Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

00 (2015) 1–30

Greening Multi-Tenant Data Center Demand Response☆

Niangjun Chen[a], Xiaoqi Ren[a], Shaolei Ren[b], Adam Wierman[a]

[a]*Computing and Mathematical Sciences Department, California Institute of Technology*
[b]*University of California, Riverside*

Abstract

Data centers have emerged as promising resources for demand response, particularly for emergency demand response (EDR), which saves the power grid from incurring blackouts during emergency situations. However, currently, data centers typically participate in EDR by turning on backup (diesel) generators, which is both expensive and environmentally unfriendly. In this paper, we focus on "greening" demand response in multi-tenant data centers, i.e., colocation data centers, by designing a pricing mechanism through which the data center operator can efficiently extract load reductions from tenants during emergency periods for EDR. In particular, we propose a pricing mechanism for both mandatory and voluntary EDR programs, ColoEDR, that is based on parameterized supply function bidding and provides provably near-optimal efficiency guarantees, both when tenants are price-taking and when they are price-anticipating. In addition to analytic results, we extend the literature on supply function mechanism design, and evaluate ColoEDR using trace-based simulation studies. These validate the efficiency analysis and conclude that the pricing mechanism is both beneficial to the environment and to the data center operator (by decreasing the need for backup diesel generation), while also aiding tenants (by providing payments for load reductions).

Keywords: demand response, mechanism design, multi-tenant data center, supply function bidding

1. Introduction

Data centers have emerged as a promising demand response opportunity. However, data center demand response today is not environmentally friendly since data centers typically participate by turning on backup (diesel) generators. In this paper, we focus on designing a pricing mechanism for multi-tenant data centers, which is a crucial class of data centers for demand response. Our pricing mechanism allows the data center operator to obtain load shedding among tenants efficiently, reducing the need for use of backup (diesel) generation and thus greening data center demand response.

Data center demand response. Power-hungry data centers have been quickly expanding in both number and scale to support the exploding IT demand, consuming 91 billion kilowatt-hour (kWh) electricity in 2013 in the U.S. alone [1]. While traditionally viewed purely as a negative, the massive energy usage of data centers has recently begun to be recognized as an opportunity. In particular, because the energy usage of data centers tends to be flexible, they are promising candidates for *demand response*, which is a crucial tool for improving grid reliability and incorporating renewable energy into the power grid. From the grid operator's perspective, a data center's flexible power demand serves as a valuable energy buffer, helping balance grid power's supply and demand at runtime [2].

☆This work is supported in part by the U.S. NSF CNS-1319820, EPAS-1307794, CNS-1423137, and CNS-1453491.

Email addresses: ncchen@caltech.edu (Niangjun Chen), xren@caltech.edu (Xiaoqi Ren), sren@ece.ucr.edu (Shaolei Ren), adamw@caltech.edu (Adam Wierman)

To this point, data centers are a promising, but still largely under-utilized opportunity for demand response. However, this is quickly changing as data centers play an increasing role in emergency demand response (EDR) programs. EDR is the most widely-adopted demand response program in the U.S., representing 87% of demand reduction capabilities across all reliability regions [3]. Specifically, during emergency events (e.g., extreme weather or natural disasters), EDR coordinates many large energy consumers, including data centers, to shed their power loads, serving as the last protection against cascading blackouts that could potentially result in economic losses of billions of dollars [4, 5]. The U.S. EPA has identified data centers as critical resources for EDR [6], which was attested to by the following example: on July 22, 2011, hundreds of data centers participated in EDR by cutting their electricity usage before a large-scale blackout would have occurred [5].

While data centers are increasingly contributing to EDR, they typically participate by turning on their on-site backup diesel generators, which is neither cost effective nor environmentally friendly. For example, in California (a major data center market), a standby diesel generator often produces 50-60 times more nitrogen oxides (a smog-forming pollutant) compared to a typical power plant for each kWh of electricity, and diesel particulate represents the state's most significant toxic air pollution problem [7].

In addition, relying on diesel generation for EDR presents emerging challenges which, if left unaddressed, may forfeit data center's EDR capability. First, as EDR becomes more frequent [4, 8], the current financial compensation offered by power grid to data centers (for committed energy reduction during EDR) may not be enough to cover the growing cost of diesel generation. Second, data center operators are aggressively cutting the huge capital investment in their power infrastructure (e.g., 10-25$/watt [9, 10]), by down-sizing the capacity of diesel generator and uninterrupted power supply (UPS) systems [11]. Such under-provisioning of diesel generation may compromise EDR capability. Therefore, to retain and encourage data center participation in EDR without contaminating the environment, it is critical and urgent that data centers seek alternative ways to shed load.

Consequently, modulating server energy for green EDR (as well as other demand response programs such as regulation service [12]) has received an increasing amount of attention in recent years, e.g., [13, 14, 15, 16, 17, 12, 2]. These studies leverage various widely-available IT computing knobs (e.g., server turning on/off and workload migration) in data centers and provide algorithms to optimize them for participation in demand response markets. Importantly, these are not simply theoretical studies. For example, a field study by Lawrence Berkeley National Laboratory (LNBL) has illustrated that data centers can reduce energy consumption by 10-25% in response to demand response signals, without noticeably impacting normal operation [18].

Demand response in collocation data centers. While existing studies on data center demand response show promising progress, they are primarily focused on owner-operated data centers (e.g., Google) whose operators have full control over both servers and facilities. Unfortunately, such companies may actually be the least likely to participate in demand response programs, because many of their workloads are extremely delay sensitive and their data centers have been optimized for minimum delay.

In this paper, we focus on another type of data centers — multi-tenant colocation data centers (e.g., Equinix). These have been investigated much less frequently, but are actually better targets for demand response than owner-operated data centers. In a colocation data center (simply called "colocation" or "colo"), multiple tenants deploy and keep full control of their own physical servers in a shared space, while the colo operator only provides facility support (e.g., high-availability power and cooling). Colos are less studied than owner-operated data centers, but they are actually more common in practice. Colos offer data center solutions to many industry sectors, and serve as physical home to many private clouds, medium-scale public clouds (e.g., VMware) [19], and content delivery providers (e.g., Akamai). Further, a recent study shows that colos consume nearly 40% of data center energy in the U.S., while Google-type data centers collectively account for less than 8%, with the remaining going to enterprise in-house data centers [1].

In addition to consuming a significant amount of energy (more than Google-type data centers), colos are often located in places more useful for demand response. While many mega-scale owner-operated data centers are built in rural areas, colos are mostly located in metropolitan areas (e.g., Los Angeles, New York) [20], which are the very places where EDR is most needed. For all these reasons, colos are key participants in EDR programs.

Further, tenants' workloads in colos are highly heterogenous, and many tenants run non-mission-critical workloads (e.g., lab computing [21]) that have very high scheduling flexibilities, different delay sensitivities, peak load periods, etc., which is ideal for demand response participation. Thus, tenants' load shedding potentials, if appropriately exploited, can altogether form a green alternative to diesel generation for colo EDR. Nonetheless, tenants manage their own servers independently and may not have incentive to cooperate with the operator for EDR, thus raising the research question: how can a colo operator *efficiently* incentivize its tenants' load shedding for EDR?[1]

Contributions of this paper. In this paper, we focus on "greening" colocation demand response by extracting load reduction from tenants instead of relying on backup diesel generation. We study both *mandatory* EDR, a type of EDR program in which participants sign contracts and are obliged to reduce loads when requested [8], and *voluntary* EDR, where participants voluntarily reduce loads for financial compensation upon grid request [4]. In both cases, we propose a new pricing mechanism with which colo operators can extract load shedding from tenants. In particular, our proposed approach, called ColoEDR, can effectively provide incentives for tenants to reduce energy consumption during EDR events, complementing (and even substituting for) the high-cost and environmentally-unfriendly diesel generation.

ColoEDR works as follows. After an EDR signal arrives at the colo operator, tenants bid using a parameterized supply function, and then the colo operator announces a market clearing price which, when plugged into the bids, specifies how much energy tenants will reduce and how much they will be paid. Participation by the tenants is easy, since they are asked to bid only one parameter, which can be viewed as a proxy of how much flexibility in energy reduction they have at that moment. This participation can be automated and so can be easily incorporated into current practice [22], and mimics the way generation resources participate in electricity markets more broadly. For example, colo operators, like Verizon Terremark, already communicate with their tenants in preparation for an EDR event.

The main technical contribution of the paper is the analysis of the efficiency of the supply function mechanism proposed in ColoEDR. In particular, while there is a large literature studying supply function bidding [23, 24, 25, 26, 27], our setting here is novel and different. For mandatory EDR, the colo operator can either satisfy the EDR request using flexibility from the tenants (as in prior supply funding literature) or through its backup diesel generator. Thus, the diesel generator is an outside option that allows for *elasticity* in the amount of response extracted from the tenants. Further, the colo operator can combine and balance between its two options (i.e., tenant load shedding and backup generator) in order to minimize costs. For voluntary EDR, the amount of response extracted from the tenants is also an elastic decision by the colo operator, since there is no obligation for the colo to reduce energy. Thus, for both mandatory and voluntary EDR, the elastic amount of response from tenants creates a multi-stage game and adds a considerable complexity as compared to the standard setting without such elasticity, e.g., [23].

Despite the added complexity, our analysis precisely characterizes the equilibrium outcome, both when tenants are price-taking and when they are price-anticipating. In both cases, our results highlight that ColoEDR suffers little performance loss compared to the socially optimal outcome, both from the operator's and the tenants' perspectives. However, our analysis does highlight one possible drawback of ColoEDR. In the worst case, it is possible that ColoEDR may result in using significantly more on-site diesel generation than would the socially optimal. However, this bad event occurs only in cases where one tenant has an overwhelmingly fraction of the servers and has a unit cost (for energy reduction) just below that of on-site diesel generation. Such an exploitation of market power is unlikely to be possible in practical multi-tenant colocation data centers where multiple tenants with comparable sizes house their servers.

In addition to our theoretical analysis, we investigate a case study of (mandatory) EDR in §6 using trace-based experiments. The results further validate the design of ColoEDR, and show that it achieves the mandatory energy reduction for EDR while benefiting tenants through financial incentives and decreasing the operator's cost. Moreover, our simulation study shows that the efficiency loss in practical settings is even lower than what is suggested by the analytic bounds. This is especially true for the amount of on-site generation, which the analytic results suggest can (in the worst-case) be significantly larger than socially optimal but in realistic settings is very close to the social optimal.

2. Modelling Multi-Tenant Data Center EDR

Our focus is the design of a mechanism for a colo operator to extract tenant load reductions in response to an EDR signal. Thus, we need to begin by describing a model for a colo operator.

Recall that the colo operator is responsible for non-IT facility support (e.g., high-availability power, cooling). We capture the non-IT energy consumption using Power Usage Effectiveness (PUE) γ, which is the ratio of the total data

[1] Tenants receive UPS-protected power from the colo operator and share cooling systems. In other words, tenants' total energy consumption is not directly provided by the grid and includes non-metered cooling energy, which makes tenants ineligible for direct participation in EDR [4].

center energy consumption to the IT energy consumption. Typically, γ ranges from 1.1 to 2.0, depending on factors such as outside temperature.

When the operator receives an EDR signal from the LSE (Load Serving Entity), it has two options for satisfying the load reduction. First, without involving the tenants, the colo operator can use its on-site backup diesel generator.[2] We denote the amount of energy reduction by diesel generation by y and the cost per kWh of diesel generation (e.g., for fuels) by α.

Alternatively, the colo operator could try to extract IT energy reductions from the tenants. We consider a setting where there are N tenants, $i \in \mathcal{N} = \{1, 2, \cdots, N\}$. When shedding energy consumption, a tenant i will incur some costs and we denote the cost from shedding s_i by a function $c_i(s_i)$. These costs could be due to wear-and-tear, performance degradation, workload shifting, etc. For the purposes of our model, we do not specify which technique reduces the IT energy, only its cost. For details on how one might model such costs, see [28, 29, 30, 31]. A standard, natural assumption on the costs is the following.

Assumption 1. *For each n, the cost function $c_n(s_n)$ is continuous, with $c_n(s_n) = 0$ if $s_n \leq 0$. Over the domain $s_n \geq 0$, the cost function c_n is convex and strictly increasing.*

Intuitively, convexity follows from the conventional assumption that the unit cost increases as tenants reduce more energy (e.g., utilization becomes higher when servers are off, leading to a faster increase in response time of tenants' workloads).

3. Pricing Tenant Load Shedding in Mandatory EDR

EDR is the last line of protection against cascading power failures, and represents 87% of demand reduction capabilities across all the U.S. reliability regions [3]. In general, there are two types of EDR programs: mandatory and voluntary (also called economic) [4, 8]. We focus on mandatory EDR first, and return to voluntary EDR in Section 5.

For mandatory EDR, participants typically sign contracts with a load serving entity (LSE) in advance (e.g., 3 years ahead in Pennsylvania-New Jersey-Maryland Interconnection (PJM) [4]) and receive financial rebates for their committed energy reduction even if no EDR signals are triggered during the participation year, whereas non-compliance (i.e., failure to cut load as required during EDR) incurs a heavy penalty [4]. If an LSE anticipates that an emergency will occur, participants are notified, usually at least 10 minutes in advance, and obliged to fulfill their contracted amounts of energy reduction for the length of the event, which may span a few minutes to a few hours.

In mandatory EDR, the colo operator can reduce load in response to an EDR signal either through tenants or by turning on an on-site generator. Since the mandatory EDR target is fixed, the operator must balance between paying tenants for reduction and using on-site generation in order to minimize cost. Note that tenants' load reduction can also reduce the usage of diesel generator, mitigating environmental impacts. Nonetheless, the challenge is that the operator does not know the tenant cost functions, and so cannot determine the cost-minimizing price.

Consequently, the operator has two options to determine the price: (i) predict the tenant supply function and compute prices based on the predictions, or (ii) allow tenants to supply some information about their cost functions through bids. Clearly, there is a tradeoff here between the accuracy of predictions and the manipulation possible in the bids. Both of these approaches have been studied in the literature [32, 16, 23, 24, 33], though not in the context of colo demand response. In general, the broad conclusion is that approach (i) is appropriate when predictions are accurate and one bidder has market power (e.g., is significantly larger than other bidders). While market power is a considerable issue for the participation of owner-operated data centers in demand response programs due to their large size compared to other participants, it is not an issue within a specific colo that houses multiple tenants (typically of comparable sizes), and so we adopt approach (ii) in this paper.

