colocation is expected to grow into a US$43-billion industry by 2018 [5]. Geographically distributed across all over the world, colocations are not only appealing options for small and medium companies that do not want to build their own in-house data centers or completely outsource to public clouds (e.g., for privacy concerns), but also satisfy the "global-footprint" needs of content distribution providers and many top-brand IT companies (e.g., Facebook, Amazon and Microsoft) to improve their last-mile latency performances. In addition, colocations are also the physical homes for many clouds, whether they are private clouds entirely serving a single enterprise or public clouds managed by medium-scale providers (e.g., Salesforce, box) for which building self-owned mega-scale data centers is out of the question due to uncertain future needs, high upfront costs and/or long construction cycle. **Second**, colocations have a great potential for demand response. As noted by a study of Google [1], "most large data centers are built to host servers from multiple companies" (i.e., colocations). Moreover, unlike large owner-operated data centers that are typically located in rural areas with low population densities, many colocations are located in densely-populated regions and even in the heart of major cities (e.g., New York), which are the very places where we need demand response.

What makes colocation demand response challenging? The full potential of colocation demand response cannot be realized without the participation of individual tenants, since the colocation operator only has control over the non-IT energy consumption (e.g., cooling energy) that offers a limited flexibility. However, because of the prevailing pricing model in colocations that often charge the tenants based on their peak power subscription (regardless of the actual energy consumption) [7],[1] tenants have little incentive to participate in demand response and instead unnecessarily always keep their servers on, resulting in an average utilization of merely 6-12% [1]. In other words, there exists a "split incentive" hurdle for colocation demand response: the incentive provided by utilities/load serving entities (LSE) to colocation operator for demand response cannot reach tenants, which own and manage servers that take up a major fraction of colocation energy consumption.

How to enable colocation demand response? We propose a first-of-its-kind incentive mechanism based on reverse auction that

*This extended abstract summarizes the work [8].

[1]Some tenants, especially large wholesale tenants, may be charged based on energy usage, but they are typically charged based on a flat rate and hence masked out by the colocation operator to participate in incentive programs for demand response [9].

financially rewards tenants for demand response. The proposed mechanism, called iCODE (incentivizing COlocation tenants for DEmand response), is "non-intrusive" to existing industry practices, as tenants' participation is fully voluntary. Fig. 1(a) illustrates the work flow of iCODE: first, upon receiving demand response requests from LSE, the colocation operator notifies the tenants to bid for energy reduction; then, participating tenants can voluntarily decide their planned energy reduction (e.g., via turning off unused servers) as well as desired payment for compensation; finally, upon receiving the tenants' bids, the colocation operator selects winning bids to maximize the total energy reduction while ensuring that the total incentive payout to tenants does not exceed the reward the colocation operator receives from LSE, and energy reduction is exercised accordingly by tenants.

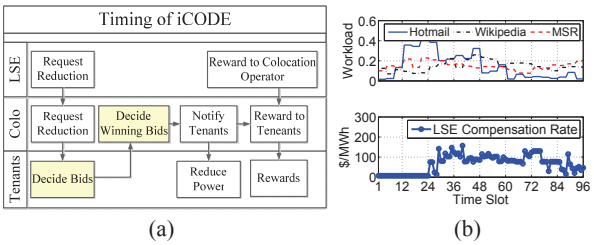

Figure 1: (a) The work flow of iCODE. (b) Data trace. "Hotmail" and "Wikipedia" are delay-sensitive, while "MSR" is delay-tolerant. Trace of compensation rate for energy reduction is taken from [6]. Each time slot is 15 minutes.

2. iCODE: Model and Algorithm

In this section, we first present the foundation of iCODE, then formalize the model, and finally formulate the problem of maximizing energy reduction for demand response.

2.1 Foundation of iCODE

We discuss both technological and economic foundations for iCODE as follows.

The tenants' servers in colocations usually have a low utilization (around 10%) [1], providing a good opportunity to reduce energy consumption via workload consolidation and turning off unused servers without any perceptible performance degradation (as extensively studied in the literature [3]). Thus, as an example, we consider the number of servers to be turned off as the energy reduction knob for the tenants, while noting that other techniques such as CPU frequency scaling can also be applied.

While dynamic pricing has been well applied in various domains (e.g., smart grid), it may not be the right choice for colocation, as it forces the tenants to follow time-varying prices which they have no control over, thereby causing business uncertainties and/or psychological concerns. Moreover, directly "reselling" energy by modifying energy price may be subject to strict regulations. Taking these into consideration, we choose a reverse auction-based incentive mechanism iCODE which, by its very nature, is "non-intrusive" to tenants, granting tenants the complete freedom of *whether* and *how* they participate in colocation demand response.

2.2 Model

We focus on one-time energy reduction for demand response which involves no sequence of *coupled* decisions, and hence we omit the time index in the model.

Tenant energy reduction. We consider a colocation data center with N tenants each having M_i homogeneous servers, while our model is extensible to heterogeneous servers. The total power consumption of the servers belonging to tenant i is $p_i = M_i \cdot \left[p_{i,s} + p_{i,d} \cdot \frac{\lambda_i}{M_i \cdot \mu_i}\right]$, where $p_{i,s}$ is the server static power, $p_{i,d}$ is the dynamic power, and μ_i is the service rate (measured in terms of the amount of workloads processed in a unit time), and λ_i is tenant i's total workload, distributed evenly among its M_i servers. Considering that tenant i turns off m_i servers when participating in demand response, the average power consumption becomes $p_i' = (M_i - m_i) \cdot \left[p_{i,s} + p_{i,d} \cdot \frac{\lambda_i}{(M_i - m_i) \cdot \mu_i}\right]$. Thus, the resulting energy reduction by tenant i is given by $\Delta e_i(m_i) = (p_i - p_i') \cdot T = m_i \cdot p_{i,s} \cdot T$, where T is the duration of each demand response period.

Tenant cost. As an *example*, we quantify the tenant's cost for turning off servers in monetary values using the notions of switching cost and delay cost. The switching cost results from, e.g., the potential wear-and-tear when powering servers up/down [3]. We denote tenant i's switching cost for one server by α_i, and thus the total switching cost for tenant i is $\alpha_i \cdot m_i$. We model the tenant's delay performance using M/M/1 queue, which provides an reasonable estimate of actual performance [3]. Hence, the average delay for tenant i's workload is $\frac{1}{\mu_i - \frac{\lambda_i}{M_i - m_i}}$. We consider a lower bound $d_{i,th}$ on the average delay due to human perception and use it as a *soft* threshold to calculate the monetary cost associated with delay performance degradation. A larger soft average delay threshold indicates that the tenant's workload is more delay-tolerant. Next, we express the total delay cost of tenant i as $d_i(m_i) = \beta_i \cdot \lambda_i \cdot [\frac{1}{\mu_i - \frac{\lambda_i}{M_i - m_i}} - d_{i,th}]^+$, where $\beta_i \geq 0$ converts additional average delay exceeding the soft threshold into monetary costs and $[\,\cdot\,]^+ = \max\{0, \cdot\}$.

Colocation operator. In addition to tenants' server energy reduction, the total energy reduction for the colocation facility also includes the energy reduction of non-IT equipments (e.g., cooling). This is captured in our model using Power Usage Effectiveness (PUE) factor γ, which typically ranges from 1.1 to 2.0. Thus, the facility level energy reduction can be expressed by $\gamma \cdot \sum_i \Delta e_i$. Denoting the compensation rate for energy reduction from LSE by q (determined via orthogonal mechanisms, e.g., [4]), the amount of reward the colocation receives is $q \cdot \gamma \cdot \sum_i \Delta e_i$.

2.3 Algorithm

We present the two main decision-making parts of iCODE as highlighted in Fig. 1(a). First, the tenants decide bids for energy reduction and corresponding financial compensation they want. Second, the colocation operator determines the winning bids to maximize the total energy reduction without losing profit.[2]

Deciding tenants' bids. We denote tenant i's requested payment from colocation operator for turning off m_i servers by $c_i(m_i) = w_i \cdot [\alpha_i \cdot m_i + d_i(m_i)]$, where $w_i \geq 1$ is referred to as *greediness* of tenant i. 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). We denote the set of tenant i's bids as $\mathbf{b}_i \subseteq \mathbf{B}_i = \{(\Delta \mathbf{e}_i, \mathbf{c}_i) \,|\, (\Delta \mathbf{e}_i(m_i), \mathbf{c}_i(m_i)), \; m_i = 0, 1, \cdots, M_{i,\max}\}$, where $M_{i,\max}$ specifies the maximum number of turned off servers (due to average delay constraint denoted by D_i) such that $\mathbf{b}_i$ only contains valid bids. Note that, the tenants decide the bids at their own discretion, and completely control the

[2]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.

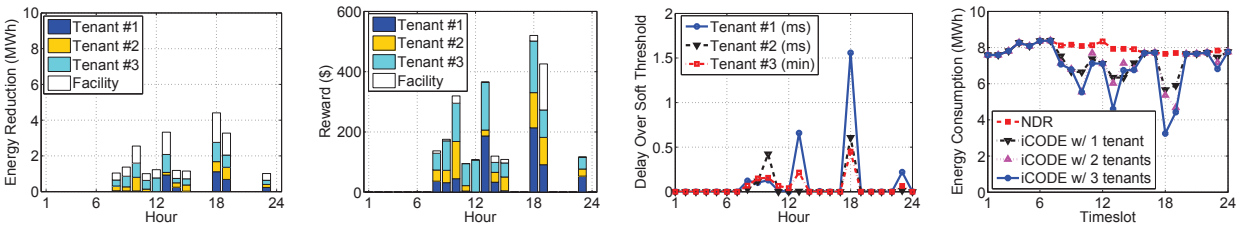

Figure 2: (a) Energy reduction by iCODE compared to NDR. (b) Incentives received by tenants. (c) Average delay exceeding the soft threshold. (d) Impact of the number of participating tenants.