Specifically, we design a mechanism, named ColoEDR, where tenants bid using parameterized supply functions and then, given the bids, the operator decides how much load to shed via tenants and how much to shed via on-site generation. In the following, we describe the mechanism and then contrast our approach with other potential alternatives.

[2] Other alternatives, e.g., battery [11], usually only last for < 5 minutes. So, diesel generation is the typical method [6].

Note that, throughout this paper, we focus on one EDR event, thus we omit the time index. In the case of multiple consecutive EDR events, ColoEDR will be executed once at the beginning of each event, as is standard in the literature [16, 34].

3.1. An overview of ColoEDR

The operation of ColoEDR is summarized below, and then discussed in detail in the text that follows.

1. The colo operator receives an EDR reduction target δ and broadcasts the supply function $S(\cdot, p)$ specified by(1) to tenants;
2. Participating tenants respond by placing their bids b_n;
3. The colo operator decides the amount of on-site generation y and market clearing price p to minimize its cost, using equations (2) to set the market clearing price p and (3) to set y in order to minimize the cost of EDR;
4. EDR is exercised. $\forall n \in \mathcal{N}$, tenant n sheds $S(b_n, p)$, and receives $pS(b_n, p)$ reward.

Given the overview above, we now discuss each step in more detail.

Step 1. Upon receiving an EDR notification of an energy reduction target δ, the colo operator broadcasts a parameterized supply function $S(b, p)$ to tenants (by, e.g., signalling to the tenants' server control interfaces, which are widely in use today [22]). The form of $S(b, p)$ is the following parameterized family[3]:

$$S(b_n, p) = \delta - \frac{b_n}{p}. \tag{1}$$

where p is an offered reward for each kWh of energy reduction and b_n is the bidding values that can be chosen by tenant n. This form is inspired by [23], where it is shown that by restricting the supply function to this parameterized family, the mechanism can guide the firms in the market to reach an equilibrium with desirable properties.[4] Note that, to be consistent with the supply function literature, we exchangeably use "price" and "reward rate" wherever applicable.

Step 2. Next, according to the supply function, each participating tenant submits its bid b_n to the colo operator. This bid specifies that, at each price p, it is willing to reduce $S(b_n, p)$ units of energy. The bid is chosen by tenants individually to maximize their own utility and can be interpreted as, e.g., the amount of IT service revenue that tenant n is willing to forgo. Note that b_n can be chosen to ensure that tenant n will not be required to reduce more energy than its capacity. To see this, note that since the operator is cost-minimizing, $p(\mathbf{b}, y) \le \alpha$ always holds, i.e., the market clearing price is lower than the unit cost of diesel generation. Hence, if K_n is the capacity of reduction for tenant n, as long as $b_n \ge \alpha(\delta - K_n)$, then

$$S(b_n, p) = \delta - \frac{b_n}{p} \le \delta - \frac{b_n}{\alpha} \le K_n.$$

An important note about the tenant bids is that the supply function is likely of a different form than the true cost function c_n, and so it is unlikely for the tenants to reveal their cost functions truthfully. This is necessary in order to provide a simple form for tenant bids. Bidding their true cost functions is too complex and intrusive. However, a consequence of this is that one must carefully analyze the emergent equilibrium to understand the efficiency of the pricing mechanism. We study both the cases of price-taking and price-anticipating equilibrium in §4.

Step 3. After tenants have submitted their bids, the colo operator decides the amount of energy y to produce via on-site generation and the clearing price p. Given y, the market clearing price has to satisfy $\Sigma_n S(p(\mathbf{b}), b_n) + y = \delta$, thus

$$p(\mathbf{b}, y) = \frac{\sum_n b_n}{(N-1)\delta + y}. \tag{2}$$

[3] The supply function allows tenants to have negative supply, i.e., tenants consume more energy intentionally, which is neither profit maximizing nor practical. We show in §4 that energy reduction of each tenant is always nonnegative in both equilibrium and social optimal outcomes.

[4] [23] studies the case where firms bid to supply an inelastic demand, which is equivalent to fixing the diesel generation $y = 0$ in our case. Allowing the operator to choose y in a cost-minimizing manner leads to significantly different results, as will be shown in §4.1 and §4.2.

To determine the amount of local generation y, the operator minimizes the cost of the two load-reduction options, i.e.,

$$y = \arg\min_{0\le y\le\delta}(\delta - y)\cdot p(\mathbf{b}, y) + \alpha y. \tag{3}$$

Step 4. Finally, EDR is exercised and tenants receive financial compensation from the colo operator via the realized price in (2), shed load $S(p, b_n)$, and on-site generation produces (3).

3.2. Discussion

To the best of our knowledge, this paper represents the first attempt to design a supply function bidding mechanism for colocation demand response. Although alternative mechanisms may be applicable, there are compelling advantages to the supply function approach. First, bidding for the tenants is simple – they only need to communicate one number, and it is already common practice for operators to communicate with tenants before EDR events [22], so the overhead is marginal. Second, the colo operator collects just enough information (i.e., how much energy reduction each tenant will contribute to EDR), while tenants' private information (i.e., how much performance penalty/cost each for energy reduction) is masked by the form of the supply function and hence not solicited. Third, ColoEDR guarantees that the colo operator will not incur a higher cost than the case where only backup generators are used. Further, ColoEDR pays a uniform price to all participating tenants and hence ensures fairness.

The most natural alternative bidding mechanism to supply function bidding is a Vickrey-Clarke-Groves (VCG)-based mechanism, as is suggested in [35]. While VCG-based mechanisms have the benefit of incentive compatibility, these mechanisms violate all the four properties discussed above. Under such approaches, tenants must submit very complex bids describing their precise cost functions, the true private cost of tenants is disclosed, payment made to tenants may be unbounded, and prices to different tenants are differentiated and thus raises unfairness issues.

Due to these shortcomings, VCG-based mechanisms are typically not adopted in complex resource allocation settings such as power markets, where supply-function based designs are common [23]. In fact, nearly all generation markets use variations of supply function bidding.

4. Efficiency Analysis of ColoEDR for Mandatory EDR

Given the ColoEDR mechanism described above, our task now is to characterize its efficiency. There are two potential causes of inefficiency in the mechanism: the cost minimizing behavior of the operator and the strategic behavior (bidding) of the tenants. In particular, since the forms of the tenant's cost functions are likely more complex than the supply function bids, tenants cannot bid their true cost function even if they wanted to. This means that evaluating the equilibrium outcome is crucial to understanding the efficiency of the mechanism.

Further, the equilibrium outcome that emerges depends highly on the behavior of the tenants – whether they are *price-taking*, i.e., they passively accept the offered market price p as given when deciding their own bids; or *price-anticipating*, i.e., they anticipate how the price p will be impacted by their own bids. We investigate both models, in §4.1 and §4.2, respectively.

In both cases, the goal of our analysis is to assess the efficiency of ColoEDR. To this end, we adopt a notion of a (socially) optimal outcome, and focus on the following social cost minimization (SCM) problem.

$$\text{SCM}: \quad \min \ \alpha y + \sum_{i\in\mathcal{N}} c_i(s_i) \tag{4a}$$

$$\text{s.t.} \quad y + \gamma \cdot \sum_{i\in\mathcal{N}} s_i = \delta \tag{4b}$$

$$s_i \ge 0, \ \forall i \in \mathcal{N}, \quad y \ge 0. \tag{4c}$$

where s_i and c_i are tenant i's energy reduction and corresponding cost, respectively.

The objective in SCM can be interpreted as the tenants' cost plus the colo operator's cost. Note that the internal payment transfer between the colo operator and tenants cancels, and does not impact the social cost. Also, note that payment from the LSE to the colo operator is not included in the social cost objective, since it is independent of how the operator obtains the amount δ of load reduction. Additionally, we do not include the option of ignoring the EDR signal and taking the penalty, since the non-compliance penalties are typically extreme [4]. Finally, the Lagrange multiplier of (4b) can be interpreted as the social optimal price p^*, i.e., given this price as reward for energy reduction, each tenant will individually reduce their energy by s_n that corresponds to the social cost minimization solution in (4).

Before moving to the analysis, in order to simplify notation, we suppress the PUE γ by, without loss of generality, setting $\gamma = 1$. To obtain results for $\gamma \neq 1$, simply take the results assuming $\gamma = 1$ and modified them in the following way: let y', δ' and α' be the diesel generation, EDR target and diesel price that appear in the results for $\gamma = 1$, replace them by $y' = y/\gamma$, $\delta' = \delta/\gamma$, and $\alpha' = \alpha\gamma$ where y, δ, α are the respective quantities when $\gamma \neq 1$.

4.1. Price-Taking Tenants

When tenants are price-taking, they maximize their net utility, which is the difference between the payment they receive and the cost of energy reduction, given the assumption that they consider their action does not impact the price. A price-taking tenant n will try to maximize the following payoff $P_n(b_n, p)$:

$$P_n(b_n, p) = pS_n(b_n, p) - c_n(S_n(b_n, p)) \tag{5a}$$

$$= p\delta - b_n - c_n\left(\delta - \tfrac{b_n}{p}\right). \tag{5b}$$

Here, the price-taking assumption implies that the variable p is considered to be as is. The price-taking assumption normally holds when the market consists of many players of similar sizes who have little power to impact the market clearing price. The other market model, when tenants are price-anticipating, is analyzed in Section 4.2.The market equilibrium for price-taking tenants is thus defined as follows:

Definition 1. *A triple* $(\mathbf{b}, p, y)$ *is a (price-taking) market equilibrium if each tenant maximizes its payoff defined in* (5)*, market is cleared by setting price p according to* (2)*, and the amount of on-site generation is decided by* (3)*, i.e.,*

$$P_n(b_n; p) \geq P_n(\bar{b}_n; p) \quad \forall \bar{b}_n \geq 0, \quad n = 1, \ldots, N. \tag{6}$$

$$p = \frac{\sum_{i \in \mathcal{N}} b_i}{(N-1)\delta + y}. \tag{7}$$

$$y = \arg\min_{0 \leq y \leq \delta} (\delta - y) \cdot p(\mathbf{b}, y) + \alpha y. \tag{8}$$

4.1.1. Market Equilibrium Characterization

The key to our analysis is the observation that the equilibrium can be characterized by an optimization problem. Once we have this optimization, we can use it to characterize the efficiency of the equilibrium outcome. This approach parallels that used in [23]; however, the optimization obtained has a different structure due to local diesel generation. Note that, though we use an optimization to characterize the equilibrium, the game is not a potential game since the objective (9a) below is not a potential function.

Our first result highlights that, given any choice for on-site generation, a unique market equilibrium exists for the tenants, and can be characterized via a simple optimization.

Proposition 1. *Under Assumption 1, when tenants are price-taking, for any on-site generation level* $0 \leq y < \delta$*, there exists a market equilibrium, i.e., a vector* $\mathbf{b}^t = (b_1^t, \ldots, b_N^t) \geq 0$ *and a scalar* $p > 0$ *that satisfies* (2)*, and the resulting allocation* $s_n = S(b_n, p)$ *is the optimal solution of the following*

$$\min_{\mathbf{s}} \quad \sum_{i \in \mathcal{N}} c_i(s_i) \tag{9a}$$

$$s.t. \quad \sum_{i \in \mathcal{N}} s_i = (\delta - y), \tag{9b}$$

$$s_i \geq 0, \ \forall i \in \mathcal{N}. \tag{9c}$$

This result is a key tool for understanding the overall market outcome. Intuitively, the operator running ColoEDR is more likely (than the social optimal) to use on-site generation, since this reduces the price paid to tenants. The following proposition quantifies this statement.

Proposition 2. *Under Assumption 1, it is optimal for price-taking tenants to use on-site generation if and only if*

$$\alpha < \frac{(\Sigma_n b_n)}{(N-1)\delta}.^{5} \tag{10}$$

However, when the operator is profit maximizing, it will turn on on-site generation if and only if

$$\alpha < \frac{N}{N-1}\frac{(\Sigma_n b_n)}{(N-1)\delta}. \tag{11}$$

This proposition is an important building block because the most interesting case to consider is when it is optimal to use some on-site generation and some tenant load shedding, i.e., $\delta > y^* > 0$. Otherwise the EDR requirement should be entirely fulfilled by tenants, and the analysis reduces to the case of an inelastic demand, as studied in [23]. Thus, subsequently, we make the following assumption, which ensures that on-site generation is valuable.

Assumption 2. *The unit cost of on-site generation is cheap enough that the optimal on-site generation is non-zero, i.e., α satisfies* (10).

Note that, when Assumption 2 holds, by first-order optimality condition of (3) we have

$$y = \sqrt{\frac{(\Sigma_{i\in\mathcal{N}} b_i)N\delta}{\alpha}} - (N-1)\delta, \tag{12}$$

and so the market clearing price for the tenants given on-site generation is

$$p = \frac{\sum_{i\in\mathcal{N}} b_i}{(N-1)\delta + y} = \sqrt{\frac{(\Sigma_{i\in\mathcal{N}} b_i)\alpha}{N\delta}}. \tag{13}$$

Using these allows us to prove a complete characterization of the market equilibrium under price-taking tenants. This theorem is the key to our analysis of market efficiency.

Theorem 3. *When Assumptions 1 and 2 hold there is a unique market equilibrium, i.e., a vector $\mathbf{b}^t = (b_1^t, \ldots, b_N^t) \geq 0$, $y^t > 0$ and a scalar $p^t > 0$ that satisfies* (6)-(8)*, and the resulting allocation $(\mathbf{s}^t, y^t)$ where $s_n^t = S(b_n^t, p^t)$ is the optimal solution of the following problem*

$$\min_{\mathbf{s},y} \quad \sum_n c_n(s_n) + \frac{\alpha}{2N\delta}(y + (N-1)\delta)^2 \tag{14a}$$

$$s.t. \quad \sum_n s_n = \delta - y, \tag{14b}$$

$$s_n \geq 0, \ \forall n, \quad y \geq 0. \tag{14c}$$

4.1.2. Bounding Efficiency Loss

We now use Theorem 3 to bound the efficiency loss due to strategic behavior in the market. Denote the socially optimal on-site generation by y^*, the optimal price that leads to the optimal allocation $s_i, \forall i \in \mathcal{N}$ by p^*, and let y^t and p^t be the allocation under the price-taking assumption.

Our first result highlights that, due to the cost-minimizing behavior of the operator, the equilibrium outcome uses more on-site generation and pays a lower price to the tenants than the social optimal.

[5] We adopt the convention that $\frac{0}{0} = 0$ and $\frac{x}{0} = +\infty$ when $x > 0$. Therefore, when $N = 1$, unless the bid is 0, the condition is always satisfied.

Proposition 4. *Suppose that Assumptions 1 and 2 hold. When tenants are price-taking, the operator running* ColoEDR *uses more on-site generation and pays a lower price for power reduction to its tenants than the social optimal. Specifically, $y^t \geq y^*$ and $\frac{N-1}{N}p^* \leq p^t \leq p^*$.*

Now, we move to more detailed comparisons. There are three components of market efficiency that we consider: social welfare, operator cost, and tenant cost.

First, let us consider the social cost.

Theorem 5. *Suppose that Assumptions 1 and 2 hold. Let $(\mathbf{s}^t, y^t)$ be the allocation when tenants are price-taking, and $(\mathbf{s}^*, y^*)$ be the optimal allocation. Then the welfare loss is bounded by: $\sum_n c_n(s_n^t) + \alpha y^t \leq \sum_n c_n(s_n^*) + \alpha y^* + \alpha\delta/2N$.*

Importantly, this theorem highlights that the market equilibrium is quite efficient, especially if the number of tenants is large (the efficiency loss decays to zero as $O(1/N)$). However, the market could maintain good overall social welfare at the expense of either the operator or the tenants. The following results show this is not true.

Let $\text{cost}_o(p, y)$ be the operator's cost, i.e.,

$$\text{cost}_o(p, y) = p(\delta - y) + \alpha y. \tag{15}$$

Then, we have the following results.

Theorem 6. *Suppose that Assumptions 1 and 2 are satisfied. The cost of colo operator with price-taking tenants is smaller than the cost in the socially optimal case. Further, we have $\text{cost}_o(p^*, y^*) - \alpha\delta/N \leq \text{cost}_o(p^t, y^t) \leq \text{cost}_o(p^*, y^*)$.*

4.2. Price-Anticipating Tenants

In contrast to the price-taking model, price-anticipating tenants realize that they can change the market price by their bids, i.e., that p is set according to (13), and adjust their bids accordingly. The price-anticipating model is suitable when the market consists of a few dominant players, who have significant power to impact the market price through their bids, i.e., the oligopoly setting. Clearly, this additional strategic behavior can lead to larger efficiency loss. However, in this section, we show that the extra loss is surprisingly small, especially when a large number of tenants participate in ColoEDR.

Given bids from the other tenants, each price-anticipating tenant n optimizes the following cost over bidding value b_n

$$Q_n(b_n, \mathbf{b}_{-n}) = p(\mathbf{b})S_n(b_n, p) - c_n(S_n(b_n, p))$$

where we use $\mathbf{b}_{-n}$ to denote the vector of bids of tenants other than n; i.e., $\mathbf{b}_{-n} = (b_1, \ldots, b_{n-1}, b_{n+1}, \ldots, b_N)$. Thus, substituting (1) and (13), we have

$$Q_n(b_n; \mathbf{b}_{-n}) = \sqrt{\frac{(\Sigma_n b_n)\alpha\delta}{N}} - b_n - c_n\left(\delta - \frac{b_n}{\sqrt{\Sigma_m b_m}}\sqrt{\frac{N\delta}{\alpha}}\right). \tag{16}$$

Note that the payoff function Q_n is similar to the payoff function P_n in the price-taking case, except that the tenants anticipate that the colo operator will set the price p according to $p = p(\mathbf{b}, y)$ from (13).

Definition 2. *A triple $(\mathbf{b}, p, y)$ is a (price-anticipating) market equilibrium if each tenant maximizes its payoff defined in (16), the market is cleared by setting the price p according to (2) and the amount of on-site generation is decided by (3), i.e.,*

$$Q_n(b_n; \mathbf{b}_n) \geq Q_n(\bar{b}_n; \mathbf{b}_n) \quad \forall \bar{b}_n \geq 0, \quad n = 1, \ldots, N \tag{17}$$

$$p = \frac{\sum_n b_n}{(N-1)\delta + y}. \tag{18}$$

$$y = \underset{0 \leq y \leq \delta}{\arg\min} (\delta - y) \cdot p(\mathbf{b}, y) + \alpha y. \tag{19}$$

Note that our analysis in this section requires one additional technical assumption about the tenant cost functions.

Assumption 3. *For all tenants, the marginal cost of energy reduction at 0 is greater than* $\frac{\alpha}{2N}$, *i.e.,* $\frac{\partial^+ c_n(0)}{\partial s_n} \geq \frac{\alpha}{2N}$, $\forall n$.

This assumption is quite mild, especially if the number of tenants N is large. Intuitively, it says that the unit cost of on-site generation is competitive with the cost of tenants reducing their server energy.

4.2.1. Market Equilibrium Characterization

Our analysis of market equilibria proceeds along parallel lines to the price-taking case. We again show that there exists a unique equilibrium and, furthermore, that the tenants and operator behave in equilibrium as if they were solving an optimization problem of the same form as the aggregate cost minimization (4), but with “modified” cost functions.

Theorem 7. *Suppose that Assumption 1-3 are satisfied, then there exists a unique equilibrium of the game defined by* $(Q_1, \ldots, Q_n)$ *satisfying* (17)-(19). *For such an equilibrium, the vector* $\mathbf{s}^a$ *defined by* $s_n^a = S(p(\mathbf{b}^a), b_n^a)$ *is the unique optimal solution to the following optimization:*

$$\min \quad \sum_n \hat{c}_n(s_n) + \frac{\alpha}{2N\delta}(y + (N-1)\delta)^2 \tag{20a}$$

$$s.t. \quad \sum_n s_n = \delta - y \tag{20b}$$

$$y \geq 0, \; s_n \geq 0, \quad n = 1, \ldots, N, \tag{20c}$$

where, for $s_n \geq 0$,

$$\hat{c}_n(s_n) = \frac{1}{2}\left(c_n(s_n) + s_n\frac{\alpha}{2N}\right) + \frac{1}{2}\int_0^{s_n} \sqrt{\left(\frac{\partial^+ c_n(z)}{\partial z} - \frac{\alpha}{2N}\right)^2 + 2\frac{\partial^+ c_n(z)}{\partial z}\frac{z\alpha}{N\delta}}dz, \tag{21}$$

and for $s_n < 0, \quad \hat{c}_n(s_n) = 0.$

Although the form of $\hat{c}_n(s_n)$ looks complicated, there is a simple linear approximation that gives useful intuition.

Lemma 8. *Suppose that Assumption 1-3 are satisfied. For all modified cost* $\hat{c}_n, n \in 1, \ldots, N$, *for any* $0 \leq s_n \leq \delta$,

$$c_n(s_n) \leq \hat{c}_n(s_n) \leq c_n(s_n) + s_n\frac{\alpha}{2N},$$

Furthermore, when the left or right derivatives of $\hat{c}(\cdot)$ *is defined, it can be bounded by*

$$\frac{\partial^- c_n(s_n)}{\partial s_n} \leq \frac{\partial^- \hat{c}(s_n)}{\partial s_n} \leq \frac{\partial^+ \hat{c}(s_n)}{\partial s_n} \leq \frac{\partial^+ c_n(s_n)}{\partial s_n} + \frac{\alpha}{2N}.$$

The form of Lemma 8 shows that the difference between the modified cost function in (21) and the true cost diminishes as N increases, and this is the key observation that underlies our subsequent results upper bounding the efficiency loss of ColoEDR.

4.2.2. Bounding Efficiency Loss

We now use Theorem 7 to bound the efficiency loss due to strategic behavior. Note that, by comparing to both the socially optimal and the price-taking outcomes, we can understand the impact of both strategic behavior by the operator and the tenants.

Our first result focuses on comparing the price-anticipating and price-taking equilibrium outcomes. It highlights that price-anticipating behavior leads to tenants receiving higher price while shedding less load.

Theorem 9. *Suppose Assumption 1-3 hold. Let (p^t, y^t) be the equilibrium price and on-site generation when tenants are price-taking, and (p^a, y^a) be those when tenants are price-anticipating, then we have, $y^t \le y^a \le y^t + \delta/2$ and $p^t \le p^a \le p^t + \alpha/2N$.*

Next, combining Theorem 9 and Proposition 4 yields the following comparison between the price-anticipating and socially optimal outcomes.

Corollary 10. *Suppose Assumption 1-3 hold. When tenants are price-anticipating, an operator running* ColoEDR *uses more on-site generation and pays lower market price than in the socially optimal case, i.e., $y^a \ge y^*$ and $\frac{N-1}{N} p^* \le p^a \le p^*$.*

Now, we move to more detailed comparisons. There are three components of market efficiency that we consider: social cost, operator cost, and tenant cost.

First, let us consider the social cost.

Theorem 11. *Suppose that Assumption 1-3 hold. Let $(\mathbf{s}^a, y^a)$ be the allocation when tenants are price-anticipating, and $(\mathbf{s}^*, y^*)$ be the optimal allocation. The welfare loss is bounded by: $\sum_n c_n(s_n^a) + \alpha y^a \le \sum_n c_n(s_n^*) + \alpha y^* + \alpha\delta/N$.*

Similarly to the price-taking case, the efficiency loss in the price-anticipating case decays to zero as $O(1/N)$, only with a larger constant. Also, as in the case of price-taking tenants, we again see that neither the tenants nor the operator suffers significant efficiency loss.

Theorem 12. *Suppose that Assumption 1-3 hold. The cost of colo operator for price-anticipating tenants is smaller than the cost in the socially optimal case. Further, we have*

$$\text{cost}_o(p^*, y^*) - \frac{\alpha\delta}{N} \le \text{cost}_o(p^a, y^a) \le \text{cost}_o(p^*, y^*),$$
$$\text{cost}_o(p^a, y^a) - \frac{\alpha\delta}{N} \le \text{cost}_o(p^t, y^t) \le \text{cost}_o(p^a, y^a)$$

Finally, let us end by considering the amount of on-site generation used in equilibrium. Here, in the worst-case, the on-site generation at equilibrium for price-anticipating tenants can be arbitrarily worse than the socially optimal, i.e., the socially optimal can use no on-site generation while the equilibrium outcome uses only on-site generation.

Theorem 13. *Suppose that Assumption 1-3 hold. For any $\varepsilon > 0$, $N \ge 1$, there exist cost functions $c_1, \ldots, c_N$, such that the on-site generation in the market equilibrium compared to the optimal is given by $y^a - y^* \ge \delta - \varepsilon$.*

This is a particularly disappointing result since a key goal of the mechanism is to obtain load shedding from the tenants. However, the proof emphasizes that this is unlikely to occur in practice. In particular, the worst-case scenario is that there exists a dominant (monopoly) tenant, which is unlikely in a multi-tenant colo, that has a cost function asymptotically linear with unit cost roughly matching the on-site generation price α. We confirm this in a case study in Section 6.

4.3. Discussion

The main results for the price-taking and price-anticipating analyses are summarized in Table 1. Note that simplified bounds are presented in the table, to ease interpretation, and the interested reader should refer to the theorems in §4.1 and §4.2 for the actual bounds. Also, note that the benchmark for social cost we consider is an ideal, but not achievable, mechanism.

Tenants	Price Ratio	Colo Saving	Welfare Loss
Price-taking	$[\frac{N-1}{N},\ 1]$	$[0,\ \alpha\delta/N]$	$[0,\ \alpha\delta/2N]$
Price-anticipating	$[\frac{N-1}{N},\ 1]$	$[0,\ \alpha\delta/N]$	$[0,\ \alpha\delta/N]$

Table 1. Performance guarantee of ColoEDR compared to the social optimal allocation.

To summarize the results in Table 1 briefly, note first that ColoEDR always benefits the operator, since the price paid to tenants to reduce energy is always less than the socially optimal price, and the total cost incurred by operator for energy reduction is also less than that of the social optimal. Secondly, ColoEDR also gives the tenants approximately the social optimal payment, since the operator's additional benefit is bounded above by $\alpha\delta/N$, and the welfare loss is bounded above by $\alpha\delta/N$. This naturally means that the loss in payment for tenants compared to the social optimal is at most $2\alpha\delta/N$, which approaches 0 as N grows. Third, regardless of tenants being price-taking or price-anticipating, ColoEDR is approximately socially cost-minimizing as the number of tenants grows.

However, while ColoEDR is good in terms of operator, tenant, and social cost, it may not use the most environmentally friendly form of load reduction: in the worst case, the upper bound on the extra on-site generation that ColoEDR uses is not decreasing with N. However, the analysis highlights that this worst-case occurs when there exists a dominant tenant with unit cost of energy reduction that is consistently just below the cost of diesel over a large range of energy reduction. As our case study in §6 shows, this is unlikely to occur in practice. So, ColoEDR can be expected to use an environmentally friendly mix in most realistic situations.

5. Pricing Tenant Load Shedding in Voluntary EDR

We now turn from mandatory EDR to voluntary EDR and show how the analysis and design of ColoEDR can be extended. Under voluntary EDR, a colo operator is offered a certain compensation rate for load reduction and can cut any amount of energy *at will* without any obligation. Voluntary EDR often supplements mandatory EDR, and both are widely adopted in practice [4, 8]. Since the colo operator can freely decide on the amount of energy to cut based on the compensation rate [4], the amount of energy reduction from tenants is *fully* elastic, differing from mandatory EDR where the total energy reduction (including diesel generation if necessary) needs to satisfy a constraint δ.

In the following, we formulate the problem and generalize ColoEDR for the voluntary EDR setting. Furthermore, we illustrate that the efficiency analysis, though more complicated, parallels that of mandatory EDR.

5.1. Problem Formulation

During a voluntary EDR event, the LSE offers a reward of u for each unit of energy reduction (or diesel generation if applicable). In our setting, the colo operator aims at maximizing its profit through extracting loads from tenants using parameterized supply function bidding, as considered for mandatory EDR.

A key difference with the case of mandatory EDR is that, since the reduction is flexible, diesel generation need not be considered. In particular, if the reward offered the LSE for reduction is larger than the cost of diesel, then the operator can contribute its whole diesel capacity and, if the reward is smaller than the cost of diesel, no diesel need be used. In the mandatory EDR setting, operator needs to use diesel generation when tenants' energy reduction (i.e., tenants' bids are high) is not enough, in order to meet the reduction requirement δ; in the voluntary EDR case, there is no mandatory energy reduction target and thus, the optimization of diesel generation by the operator is separable from the optimization of tenant reduction.