Table 1: Tenant model parameters (in U.S. currency).

	Tenant #1	Tenant #2	Tenant #3
Delay cost β (¢/ms/10^6jobs)	30	20	0.4
Switching cost (¢/server/15min)	0.5	0.5	0.5
Service rate (jobs/hour)	360,000	180,000	30
Soft avg. delay threshold	12 ms	25 ms	175 s
Avg. delay constraint	20 ms	40 ms	300 s

parameters (e.g., α_i, β_i). However, asking for unreasonably high payments may result in that the tenants receive no incentives at all (without noticeably improving their delay performance), since in such cases their bids are unlikely to be accepted due to constraint of no profit loss.

Deciding winning bids. We formulate the process of deciding the winning bids into an optimization problem, in which the colocation operator aims at maximizing the energy reduction while the total incentive payout to tenants does not exceed the total reward provided by LSE. Mathematically, the problem of deciding winning bids (DWB) can be formalized as:

$$\text{DWB}: \quad \max_{(\Delta e_i, c_i), \forall i \in I} \gamma \cdot \sum_{i \in I} \Delta e_i \tag{1}$$

$$\text{s.t.} \quad \sum_{i \in I} c_i \leq q \cdot \gamma \cdot \sum_{i \in I} \Delta e_i, \tag{2}$$

$$(\Delta e_i, c_i) \in \mathbf{b}_i \cup \{(0,0)\}, \quad \forall\, i \in I, \tag{3}$$

where I is the set of participating tenants who submit their bids to the colocation operator, (1) specifies the objective of maximizing energy reduction, (2) sets the limit on total maximum incentive payment, and (3) defines the set of valid winning bids. $\{(0,0)\}$ in (3) indicates that the colocation operator may chose to reject all bids of a particular tenant (e.g., when the bids come with unreasonably high incentive payment requests). More details of solving DWB are available [8].

3. Performance Evaluation

We present a trace-based simulation study to evaluate iCODE. We consider a colocation with 3 *consolidated* tenants, each having 10,000 homogeneous servers with a static power of 150W and dynamic power of 100W per server. The simulation settings are chosen such that tenant #1 and #2 process delay-sensitive workloads, and tenant #3 processes delay-tolerant workloads. The colocation is considered to be located in New York with a PUE of 1.6 (typical for colocations). The workload traces of the 3 tenants normalized to their respective maximum server capacities and the LSE compensation rates are shown in Fig. 1(b), while the simulation parameters used are listed in Table. 1. We note that the switching cost of 0.5 cents for turning off one server for 15 minutes is already higher than the cost saving in owner-operated data centers (assuming a fair electricity price of 10 cents/KWh). This indicates that there is better motivation for turning off servers in the colocation than in owner-operated data centers (which has been extensively studied [3]). In our study, we use NDR (Non-Demand Response) as the benchmark algorithm where the tenants always keeps their servers on without incentives for demand response.

We present our results in Fig. 2, where hourly values are shown for clearer presentation. In Fig. 2(a), we show that by applying iCODE, the colocation can reduce energy consumption by as much as 4.4MWh, which is more than 50% of the NDR energy demand. Fig. 2(b) shows the financial rewards received by the tenants during different hours. This figure also includes the net-reward for the colocation operator, when there is a greater reward from the LSE than the total incentive payout to tenants. In Fig. 2(c), the delay performance of the tenants are shown, showing a negligible increase in delay (e.g., less than 2ms for tenant #1) even when there is a high energy reduction. The impact of tenants' participation in iCODE is examined in Fig. 2(d), showing a declining energy saving when fewer tenants take part in iCODE. Nonetheless, even though only 1 out of 3 tenants participate in iCODE, a non-negligible energy reduction can still be achieved.

4. References

[1] 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.

[2] M. Ghamkhari and H. Mohsenian-Rad. Data centers to offer ancillary services. In *SmartGridCom*, 2012.

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

[4] Z. Liu, I. Liu, S. Low, and A. Wierman. Pricing data center demand response. In *Sigmetrics*, 2014.

[5] MarketsAndMarkets. Colocation market - worldwide market forecast and analysis (2013 - 2018).

[6] New York ISO. `http://www.nyiso.com/`.

[7] 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.

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

[9] Silicon Valley Power. Data center program, `http://siliconvalleypower.com/index.aspx?page=2088`.