This yields a situation where the net profit (from tenant reduction) received by the colo operator is:

$$u \cdot d - p \cdot d \tag{22}$$

where p is the unit price the colo operator pays to the tenants to solicit d units of reduction in aggregate from N tenants, where tenant i has reduction capacity D_i.

An overview of ColoEDR. It is straightforward to adapt ColoEDR to this setting. We outline its operation in four steps below, which parallel the steps in the case of mandatory EDR.

1. The colo operator receives the voluntary EDR reduction price u and broadcasts the supply function $S(b_n, p)$ to tenants according to

$$S_i(b_i, p) = D_i - \frac{b_i}{p}, \tag{23}$$

   where D_i is the capacity of tenant i for reduction determined exogenously.

2. Participating tenants respond by placing their bids b_n in order to maximize their own payoff;

3. The colo operator decides the total amount of reduction from tenants d and market clearing price p to maximize its utility. Given the bids $\mathbf{b} = (b_1, \ldots, b_n)$, if the operator decides to offer d amount of energy reduction to the utility, then the market clearing price p needs to satisfy $\sum_{i=1}^n S_i(b_i, p) = d$ and hence, p will be

$$p = \frac{\sum_{i=1}^n b_i}{\sum_{i=1}^n D_i - d}. \tag{24}$$

Hence, to maximize the operator's profit, the operator will chooose d such that

$$d = \underset{0 \le d \le \sum_{i=1}^n D_i}{\arg\max}\ (u - p)d = \left(u - \frac{\sum_{i=1}^n b_i}{\sum_{i=1}^n D_i - d}\right) d. \tag{25}$$

4. Voluntary EDR is exercised. $\forall n \in \mathcal{N}$, tenant n sheds $S(b_n, p)$, and receives $pS(b_n, p)$ reward.

Discussion. A key difference in the operation of ColoEDR for mandatory EDR and voluntary EDR is in the form of the supply function (23). In particular, for voluntary EDR, we allow heterogeneity in the supply function for tenants in terms of their capacity D_n. Recall, however, that in the case of mandatory EDR the required reduction capacity δ was used. This difference stems from the fact that the reduction target is flexible for voluntary EDR and also creates significant challenges – both in terms of efficiency, since it allows the chance of market power to emerge because of capacity differences, and for analysis, since it adds considerable complexity.

5.2. *Efficiency Analysis of* ColoEDR *for Voluntary EDR*

Given the adaptation of ColoEDR to the voluntary EDR setting, it is natural to ask how the efficiency of the mechanism changes when the operator has full flexibility in deciding the amount of response to a voluntary EDR signal. Intuitively, the increased flexibility leads to the possibility of more inefficiency, but how large is this effect?

We again quantify efficiency through a comparison with the (socially) optimal outcome. Assume that each tenant has a cost $c_i(\cdot)$ associated with energy reduction that is convex, increasing, and $c_i(x) = 0, \forall x \le 0$ (Assumption 1). Then, the allocation that maximizes social utility (the sum of operator's and tenants' utility) solves the following problem

$$\max_{d,\mathbf{s}} \quad ud - \sum_{i=1}^n c_i(s_i) \tag{26a}$$

$$\text{subject to} \quad \sum_{i=1}^n s_i = d \tag{26b}$$

$$0 \le s_i \le D_i. \tag{26c}$$

Finally, note that our analysis makes the following natural assumptions on the unit price u and the marginal cost of each tenant. Note that these are analogous to Assumption 2 and Assumption 3, respectively.

Assumption 4. *The market clearing price p is lower than the price offered by the utility for any $d > 0$, i.e., $u \ge \frac{\sum_{i=1}^n b_i}{\sum_{i=1}^n D_i}$.*

Assumption 5. *The marginal cost of each tenants satisfies $\left.\frac{\partial^+ c_n(z)}{\partial z}\right|_{z=0} \ge \frac{v_n u}{2}, \forall n$.*

Before moving to the main results, let us first define some notations. Letting $v_n = \frac{D_n}{\sum_{i=1}^n D_i}$, we have $\sum_n v_n = 1$. Here v_n behaves like "market share" of tenant n in the voluntary EDR market. In the mandatory EDR case, $v_n = 1/N$ for all n. Furthermore, define $v = \max_n v_n$, as the "dominant share" in load reduction among the tenants, and $D = \max_n D_n$.

5.3. *Market Equilibrium Characterization*

As in the case of mandatory EDR, we consider both price-taking and price-anticipating tenants.

5.3.1. Price-taking Tenants

Recall that a price-taking tenant considers the price as is without accounting for the impact of its bidding decision on the market clearing price. Hence, given the other tenants' bidding decisions, each price-taking tenant n optimizes the following payoff over bidding value b_n,

$$P_n(b_n, \mathbf{b}_{-n}) = pS_n(b_n, p) - c_n(S_n(b_n, p)) = pD_n - b_n - c_n(D_n - \frac{b_n}{p})$$

So, in a price-taking equilibrium $(\mathbf{b}, d, p)$, $P_n(b_n; \mathbf{b}_{-n}) \geq P_n(\bar{b}_n; \mathbf{b}_{-n})$ holds for each tenant n over all $\bar{b}_n \geq 0$. Also, the market clearing price p must satisfy (24) and the total reduction d must satisfy (25). Using techniques similar to the proof of Theorem 3, we can completely characterize the price-taking equilibrium of ColoEDR in voluntary EDR as follows:

Theorem 14. *There exists a unique equilibrium of the game defined by* $(P_1, \ldots, P_N)$ *for* ColoEDR *in voluntary EDR. For such an equilibrium, the vector* $\mathbf{s}^t$ *defined by* $s_n^t = S(p(\mathbf{b}^t), b_n^t)$ *is the unique optimal solution to the following optimization:*

$$\max \quad ud - \frac{ud^2}{2\sum_n D_n} - \sum_n c_n(s_n) \tag{27a}$$

$$s.t. \quad \sum_n s_n = d \tag{27b}$$

$$d \geq 0,\ 0 \leq s_n \leq D_n, \quad n = 1, \ldots, N, \tag{27c}$$

5.3.2. Price-anticipating Tenants

Recall that a price-anticipating tenant actively seek to change market price through its bid to maximize payoff. Hence, given the other tenants' bidding decisions, each price-anticipating tenant n optimizes the following payoff over bidding value b_n, the payoff function $Q_n(b_n, \mathbf{b}_{-n})$ can be derive in a similar manner as (16):

$$Q_n(b_n, \mathbf{b}_{-n}) = p(\mathbf{b})S_n(b_n, p) - c_n(S_n(b_n, p)) = v_n \sqrt{\Sigma_m b_m} \sqrt{u \sum_{i=1}^{n} D_i} - b_n - c_n(D_n - \frac{b_n}{\Sigma_m b_m} \sqrt{\frac{\sum_{i=1}^{n} D_i}{u}}),$$

So, in a price-anticipating equilibrium $(\mathbf{b}, d, p)$, we must have $Q_n(b_n; \mathbf{b}_{-n}) \geq Q_n(\bar{b}_n; \mathbf{b}_{-n})$ for all n over all $\bar{b}_n$. Also, the market clearing price p must satisfy (24) and the total reduction d must satisfy (25).

Using techniques similar to the proof of Theorem 7, we can completely characterize the price-anticipating equilibrium of ColoEDR in voluntary EDR as follows.

Theorem 15. *There exists a unique equilibrium of the game defined by* $(Q_1, \ldots, Q_N)$ *for* ColoEDR *in voluntary EDR. For such an equilibrium, the vector* $\mathbf{s}^a$ *defined by* $s_n^a = S(p(\mathbf{b}^a), b_n^a)$ *is the unique optimal solution to the following optimization:*

$$\max \quad ud - \frac{ud^2}{2\sum_n D_n} - \sum_n \hat{c}_n(s_n) \tag{28a}$$

$$s.t. \quad \sum_n s_n = d \tag{28b}$$

$$d \geq 0,\ 0 \leq s_n \leq D_n, \quad n = 1, \ldots, N, \tag{28c}$$

Tenants	Price Ratio	Colo Extra Profit	Welfare Loss
Price-taking	$[1-\frac{d^*}{\Sigma_n D_n},\ 1]$	$[0,\ ud^{*2}/\Sigma_n D_n]$	$[0,\ ud^{*2}/2\Sigma_n D_n]$
Price-anticipating	$[1-\frac{d^*}{\Sigma_n D_n},\ 1]$	$[0,\ ud^{*2}/\Sigma_n D_n]$	$[0,\ u(\Sigma_n D_n \nu_n + d^{*2}/\Sigma_n D_n)/2]$

Table 2. Performance guarantee of ColoEDR compared to the social optimal allocation.

where, for $s_n \geq 0$,

$$\hat{c}_n(s_n) = \frac{1}{2}\left(s_n \frac{\nu_n u}{2} + c_n(s_n)\right) + \frac{1}{2}\int_0^{s_n} \sqrt{\left(\frac{\nu_n u}{2} - \frac{\partial^+ c_n(z)}{\partial z}\right)^2 + 2\frac{\partial^+ c_n(z)}{\partial z}\frac{zu}{\Sigma_i D_i}}dz, \tag{29}$$

and for $s_n < 0$, $\quad \hat{c}_n(s_n) = 0$.

Like in the case of mandatory EDR, the above characterization can be approximated using a modified cost function when ν_n is small, i.e., when there are a large number of tenants and all tenants have similar market shares.

Lemma 16. *For $0 \leq s_n \leq D_n$, the modified cost in* (29) *can be upper and lower bounded by,*

$$c_n(s_n) \leq \hat{c}_n(s_n) \leq c_n(s_n) + s_n \frac{\nu_n u}{2},$$

Furthermore, where the left or right derivatives are defined, we have

$$\frac{\partial^- c_n(s_n)}{\partial s_n} \leq \frac{\partial^- \hat{c}_n(s_n)}{\partial s_n} \leq \frac{\partial^+ \hat{c}_n(s_n)}{\partial s_n} \leq \frac{\partial^+ c_n(s_n)}{\partial s_n} + \frac{\nu_n u}{2}. \tag{30a}$$

5.4. Bounding Efficiency Loss

We now use the characterization results in Theorem 14 and Theorem 15 to analyze the social efficiency of ColoEDR in the voluntary EDR setting for both price-taking and price-anticipating tenants.

Theorem 17. *For price taking tenants, the welfare loss of* ColoEDR *in voluntary EDR is bounded by $ud^t - \sum_n c_n(s_n^t) \geq ud^* - \sum_n c_n(s_n^*) - \frac{ud^{*2}}{2\sum_n D_n}$. Moreover, the bound is tight.*

Theorem 18. *For price anticipating tenants, the welfare loss of* ColoEDR *in voluntary EDR is bounded by $ud^a - \sum_n c_n(s_n^a) \geq ud^* - \sum_n c_n(s_n^*) - \frac{u}{2}\left(\Sigma_n D_n \nu_n + \frac{d^{*2}}{\Sigma_n D_n}\right)$.*

Theorem 17 highlights that the price-taking market equilibrium is efficient when the optimal energy reduction d^* is small. This is due to the profit maximizing behavior of the operator: when the social optimal d^* is large, the operator has greater opportunity to raise his profit by lowering the market price.

Comparing Theorem 18 with Theorem 17, we can see that the additional welfare loss due to price-anticipating behavior of tenants is a function of ν_n, the market share of the tenants. It is easy to see that the additional loss of social welfare is minimized when $\nu_n = 1/N$ for all n, i.e., when the reduction capacity of each tenant is equal.

Additionally, we can obtain tight bounds on the market clearing price, energy reduction quantity, and operator's profit in a similar fashion as our analysis done for the mandatory EDR case using Theorem 14 and Theorem 15. Due to space constraints, we summarize the results in Table 2 and Table 3.

Table 2 shows that as the optimal reduction d^* increases, there is more opportunity for the operator to profitably reduce market price and increase his own profit. Table 3 shows further that, when tenants are price-anticipating, they will drive the market clearing price up, provide less energy reduction and reduce the operator's profit. However, all these additional losses can be bounded by linear functions of ν, the dominant share of the energy reduction capacity. Hence, the loss due to price-anticipating behavior of tenants is minimized $D_1 = D_2 = \cdots = D_N$.

Price Markup	Load Reduction	Operator's cost
[0, $uv/2$]	[$-D/2$, 0]	[0, uD]

Table 3. Performance guarantee of ColoEDR when tenants are price-anticipating compared to them being price-taking.

6. Case Study

Our goal in this section is to investigate ColoEDR in a realistic scenario. Given the theoretical results in the prior sections, we know that ColoEDR is efficient for both the operator and tenants when the number of tenants is large, but that it may use excessive on-site generation (in the worst case). Thus, two important issues to address in the case study are: *How efficient is the pricing mechanism in small markets, i.e., when N is small? What is the impact of the pricing mechanism on on-site generation in realistic scenarios?* Additionally, the case study allows us to better understand when it is feasible to obtain load shedding from tenants, i.e., *how flexible must tenants be in order to actively participate in a load shedding program?*

Due to space constraints, we discuss only mandatory EDR in this section. The results in the case of voluntary EDR are parallel and hence omitted for brevity.

6.1. Simulation Settings

We use trace-based simulations in our case study. Our simulator takes the tenants' workload trace and a trace of mandatory EDR signals from Pennsylvania-New Jersey-Maryland Interconnection (PJM) [4] as its inputs. It then executes ColoEDR (by emulating the bidding process and tenants' energy reduction for EDR) at each timestamp of the EDR signal, and outputs the resulting equilibrium. The settings we use for modeling the colocation data center and the tenant costs follow.

Colocation data center setup. We consider a colocation data center located in Ashburn, VA, which is a major data center market served by PJM Interconnection. By default, there are three participating tenants interested in EDR, though we vary the number of participating tenants during the experiments.

Each participating tenant has 2,000 servers, and each server has an idle and peak power of 150W and 250W, respectively. The default PUE of the colo is set to 1.5 (typical for colo), and hence, whenever a tenant reduces 1kWh energy, the corresponding energy reduction at the colo level amounts to 1.5kWh. Thus, the maximum possible power reduction is 2.25MW (i.e., 1.5MW IT plus 0.75MW non-IT). We assume that the colo operator counts the extra energy reduction at the colo level as part of the tenants' contributions, and rewards the tenants accordingly.

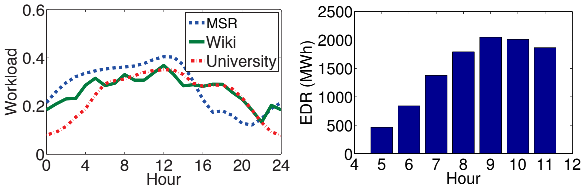
The colo has an on-site diesel generator, which has cost \$0.3/kWh estimated based on typical fuel efficiency [36].

For setting the energy reduction target received by the colo, we follow the EDR signals issued by PJM Interconnection from 5:00am to 11:00am on January 7, 2014, when many states in the eastern U.S. experienced an extremely cold weather and faced an electricity production shortage [37]. Fig. 1(b) shows the total energy reduction target set by PJM during that day for all participating colos. In our simulation, we keep the shape of the energy reduction target but scale down the reduction amount based on real power consumption in our considered colo.

Tenant workloads characteristics. We choose three representative types of workloads for participating tenants: tenant 1 is running high delay-sensitive workloads (e.g., user-facing web service), tenant 2 is running low delay-sensitive workloads (e.g., enterprise's internal services), and tenant 3 is running delay-tolerant workload (e.g., back-end processing).

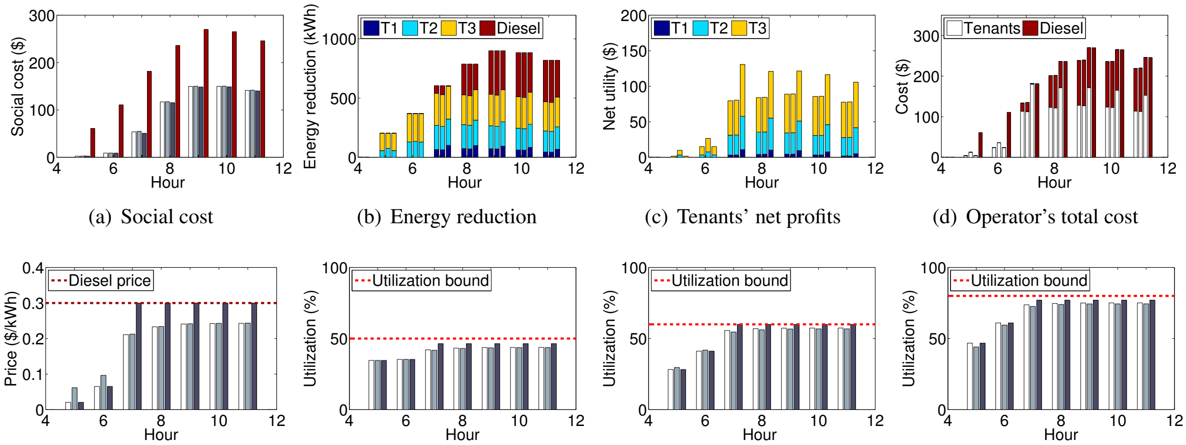
The workload traces for the three participating tenants were collected from server utilization log of MSR [38], Wiki [39], and Florida International University, respectively. Fig. 1(a) illustrates a snapshot of the traces of server cluster utilization over 24 hours, where the workloads are normalized with respect to each tenant's maximum service capacity. For our evaluation based on PJM EDR signals, we only use the traces from Hour 5–11 (i.e., 5:00am–11:00am). The illustrated results use an average utilization for each tenant of 30%, consistent with reported values from real systems [9]. Our results are not particularly sensitive to this choice.

There are various power management techniques, e.g., load migration/scheduling, that can be used for reducing tenants' server energy. Here, as a concrete example, we consider that tenants dynamically consolidate workloads and turn on/off servers for energy saving subject to SLA [40]. This power-saving technique has been widely studied [40, 41] and also recently applied in real systems (e.g., Facebook's AutoScale [42]).



(a) (b)

Figure 1. **(a)** Workload traces. **(b)** Energy reduction for PJM's EDR on January 7, 2014 [37].



(a) Social cost (b) Energy reduction (c) Tenants' net profits (d) Operator's total cost

(e) Market clearing price (f) Tenant 1's utilization (g) Tenant 2's utilization (h) Tenant 3's utilization

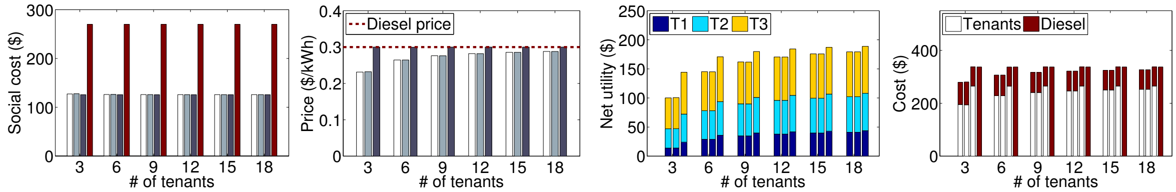
Figure 2. Performance comparison under default settings. Throughout this and later plots, the bars in each cluster are the price-taking, price-anticipating, socially optimal, and diesel only (if applicable) outcomes.

When tenants save energy for EDR by turning off some servers, their application performance might be affected. We adopt a simple yet common model based on an M/G/1/Processor-Sharing queueing model, as follows. For a tenant with M servers each with a service rate of μ, denote the workload arrival rate by λ. When m servers are shut down, we model the total delay cost as $\bar{c}(m) = \lambda \cdot \beta \cdot T \cdot \text{delay}(m) = \frac{\beta T}{\frac{1}{\nu M} - \frac{1}{M-m}}$, where $\nu = \frac{\lambda}{\mu M}$ denotes the normalized workload arrival (i.e., utilization without turning off servers), T is the duration of an EDR event, and β is a cost parameter ($/time unit/job). In our simulations, we set the cost parameter for tenant 1, tenant 2 and tenant 3 as 0.1, 0.03, 0.006, respectively, which are already higher than those considered in the prior context of turning off servers for energy saving [40]. Note that we have experimented with a variety of other models as well and the results do not qualitatively change.

We use a standard model for energy usage [9] and take the energy reduction s as linear in the number of servers shut down, i.e., $s = \theta \cdot m$, where θ is a constant decided by server's idle power and T. Then, it yields the following cost function for a tenant's energy reduction $c(s) = \bar{c}(\frac{s}{\theta}) - \bar{c}(0)$, where $\bar{c}(\,\cdot\,)$ is defined in the above paragraph. Note that we have experimented with a variety of other forms, and our results are not sensitive to the details of this cost function.

Finally, note that tenants typically have a delay performance requirement. Based on the above queueing model, this can be translated as an utilization upper bound. Such a translation is also common in real systems (e.g., default policy for auto-scaling virtual machines [43]). In our simulation, we capture the performance constraint by setting utilization upper bounds for tenant 1, tenant 2, and tenant 3 as 0.5, 0.6, and 0.8, respectively.

Efficiency benchmarks. Throughout our experiments, we consider the price-taking, price-anticipating, and social optimal outcomes. Additionally, we consider one other benchmark, *diesel only*, which is meant to capture common practice today and to highlight that any tenant response extracted "greens" data center demand response. Under diesel only, the full EDR response is provided by the on-site diesel generator. *Throughout, our results are presented in grouped bar plots with the bars representing (from left to right) the price-taking, price-anticipating, social optimal,*



(a) Social cost (b) Market price (c) Tenants' net profits (d) Operator's total cost

Figure 3. Impact of number of tenants.

and diesel only (if applicable) outcomes.

While other mechanisms (e.g., direct pricing [16], auction [35]) have been introduced in recent papers, we do not compare ColoEDR with them here because ColoEDR is already typically indistinguishable from the social optimal cost.

6.2. *Performance Evaluation*

We now discuss our main results, shown in Fig. 2.

Social cost. We first compare in Fig. 2(a) the social costs (4) incurred by different algorithms. Note that ColoEDR is close to the social cost optimal under both price-taking and price-anticipating cases even though there are only three participating tenants. Further, the resulting social costs in both the price-taking and price-anticipating scenarios are significantly lower than that of the diesel only outcome. This shows a great potential of tenants' IT power reduction for EDR, which is consistent with the prior literature on owner-operated data center demand response [15, 16, 2].

Energy reduction contributions. Fig. 2(b) plots EDR energy reduction contributions from tenants and the diesel generator. As expected from analytic results, both price-taking and price-anticipating tenants tend to contribute less to EDR (compared to the social optimal) because of their self-interested decisions. In other words, given self-interested tenants, the colo operator needs more diesel generation than the social optimal. Nonetheless, the difference is fairly small, much smaller than predicted by the worst-case analytic results. This highlights that worst-case results were too pessimistic in this case. Of course, one must remember that all tenant reduction extracted is in-place of diesel generation, and so serves to make the demand response more environmentally friendly.

Benefits for tenants and colocation operator. We show in Fig. 2(c) and Fig. 2(d) that both the tenants and the colo operator can benefit from ColoEDR. Specifically, Fig. 2(c) presents net profit (i.e., payment made by colo operator minus performance cost) received by tenants, showing that all participating tenants receive positive net rewards. While price-anticipating tenants can receive higher net rewards than when they are price-taking, the extra reward gained is quite small. Similarly, Fig. 2(d) shows cost saving for the colo operator, compared to the "diesel only" case.

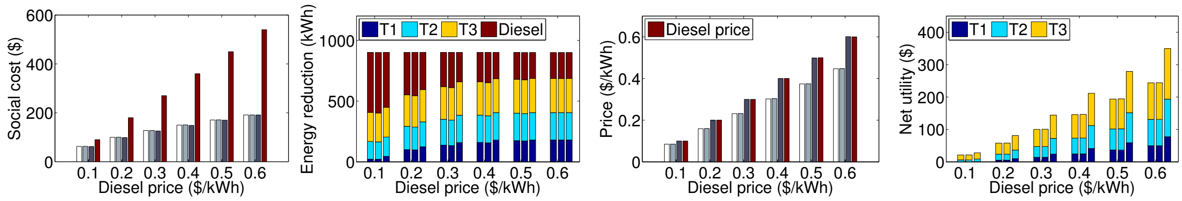
Market clearing price. Fig. 2(e) shows the market clearing price. Naturally, when using ColoEDR to incentivize tenants for EDR while minimizing the total cost, the colo operator will not pay the tenants at a price higher than its diesel price (shown via the red horizontal line). We also note that the price under the price-anticipating case is higher than that under the price-taking case, because the price-anticipating tenants are more strategic. However, the price difference between price-anticipating and price-taking cases is quite small, which again confirms our analytic results.

Tenant' server utilization. Tenants' server utilizations are shown in Figs. 2(f), 2(g) and 2(h), respectively. These illustrate that, while tenants reduce energy for EDR, their server utilizations still stay within their respective limits (shown via the red horizontal lines), satisfying performance constraints. This is because tenants typically provision their servers based on the maximum possible workloads (plus a certain margin), while in practice their workloads are usually quite low, resulting in a "slackness" that allows for saving energy while still meeting their performance requirements.

6.3. *Sensitivity Analysis*

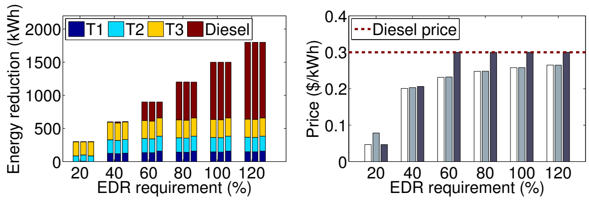
To complete our case study, we investigate the sensitivity of the conclusions discussed above to the settings used. For each study, we only show results that are most significantly different than those in Fig. 2.

Impact of the number of tenants. First, we vary the number of participating tenants and show the results in Fig. 3. To make results comparable, we fix the EDR energy reduction requirement as well as total number of servers:



(a) Social cost (b) Energy reduction (c) Market price (d) Tenants' net profits

Figure 4. Impact of diesel price.



(a) Energy reduction (b) Market price

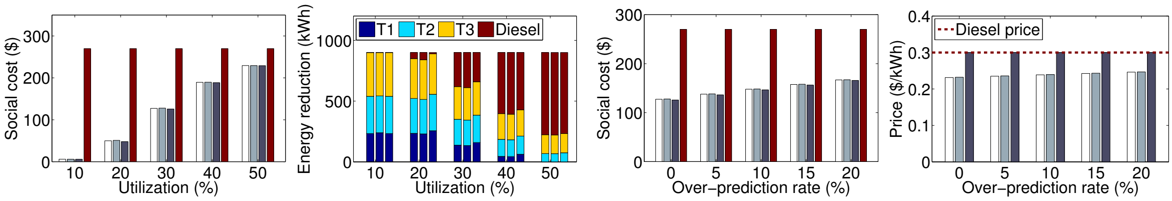
Figure 5. Impact of EDR energy reduction target.

tenant 1, tenant 2 and tenant 3 are each equally split into multiple smaller tenants, each having fewer servers with the same workload arrival rate scaled down accordingly. We then aggregate replicas of the same tenant together for an easy viewing in the figures, e.g., "tenant 1" in the figures represent the whole group of tenants that are obtained by splitting the original tenant 1. One interesting observation is that as more tenants participate in EDR, the market becomes more "competitive". Hence, each individual tenant can only gain less net reward, but both the price and the aggregate net reward become higher (see Figs. 3(b) and 3(c)). Motivated by this, one might suggest a possible trick: a tenant may gain more utility by splitting its servers and pretending to be multiple tenants. In practice, however, each tenant has only one account (for billing, etc.) which requires contracts and base fees, and thus pretending as multiple tenants is not viable in a colo.

Impact of the price of diesel. Fig. 4 illustrates how our result changes as the diesel price varies. Intuitively, as shown in Fig. 4(a), the social cost (which includes diesel cost as a key component) increases with the diesel price. We see from Figs. 4(b) and 4(c) that, when diesel price is very low (e.g., $0.1/kWh), the colo operator is willing to use more diesel and offers a lower price to tenants. As a result, tenants contribute less to EDR. As the diesel price increases (e.g., from $0.2/kWh to $0.3/kWh), the colo operator increases the market price (but still below the diesel price) to encourage tenants to cut more energy for EDR. Nonetheless, tenants' energy reduction contribution cannot increase arbitrarily due to their performance constraints. Specifically, after the diesel price exceeds $0.4/kWh, tenants will not contribute more to EDR (i.e., almost all their IT energy reduction capabilities have been exploited), even though the colo operator increases the reward. In this case, tenants simply receive higher net rewards without further contributing to EDR, as shown in Fig. 4(d).

Impact of EDR requirement. Fig. 5 varies the EDR energy reduction target, with the maximum reduction ranging from 20% to 120% of the colo's peak IT power consumption. As the EDR energy reduction target increases, tenants' energy reduction for EDR also increases; after a certain threshold, diesel generation becomes the main approach to EDR, while the increase in tenant's contribution is diminishing (even though the colo operator increases the market price), because of tenants' performance requirements that limit their energy reduction capabilities.

Impact of tenants' workloads. In Fig. 6(a)-6(b), we vary the tenants' workload intensity (measured in terms of the average server utilization when all servers are active) from 10% to 50%, while still keeping the maximum utilization bounds to 50%, 60% and 80% as the performance requirements for the three tenants, respectively. While it is straightforward that when tenants have more workloads, they tend to contribute less to EDR, because they need to keep more servers active to deliver a good performance. Nonetheless, even when their average utilization without turning off servers is as high as 50% (which is quite high in real systems, considering that the average utilization



(a) Social cost (b) Energy reduction (c) Social cost (d) Market price

Figure 6. Impact of tenants' workloads and the workload prediction errors.

is only around 10-30% [9]), tenants can still contribute more than 20% of EDR energy reduction under ColoEDR, showing again the potential of IT power management for EDR.

Impact of workload prediction error. In practice, tenants may not perfectly estimate their own workload arrival rates. To cope with possible traffic spikes, tenants can either keep more servers active as a backup or deliberately overestimate the workload arrival rate by a certain overestimation factor. We choose the later approach in our simulation. Fig. 6(c)-6(d) shows the result under workload prediction errors. We see that both the social cost and market price are fairly robust against tenants' workload over-predictions. For example, the social cost increases by less than 10%, even when tenants overestimate their workloads by 20% (which is already sufficiently high in practice, as shown in [41]). Other results (e.g., tenants' net reward, colo operator's total cost) are also only minimally affected, thereby demonstrating the robustness of ColoEDR against tenants' workload over-predictions.

7. Related Work

Our work contributes both to the growing literature on data center demand response, and to the literature studying supply function equilibria. We discuss each in turn below.

Recently, data center demand response has received a growing amount of attention. A variety of approaches have been considered, such as optimizing a grid operator's pricing strategies for data centers [16] and tuning computing (e.g., server control and scheduling) and/or non-computing knobs (e.g., cooling system) in data centers for various types of demand response programs [15, 44, 13, 14]. Field tests by LBNL also verify the practical feasibility of data center demand response using a combination of existing power management techniques (e.g., load migration) [18]. These studies, however, have all focused on large owner-operated data centers.

In contrast, to the best of our knowledge, colocation demand response has been investigated by only a few previous works. The first is [34], which proposes a simple mechanism, called iCODE, to incentivize tenants' load reduction. But, iCODE is purely for voluntary EDR and does not include any energy reduction target (needed for mandatory EDR). More importantly, iCODE is designed without considering strategic behavior by tenants, and can be compromised by price-anticipating tenants [34]. More relevant to the current work is [35], which proposes a VCG-type auction mechanism where colo participates in EDR programs. While the mechanism is approximately truthful, it asks participating tenants to reveal their private cost information through complex bidding functions. Further, the colo operator may be forced to make arbitrarily high payments to tenants. In contrast, our proposed solution provides a simple bidding space, protects tenants' private valuation, and ensures that the colo operator does not incur a higher cost for EDR than the case tenant contributions. Thus, unlike [35], ColoEDR benefits both colo operator and tenants, giving both parties incentives to cooperate for EDR. Further, ColoEDR is applicable to both mandatory and voluntary EDR, both of which are important EDR programs [8].

Finally, it is important to note that our approach builds on, and adds to, the supply function mechanism literature. Supply function bidding (c.f. the seminal work by [45]) is frequently used in electricity markets due to its simple bidding language and the avoidance of the unbounded payments typical in VCG-like mechanisms. Supply function bidding mechanisms have been extensively studied, e.g., [24, 25, 26, 27, 33, 46]. The literature primarily focuses on existence and computation of supply function equilibrium, sometimes additionally proving bounds on efficiency loss. Our work is most related to [23], which considers an inelastic demand δ that must be satisfied via extracting load shedding from consumers and proves efficient bounds on supply function equilibrium. In contrast, our work assumes that the operator has an outside option (diesel) that can be used to satisfy the inelastic demand. This leads to a multistage game between the tenants and the profit-maximizing operator, a dynamic which has not been studied previously in the supply function literature.

8. Conclusion

In this paper, we focused on "greening" colocation demand response by designing a pricing mechanism that can extract load reductions from tenants during EDR events. Our mechanism, ColoEDR, can be used in both mandatory and voluntary EDR programs and is easy to put in place given systems available in colos today. The main technical contribution of the work is the analysis of the ColoEDR mechanism, which is a supply function mechanism for an elastic demand, a setting for which efficiency results have not previously been attained in the supply function literature. Our results highlight that ColoEDR provides provably near-optimal efficiency guarantees, both when tenants are price-taking and when they are price-anticipating. We also evaluate ColoEDR using trace-based simulation studies and validate that ColoEDR is not only "greens" multi-tenant EDR by reducing diesel generation, it also benefits the colo operator by reducing costs and the tenants by providing payments for reductions.

References

[1] NRDC, Scaling up energy efficiency across the data center industry: Evaluating key drivers and barriers, Issue Paper.

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

[3] K. Managan, Demand repsonse: A market overview (2014, `http://enaxisconsulting.com`).

[4] PJM, Emergency demand response (load management) performance report – 2012/2013.

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

[6] EnerNOC, Ensuring U.S. grid security and reliability: U.S. EPA's proposed emergency backup generator rule (2013).

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

[8] PJM, Retail electricity consumer opportunities for demand response in PJM's wholesale markets (`http://www.pjm.com`).

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

[10] H. Lim, A. Kansal, J. Liu, Power budgeting for virtualized data centers, in: USENIX ATC, 2011.

[11] D. Wang, S. Govindan, A. Sivasubramaniam, A. Kansal, J. Liu, B. Khessib, Underprovisioning backup power infrastructure for datacenters, in: ASPLOS, 2014.

[12] S. Li, M. Brocanelli, W. Zhang, X. Wang, Data center power control for frequency regulation, in: PES, 2013.

[13] B. Aksanli, T. S. Rosing, Providing regulation services and managing data center peak power budgets, in: DATE, 2014.

[14] H. Chen, A. K. Coskun, M. C. Caramanis, Real-time power control of data centers for providing regulation service, in: CDC, 2013.

[15] D. Aikema, R. Simmonds, H. Zareipour, Data centres in the ancillary services market, in: IGCC, 2012.

[16] Z. Liu, I. Liu, S. Low, A. Wierman, Pricing data center demand response, in: SIGMETRICS, 2014.

[17] M. Ghamkhari, H. Mohsenian-Rad, Data centers to offer ancillary services, in: SmartGridCom, 2012.

[18] G. Ghatikar, V. Ganti, N. E. Matson, M. A. Piette, Demand response opportunities and enabling technologies for data centers: Findings from field studies (2012).

[19] Data Center Knowledge, Inside SuperNAP 8: Switch's Tier IV data fortress, Feb. 11, 2014.

[20] DatacenterMap, Colocation USA (`http://www.datacentermap.com/usa/`).

[21] J. Verge, Symantec Signs Multi-Megawatt Lease at Santa Clara Data Center (2015, `http://www.datacenterknowledge.com/`).

[22] Equinix, Customer portal (`http://www.equinix.com/services/support/customer-portal/`).

[23] R. Johari, J. N. Tsitsiklis, Parameterized supply function bidding: Equilibrium and efficiency, Operations research 59 (5) (2011) 1079–1089.

[24] C. J. Day, B. F. Hobbs, J.-S. Pang, Oligopolistic competition in power networks: a conjectured supply function approach, Power Systems, IEEE Transactions on 17 (3) (2002) 597–607.

[25] R. Baldick, R. Grant, E. Kahn, Theory and application of linear supply function equilibrium in electricity markets, Journal of Regulatory Economics 25 (2) (2004) 143–167.

[26] R. J. Green, D. M. Newbery, Competition in the british electricity spot market, J. of Political Economy (1992) 929–953.

[27] R. Green, Increasing competition in the british electricity spot market, J. of Industrial Economics (1996) 205–216.

[28] S. Ong, P. Denholm, E. Doris, The impacts of commercial electric utility rate structure elements on the economics of photovoltaic systems, National Renewable Energy Laboratory, 2010.

[29] X. Fan, W.-D. Weber, L. A. Barroso, Power provisioning for a warehouse-sized computer, in: ACM SIGARCH Computer Architecture News, Vol. 35, ACM, 2007, pp. 13–23.

[30] L. L. Andrew, M. Lin, A. Wierman, Optimality, fairness, and robustness in speed scaling designs, SIGMETRICS Perform. Eval. Rev. 38 (1) (2010) 37–48. doi:10.1145/1811099.1811044.
URL `http://doi.acm.org/10.1145/1811099.1811044`

[31] A. Wierman, L. L. Andrew, A. Tang, Power-aware speed scaling in processor sharing systems, in: INFOCOM 2009, IEEE, IEEE, 2009, pp. 2007–2015.

[32] A.-H. Mohsenian-Rad, A. Leon-Garcia, Optimal residential load control with price prediction in real-time electricity pricing environments, Smart Grid, IEEE Transactions on 1 (2) (2010) 120–133.

[33] E. J. Anderson, X. Hu, Finding supply function equilibria with asymmetric firms, Oper. Res. 56 (3) (2008) 697–711.

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

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

[36] Wikipedia (`http://en.wikipedia.org/wiki/Diesel_generator`).

[37] PJM, Demand response activity on January 7-8, 2014 (`http://www.pjm.com`).

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

[39] G. Urdaneta, G. Pierre, M. Van Steen, Wikipedia workload analysis for decentralized hosting, Computer Networks.

[40] M. Lin, A. Wierman, L. L. H. Andrew, E. Thereska, Dynamic right-sizing for power-proportional data centers, in: IEEE Infocom, 2011.

[41] A. Gandhi, M. Harchol-Balter, R. Raghunathan, M. A. Kozuch, Autoscale: Dynamic, robust capacity management for multi-tier data centers, ACM Trans. Comput. Syst. 30 (4) (2012) 14:1–14:26.

[42] Q. Wu, Making facebook's software infrastructure more energy efficient with autoscale (2014).

[43] Microsoft Azure, How to use the autoscaling application block (`http://azure.microsoft.com`).

[44] H. Wang, J. Huang, X. Lin, H. Mohsenian-Rad, Exploring smart grid and data center interactions for electric power load balancing, SIGMETRICS Perform. Eval. Rev. 41 (3) (2014) 89–94.

[45] P. D. Klemperer, M. A. Meyer, Supply function equilibria in oligopoly under uncertainty, Econometrica: Journal of the Econometric Society (1989) 1243–1277.

[46] X. Vives, Strategic supply function competition with private information, Econometrica 79 (6) (2011) 1919–1966.

[47] N. Chen, X. Ren, S. Ren, A. Wierman, Greening multi-tenant data center demand response, arXiv preprint arXiv:1504.07308.

Due to space constraints, we only include proofs of the results for the mandatory EDR scenario. The analysis for voluntary EDR is analogous to the mandatory case, though more complex. Full proofs for all the results can be found in the technical report [47].

Appendix A. Price taking tenants

Appendix A.1. Proof of Proposition 1

When tenants are price takers, they maximize the payout $P_n(b_n, p) = pS_n(b_n, p) - c_n(s_n)$ over the bid b_n. Note that $b_n \in [0, p\delta]$ as no tenant will bid beyond $p\delta$ otherwise the payout $P_n < 0$. Hence b $= (b_1, \dots, b_n)$ is an equilibrium if and only if the following condition is satisfied

$$\frac{\partial^- c_n(s_n)}{\partial s_n} \le p, \quad 0 \le b_n < p\delta, \tag{A.1a}$$

$$\frac{\partial^+ c_n(s_n)}{\partial s_n} \ge p, \quad 0 < b_n \le p\delta. \tag{A.1b}$$

At least one feasible solution to (9) exists because it is minimizing a continuous function over a compact set. Furthermore, (9b) - (9c) satisfy standard constraint qualification, hence for the Lagrangian

$$L(\mathrm{s}, \mu) = \sum_n c_n(s_n) + \mu((\delta - y) - \sum_n s_n),$$

there exists optimal primal dual pair (s, μ), such that (9b) and (9c) are satisfied, and

$$\frac{\partial^- c_n(s_n)}{\partial s_n} \le \mu, \quad s_n > 0, \tag{A.2a}$$

$$\frac{\partial^+ c_n(s_n)}{\partial s_n} \ge \mu, \quad s_n \ge 0. \tag{A.2b}$$

Given the optimal (s, μ), let $p = \mu$, and $b_n = p(\delta - s_n)$, then (9b) implies p satisfies (2), and (A.2a)-(A.2b) implies (A.1a) - (A.1b), hence an equilibrium exists.

Conversely, if (b, p) is an equilibrium and p satisfies (2), the resulting allocation s is optimal to (9). To see this, if $0 \le s_n < \delta - y$ for all n, (A.1a)-(A.1b) is equivalent to (A.2a)-(A.2b) if we set $\mu = p$, hence (s, μ) is primal dual optimal pair for (9). If $s_n = (\delta - y)$, then $s_m = 0, \forall m \neq n$. In this case, we set $\bar{\mu} = \min\{p, \partial^+ c_n(s_n)/\partial s_n\}$, and we can check that $(\mathrm{s}, \bar{\mu})$ is the primal dual optimal solution for (9).

Appendix A.2. Proof of Theorem 3

By Proposition 1, when tenants are price-taking, for any y, the there is always an equilibrium, and the resulting $\mathbf{s}$ is always the optimal allocation to provide $(\delta - y)$ energy reduction.

Hence we only need to verify that the on-site generation level y is the solution to (14a)-(14c). Similar to the proof of Proposition 1, by Assumption 2, the first order optimality condition for the y in (14a)-(14c) is $\frac{\alpha}{N\delta}(y + (N-1)\delta) = p$. By Proposition 1, p satisfies the relation (2), substitute the left-hand-side into (2) and solve for y, we have $y = \sqrt{\frac{\Sigma_n b_n N\delta}{\alpha}} - (N-1)\delta$. This is exactly the on-site generation y that minimizes $\text{cost}_o(\mathbf{b}, y)$ given in (12). Hence the datacenter will always pick y that is optimal for (14a)-(14c), together with Proposition 1, an equilibrium exists, and the resulting allocation $(\mathbf{s}, y)$ is optimal for (14a)-(14c).

Appendix A.3. Proof of Proposition 4

Since $y \ge 0$, it suffices to prove that whenever the optimal on-site generation is non-zero, $y^* > 0$, $y^t \ge y^*$. From (4), the Lagrangian of SCM is

$$L(\mathrm{s}, y, \mu^*, \lambda^*) = \sum_n c_n(s_n) + \alpha y + \mu^*((\delta - y) - \sum_n s_n) - \lambda^* y.$$

By constraint qualification and the KKT conditions, assuming $y^* > 0$, then $\lambda = 0$, $\mu^* = \alpha$, hence the market clearing price in the optimal allocation should be $p^* = \alpha$.

Next, consider the market price for price taking tenants. From (13),

$$p^t = \frac{\sum_{i\in\mathcal{N}} b_i^t}{(N-1)\delta + y^t} = \sqrt{\frac{(\Sigma_{i\in\mathcal{N}} b_i^t)\alpha}{N\delta}}. \tag{A.3}$$

The second equality yields $\sum_{i\in\mathcal{N}} b_i^t = \frac{((N-1)\delta+y^t)^2}{N\delta}\alpha$. Substitute this back to (A.3),

$$p^t = \frac{\sum_{i\in\mathcal{N}} b_i^t}{(N-1)\delta + y^t} = \frac{(N-1)\delta + y^t}{N\delta}\alpha. \tag{A.4}$$

And note that $y_t \in [0,\delta]$ and $p^* = \alpha$, thus (A.4) yields $\frac{N-1}{N}p^* \leq p^t \leq p^*$.

Finally, from (14), the Lagrangian of the price-taking characterization optimization is,

$$L(\mathrm{s}, y, \mu^t, \lambda^t) = \sum_n c_n(s_n) + \frac{\alpha}{2N\delta}(y + (N-1)\delta)^2 + \mu^t((\delta - y) - \sum_n s_n) - \lambda^t y.$$

By examining the KKT condition and using a similar argument to the proof of Proposition 1, we have $p^t = \mu^t$, also, $\frac{\partial^- c_n(s_n^t)}{\partial s_n^t} \leq p^t \leq p^* \leq \frac{\partial^+ c_n(s_n^*)}{\partial s_n^*}$. Thus, $\forall n, s_n^t \leq s_n^*$. Since $y = \delta - \sum s_n$, $y^t \geq y^*$.

Appendix A.4. Proof of Proposition 2

From the proof of Proposition 4, we see that when $y^* > 0$, $\lambda^* = 0$, and $\mu^* = \alpha$. Furthermore, we have $\sum_n s_n < \delta$, but $s_n = \delta - \frac{b_n}{\mu^*}$. Hence $(N\delta - \frac{\Sigma_n b_n}{\alpha}) < \delta$. Conversely, if (10) holds, then $\alpha(N-1)\delta < \sum_n b_n$. But by Proposition 1 and (2), we have $\sum_n b_n = (p^*(N-1)\delta + y)$. By combining the two equations above: $\alpha(N-1)\delta < p^*((N-1)\delta + y^*)$. However, from the proof in Proposition 1, we have $p^* \leq \alpha$, hence we must have $y^* > 0$.

On the other hand, when the data center operator is profit maximizing, the cost to the operator $\mathrm{cost}_o(\mathbf{b}, y) = \frac{(\Sigma_n b_n)(\delta - y)}{(N-1)\delta + y} + \alpha y$ is a convex function in y over the domain $y \geq 0$. By first order condition, the cost is minimized when

$$y' = \sqrt{\frac{N\ \delta \Sigma_n b_n}{\alpha}} - (N-1)\delta, \tag{A.5}$$

then $y = y'$ if and only if $y' \in [0,\delta]$. However, $\Sigma_n b_n = \Sigma_n p(\delta - s_n) = p((N-1)\delta + y) \leq \alpha(N\delta)$, where the last inequality is because $y \leq \delta$, and $p \leq \alpha$, since operator always has the option to use on-site generation to get unit cost of energy reduction at α. Hence we always have $y' \leq \delta$. So, if $y > 0$, by (A.5), (11) must hold, conversely, if (11) holds, then by (A.5), $y' > 0$, so operator will use $y = y'$.

Appendix A.5. Proof of Theorem 5

Note that $(\mathbf{s}^*, y^*)$ is a feasible solution to (14). By Theorem 3, we have $\sum_n c_n(s_n^t) + \frac{\alpha}{2N\delta}(y^t + (N-1)\delta)^2 \leq \sum_n c_n(s_n^*) + \frac{\alpha}{2N\delta}(y^* + (N-1)\delta)^2$. Rearranging, we have

$$\begin{aligned}
&\sum_n c_n(s_n^t) + \alpha y^t - \left(\sum_n c_n(s^*) + \alpha y^*\right) \leq \frac{\alpha}{2N\delta}(y^t - y^*)\left(2\delta - (y^t + y^*)\right) \\
=&\frac{\alpha}{2N\delta}[-(y^t - y^*)^2 + 2(\delta - y^*)(y^t - y^*)] \leq \frac{\alpha}{2N\delta}[-(y^t - y^* - (\delta - y^*))^2 + (\delta - y^*)^2] \\
=&\frac{\alpha}{2N\delta}(\delta - y^*)^2 \leq \frac{\alpha\delta}{2N}.
\end{aligned}$$

Appendix A.6. Proof of Theorem 6

From Proposition 4, we have $\frac{N-1}{N}\alpha \leq p^t \leq p^* = \alpha$, and $0 \leq y^t \leq \delta$, which yields

$$\mathrm{cost}_o^*(p^*, y^*) - \mathrm{cost}_o(p^t, y^t) = p^*(\delta - y^*) + \alpha y^* - \left(p^t(\delta - y^t) + \alpha y^t\right) = (\alpha - p^t)(\delta - y^t)$$

Substituting the above bounds for p^t and y^t gives $0 \leq \mathrm{cost}_o^*(p^*, y^*) - \mathrm{cost}_o(p^t, y^t) \leq \frac{\alpha\delta}{N}$.

Appendix B. price-anticipating tenants

Appendix B.1. Proof of Theorem 7

The proof proceeds in a number of steps. We first show that the payoff function Q_n is a concave and continuous function for each firm n. We then establish necessary and sufficient conditions for $\mathbf{b}$ to be an equilibrium; these conditions look similar to the optimality conditions (A.1a)-(A.1b) in the proof of Proposition 1, but for a "modified" cost function defined according to (21). We then show the correspondence between these conditions and the optimality conditions for the problem (20a)-(20c). This correspondence establishes existence of an equilibrium, and uniqueness of the resulting allocation.

Step 1: *If* $\mathbf{b}$ *is an equilibrium, and Assumption 2 is satisfied, at least one coordinate of* $\mathbf{b}$ *is positive.*
By Assumption 2, $0 < \alpha < \frac{\Sigma_n b_n}{(N-1)\delta}$, hence at least one coordinate of $\mathbf{b}$ must be positive.

Step 2: *The function* $Q_n(\bar{b}_n; \mathbf{b}_{-n})$ *is concave and continuous in* $\bar{b}_n$*, for* $\bar{b}_n \geq 0$. From (16) and by plugging $p(\mathbf{b})$ into s_n in (1), we have

$$Q_n(\bar{b}_n; \mathbf{b}_{-n}) = \sqrt{\frac{(\Sigma_{m\neq n} b_m + \bar{b}_n)\alpha\delta}{N}} - \bar{b}_n - c_n\left(\delta - \frac{\bar{b}_n}{\sqrt{\Sigma_{m\neq n} b_m + \bar{b}_n}}\sqrt{\frac{N\delta}{\alpha}}\right).$$

When $\Sigma_{m\neq n} b_m + \bar{b}_n > 0$, the function $\bar{b}_n / \sqrt{\Sigma_{m\neq n} b_m + \bar{b}_n}$ is a strictly concave function of $\bar{b}_n$ (for $\bar{b}_n \geq 0$). Since c_n is assumed to be convex and nondecreasing (and hence continuous), it follows that $Q_n(\bar{b}_n, \mathbf{b}_{-n})$ is concave and continuous in $\bar{b}_n$, for $\bar{b}_n \geq 0$.

It is easy to show that for s_n to be positive, we need $b_n \leq \overline{b_n}$ where $\overline{b_n} = \frac{1}{2}\left(\frac{\alpha\delta}{N} + \sqrt{\frac{\alpha\delta}{N}(\frac{\alpha\delta}{N} + 4\Sigma_{m\neq n} b_m)}\right)$.

Step 3:*In an equilibrium,* $0 \leq b_n \leq \overline{b_n}, \forall n.$
Tenant n would never bid more than $\bar{b}_n$ given $\mathbf{b}_{-n}$. If $b_n > \overline{b_n}$, then $S(p(\mathbf{b}), b_n) = \delta - \frac{b_n}{\sqrt{b_n + \Sigma_{m\neq n} b_m}} \frac{N\delta}{\alpha} < 0$. so the payoff $Q_n(b_n; \mathbf{b}_{-n})$ becomes negative; on the other hand, $Q_n(\overline{b_n}; \mathbf{b}_{-n}) = 0$.
We specify the following condition when marginal cost of production is not less than the price:

$$\forall n, \quad \frac{\partial^- c_n(s_n)}{\partial s_n} \leq p(\mathbf{b}), \quad s_n > 0. \tag{B.1}$$

This condition is satisfied when tenants are price-taking, in the next step, we show that (B.1) also holds in an equilibrium outcome when tenants are price-anticipating.

Step 4: *The vector b is an equilibrium if and only if* (B.1) *is satisfied, at least one component of* $\mathbf{b}$ *is positive, and for each n,* $b_n \in [0, \overline{b_n}]$*, and the following conditions hold:*

$$\textit{if } 0 < b_n \leq \overline{b_n}, \quad \frac{1}{2}\left(\frac{\partial^+ c_n(s_n)}{\partial s_n} + \frac{\alpha}{2N}\right) + \frac{1}{2}\sqrt{\left(\frac{\partial^+ c_n(s_n)}{\partial s_n} - \frac{\alpha}{2N}\right)^2 + \frac{\partial^+ c_n(s_n)}{\partial s_n}\frac{2s_n\alpha}{N\delta}} \geq p(\mathbf{b}), \tag{B.2a}$$

$$\textit{if } 0 \leq b_n < \overline{b_n}, \quad \frac{1}{2}\left(\frac{\partial^- c_n(s_n)}{\partial s_n} + \frac{\alpha}{2N}\right) + \frac{1}{2}\sqrt{\left(\frac{\partial^- c_n(s_n)}{\partial s_n} - \frac{\alpha}{2N}\right)^2 + \frac{\partial^- c_n(s_n)}{\partial s_n}\frac{2s_n\alpha}{N\delta}} \leq p(\mathbf{b}). \tag{B.2b}$$

By Step 2, $Q_n(b_n; \mathbf{b}_{-n})$ is concave and continuous for $b_n \geq 0$. By Step 3, $b_n \in [0, \overline{b_n}]$. b_n must maximize $Q_n(b_n; \mathbf{b}_{-n})$ over $0 \leq b_n \leq \overline{b_n}$, and satisfy the following first order optimality conditions:

$$\frac{\partial^+ Q_n(b_n; \mathbf{b}_{-n})}{\partial b_n} \leq 0, \quad \text{if } 0 < b_n \leq \overline{b_n};$$
$$\frac{\partial^- Q_n(b_n; \mathbf{b}_{-n})}{\partial b_n} \geq 0, \quad \text{if } 0 \leq b_n < \overline{b_n};$$

Recalling the expression for $p(\mathbf{b})$ given in (13), and note that by (13) and (1), we have : $\frac{1}{\sqrt{\Sigma_m b_m}} = \frac{1}{p(\mathbf{b})}\sqrt{\frac{\alpha}{N\delta}}$, and $\frac{b_n}{\sqrt{\Sigma_m b_m}} = (\delta - s_n)\sqrt{\frac{\alpha}{N\delta}}$. Expanding the first order optimality conditions with (13) and simplify with the two equations into the above, we have

$$\frac{1}{2p(\mathbf{b})}\frac{\alpha}{N} - 1 + \frac{\partial^- c_n(s_n)}{\partial s_n}\frac{1}{p(\mathbf{b})}\left(1 - \frac{1}{2p(\mathbf{b})}\frac{\alpha}{N}\frac{\delta - s_n}{\delta}\right) \leq 0. \tag{B.3a}$$

$$\frac{1}{2p(\mathbf{b})}\frac{\alpha}{N} - 1 + \frac{\partial^+ c_n(s_n)}{\partial s_n}\frac{1}{p(\mathbf{b})}\left(1 - \frac{1}{2p(\mathbf{b})}\frac{\alpha}{N}\frac{\delta - s_n}{\delta}\right) \geq 0. \tag{B.3b}$$

To show (B.1) holds, we divide into two cases, when $N \geq 2$, by rearranging (B.3a), we have

$$\frac{\partial^- c_n(s_n)}{\partial s_n}\frac{1}{p(\mathbf{b})} \leq \frac{2Np(\mathbf{b}) - \alpha}{2Np(\mathbf{b}) - \alpha\frac{\delta - s_n}{\delta}} \leq 1.$$

This is because by Assumption 2, $2Np(\mathbf{b}) - \alpha > 0$ when $N \geq 2$. Also, we have $2Np(\mathbf{b}) - \alpha\frac{\delta - s_n}{\delta} \geq 2Np(\mathbf{b}) - \alpha$. Hence (B.1) holds for $N \geq 2$.

When $N = 1$, we can simplify (B.3a) further to

$$\frac{1}{2p(\mathbf{b})}\alpha - 1 + \frac{\partial^- c_n(s_n)}{\partial s_n}\frac{1}{2p(\mathbf{b})} \leq 0, \Rightarrow p(\mathbf{b}) \geq \frac{1}{2}\left(\alpha + \frac{\partial^- c_n(s_n)}{\partial s_n}\right) \geq \frac{\partial^- c_n(s_n)}{\partial s_n}.$$

The last inequality is because $\alpha \geq \frac{\partial^- c_n(s_n)}{\partial s_n}$, otherwise $p(\mathbf{b}) > \alpha$, but profit maximizing operator will not pay for price more than α, contradiction. Hence (B.1) must hold for all N. After multiplying through (B.3a)-(B.3b) by $p(\mathbf{b})$ and rearranging, we have two quadratic inequalities in terms of $p(\mathbf{b})$. Solving the inequalities lead to two sets of conditions of $p(\mathbf{b})$ that satisfy the first order optimality conditions, they are:

$$\text{if } 0 \leq b_n < \overline{b_n}, \quad \frac{1}{2}\left(\frac{\partial^- c_n(s_n)}{\partial s_n} + \frac{\alpha}{2N}\right) \pm \frac{1}{2}\sqrt{\left(\frac{\partial^- c_n(s_n)}{\partial s_n} - \frac{\alpha}{2N}\right)^2 + 4\frac{\partial^- c_n(s_n)}{\partial s_n}\frac{s_n\alpha}{2N\delta}} \leq p(\mathbf{b}) \tag{B.4a}$$

$$\text{if } 0 < b_n \leq \overline{b_n}, \quad \frac{1}{2}\left(\frac{\partial^+ c_n(s_n)}{\partial s_n} + \frac{\alpha}{2N}\right) \pm \frac{1}{2}\sqrt{\left(\frac{\partial^+ c_n(s_n)}{\partial s_n} - \frac{\alpha}{2N}\right)^2 + 4\frac{\partial^+ c_n(s_n)}{\partial s_n}\frac{s_n\alpha}{2N\delta}} \geq p(\mathbf{b}) \tag{B.4b}$$

However, only the conditions with plus sign satisfies (B.1), the conditions with minus sign violates (B.1) because since

$$\forall s_n > 0, \quad p(\mathbf{b}) \leq \frac{\alpha}{2N} \leq \frac{\partial^+ c_n(0)}{\partial s_n} < \frac{\partial^- c_n(s_n)}{\partial s_n}.$$

Hence we discard the conditions with minus sign and note that (B.4b)-(B.4a) corresponds to (B.2a)-(B.2b).

Conversely, suppose that $\mathbf{b}$ has at least one strictly positive component, that $0 \leq b_n \leq \overline{b_n}$, and that $\mathbf{b}$ satisfies (B.1) and (B.2a)-(B.2b). Then we may simply reverse the argument: by Step 2, $Q_n(b_n; \mathbf{b}_{-n})$ is concave and continuous in $b_n \geq 0$, and in this case the conditions (B.2a)-(B.2b) imply that b_n maximizes $Q_n(b_n; \mathbf{b}_{-n})$ over $0 \leq b_n \leq \overline{b_n}$. Since we have already shown that choosing $b_n > \overline{b_n}$ is never optimal for firm n, we conclude that $\mathbf{b}$ is an equilibrium, and it is easy to check that in this case condition (B.1) is satisfied.

Step 5: *If Assumption 2 holds, then the function* $\hat{c}_n(s_n)$ *defined in* (21) *is continuous, and strictly convex and strictly increasing over* $s_n \geq 0$, *with* $\hat{c}(s_n) = 0$ *for* $s_n \leq 0$.

$\hat{c}_n(s_n)$ is continuous on $s_n > 0$ by continuity of c_n and on $s_n < 0$ by definition. We only need to show that $\hat{c}_n(0) = 0$, this is because when $s_n = 0$, $c_n(s_n) = 0$, $s_n\frac{\alpha}{2N} = 0$, and integrating from 0 to s_n is 0. Hence $\hat{c}_n(s_n) = 0$ for $s_n \leq 0$.

For $s_n \geq 0$, we simply compute the directional derivatives of $\hat{c}_n$:

$$\frac{\partial^+ \hat{c}_n(s_n)}{\partial s_n} = \frac{1}{2}\left(\frac{\alpha}{2N} + \frac{\partial^+ c_n(s_n)}{\partial s_n}\right) + \frac{1}{2}\sqrt{\left(\frac{\alpha}{2N} - \frac{\partial^+ c_n(s_n)}{\partial s_n}\right)^2 + 2\frac{\partial^+ c_n(s_n)}{\partial s_n}\frac{s_n \alpha}{N\delta}},$$

$$\frac{\partial^- \hat{c}_n(s_n)}{\partial s_n} = \frac{1}{2}\left(\frac{\alpha}{2N} + \frac{\partial^- c_n(s_n)}{\partial s_n}\right) + \frac{1}{2}\sqrt{\left(\frac{\alpha}{2N} - \frac{\partial^+ c_n(s_n)}{\partial s_n}\right)^2 + 2\frac{\partial^+ c_n(s_n)}{\partial s_n}\frac{s_n \alpha}{N\delta}}.$$

Since c_n is strictly increasing and convex, for $0 \leq s_n < \bar{s}_n$, we will have

$$0 \leq \frac{\partial^+ \hat{c}(s_n)}{\partial s_n} < \frac{\partial^- \hat{c}(\bar{s}_n)}{\partial s_n} \leq \frac{\partial^+ \hat{c}(\bar{s}_n)}{\partial s_n}.$$

This guarantees that $\hat{c}_n$ is strictly increasing and strictly convex over $s_n \geq 0$.

Step 6: *There exists a unique vector* $\mathbf{s} \geq 0, y \geq 0$ *and at least one scalar* $\rho > 0$ *such that:*

$$\frac{1}{2}\left(\frac{\partial^+ c_n(s_n)}{\partial s_n} + \frac{\alpha}{2N}\right) + \frac{1}{2}\sqrt{\left(\frac{\partial^+ c_n(s_n)}{\partial s_n} - \frac{\alpha}{2N}\right)^2 + \frac{\partial^+ c_n(s_n)}{\partial s_n}\frac{2s_n\alpha}{N\delta}} \geq \rho, \quad \textit{if } s_n \geq 0; \tag{B.5a}$$

$$\frac{1}{2}\left(\frac{\partial^- c_n(s_n)}{\partial s_n} + \frac{\alpha}{2N}\right) + \frac{1}{2}\sqrt{\left(\frac{\partial^+ c_n(s_n)}{\partial s_n} - \frac{\alpha}{2N}\right)^2 + \frac{\partial^+ c_n(s_n)}{\partial s_n}\frac{2s_n\alpha}{N\delta}} \leq \rho, \quad \textit{if } s_n > 0; \tag{B.5b}$$

$$\frac{\alpha}{N\delta}(y + (N-1)\delta) = \rho; \tag{B.5c}$$

$$\sum_n s_n = (\delta - y). \tag{B.5d}$$

The vector $\mathbf{s}$ *and* y *is then the unique optimal solution to* (20a)-(20c).

By Step 5, since $\hat{c}_n$ is continuous and strictly over the convex, compact feasible region for each n, we know that (20a)-(20c) have a unique optimal solution $\mathbf{s}, y$. As in the proof of Proposition 1, form the Lagrangian

$$L(\mathbf{s}, y; \rho) = \sum_n \hat{c}_n(s_n) + \frac{\alpha}{2N\delta}(y + (N-1)\delta)^2 + \rho((\delta - y) - \sum_n s_n).$$

By assumption 2, $y > 0$, and by the fact that $\hat{c}_n(s_n) = 0$ for $s_n \leq 0$, $s_n \geq 0$. there exists a Lagrange multiplier ρ such that $(\mathbf{s}, y, \rho)$ satisfy the stationarity conditions which corresponds to (B.5a)-(B.5c) when we expand the definition of $\hat{c}_n(s_n)$, together with the constraint (B.5d). The fact that $\rho > 0$ follows by (B.5c) as $y > 0$.

Step 7: *If* $\mathbf{s} \geq 0, y \geq 0$ *and* $\rho > 0$ *satisfy* (B.5a)-(B.5d)*, then the triple* $(\mathbf{b}, \rho, y)$ *defined by* $b_n = (\delta - s_n)\rho$ *is an equilibrium as defined in* (17) *and* (18).

First observe that with this definition, together with (B.5d) and the fact that $s_n \geq 0$, we have $b_n \geq 0$ for all n. Furthermore, we can show $b_n \leq \overline{b_n}$, since $s_n \geq 0$, $b_n \leq \rho\delta$, but by (B.5c)-(B.5d), we have

$$\rho = \frac{\alpha}{N\delta}(y + (N-1)\delta) = \frac{\alpha}{N\delta}(N\delta - \sum_n s_n) \tag{B.6}$$

Substitute the definition $s_n = \delta - \frac{b_n}{\rho}$ into (B.6), we have

$$\rho = \frac{\alpha}{N\delta}\frac{\Sigma_n b_n}{\rho} \Rightarrow \rho = \sqrt{\frac{\Sigma_n b_n \alpha}{N\delta}}. \tag{B.7}$$

Substituting (B.7) into $b_n \leq \rho\delta$, we have $b_n \leq \sqrt{\frac{(\Sigma_{m\neq n} b_m + b_n)\alpha\delta}{N}}$, Solving this inequality we have $b_n \leq \overline{b_n}$.

Finally, at least one component of **b** is strictly positive, since otherwise we have $s_{n1} = s_{n2} = \delta$ for some $n1 \neq n2$, in which case $\Sigma_n s_n > \delta$, which contradicts (B.5d). (or $s_n = \delta$, $y = 0$, contradicting our assumption that $y > 0$.)

By Step 4, to check that **b** is an equilibrium, we must only check the stationarity conditions (B.2a)-(B.2b). We simply note that under the identification $b_n = \rho(\delta - s_n)$, using (B.7) and (B.5c), we have

$$y = \sqrt{\frac{\Sigma_n b_n N\delta}{\alpha}} - (N-1)\delta; \quad \rho = \frac{\Sigma_n b_n}{(N-1)\delta + y} = p(\mathbf{b}).$$

Substitute $p(\mathbf{b})$ into (B.5a) will correspond to (B.2a), and (B.5b) implies (B.2b) and (B.1) because $\frac{\partial^- c_n(s_n)}{\partial s_n} \leq \frac{\partial^+ c_n(s_n)}{\partial s_n}$. Thus $(\mathbf{b}, \rho, y)$ is an equilibrium.

Step 8: *If* $(\mathbf{b}, p(\mathbf{b}), y)$ *is an equilibrium, then there exists a scalar* $\rho \geq 0$ *such that the vector* **b** *defined by* $s_n = S(p(\mathbf{b}), b_n)$ *satisfies* (B.5a)-(B.5d).

We simply reverse the argument of Step 7. Since **b** is an equilibrium bids, by (18) and $s_n = S(p(\mathbf{b}), b_n)$, we have $\sum_n s_n = (\delta - y)$, i.e., (B.5d) is satisfied. By Step 4, **b** satisfies (B.2a)-(B.2b). Since $y > 0$ by Assumption 2, $0 \leq s_n < \delta$ for all n, let

$$\rho = \max\left\{p(\mathbf{b}), \frac{1}{2}\left(\frac{\partial^- c_n(s_n)}{\partial s_n} + \frac{\alpha}{2N}\right) + \frac{1}{2}\sqrt{(\frac{\partial^+ c_n(s_n)}{\partial s_n} - \frac{\alpha}{2N})^2 + \frac{\partial^+ c_n(s_n)}{\partial s_n}\frac{2 s_n \alpha}{N\delta}}\right\}.$$

In this case $\rho > 0$ and $0 \leq b_n \leq \overline{b_n}$ for all n, so (B.2b) implies (B.5b) by definition of ρ, and (B.5a) holds by (B.2a) and the fact that $\partial^- c_n(s_n) \leq \partial^+ c_n(s_n)$ (by convexity).

Step 9: *There exists an equilibrium* **b***, and for any equilibrium that price is greater than marginal cost, the vector* **s** *defined by* $s_n = S(p(\mathbf{b}), b_n)$ *is the unique optimal solution of* (B.5a)-(B.5d).

The conclusion is now straightforward. Existence follows from Steps 6 and 7. Uniqueness of the resulting production vector **s**, and the fact that **s** is an optimal solution to (20a)-(20c), follows by Steps 6 and 8.

Appendix B.2. Proof of Lemma 8

We exploit the structure of the modified cost $\hat{c}_n$ to prove the result. Note that, for all n, $s_n \geq 0$, if we define $G_n(s_n) = \int_0^{s_n} \sqrt{(\frac{\partial^+ c_n(z)}{\partial z} - \frac{\alpha}{2N})^2 + \frac{\partial^+ c_n(z)}{\partial z}\frac{2z\alpha}{N\delta}}dz$, then

$$G_n(s_n) \geq \int_0^{s_n} \sqrt{\left(\frac{\partial^+ c_n(z)}{\partial z} - \frac{\alpha}{2N}\right)^2} dz = c_n(s_n) - s_n\frac{\alpha}{2N}.$$

First inequality is because $z \geq 0$, last equality is because by convexity and Assumption 3, we have $\frac{\partial^+ c_n(z)}{\partial z} \geq \frac{\partial^+ c_n(0)}{\partial s_n} \geq \frac{\alpha}{2N}$.

Hence we have $\hat{c}_n(s_n) = \frac{1}{2}\left(c_n(s_n) + s_n\frac{\alpha}{2N}\right) + \frac{1}{2}G_n(s_n) \geq c_n(s_n)$. On the other hand, notice that $s_n \leq \delta$, we have:

$$\begin{aligned} G_n(s_n) &\leq \int_0^{s_n} \sqrt{\left(\frac{\partial^+ c_n(z)}{\partial z} - \frac{\alpha}{2N}\right)^2 + \frac{\partial^+ c_n(z)}{\partial z}\frac{2\delta\alpha}{N\delta}}dz \\ &= \int_0^{s_n} \sqrt{\left(\frac{\partial^+ c_n(z)}{\partial z} + \frac{\alpha}{2N}\right)^2} dz = c_n(s_n) + s_n\frac{\alpha}{2N}. \end{aligned}$$

Hence we have $\hat{c}_n(s_n) = \frac{1}{2}\left(c_n(s_n) + s_n\frac{\alpha}{2N}\right) + \frac{1}{2}G_n(s_n) \leq c_n(s_n) + s_n\frac{\alpha}{2N}$. The bounds for the left and right derivatives can be obtained from taking the left (or right) derivatives at the bounds of $G_n(s_n)$.

Appendix B.3. Proof of Theorem 9

Firstly we will prove one side of the inequality $p^t \leq p^a, y^t \leq y^a$. Recall that by the examinging the Lagrangians of the optimizations in Proposition 4 in and Theorem 7, we have $p^t \geq \partial^- c_n(s_n^t)/\partial s_n$, $p^t \leq \partial^+ c_n(s_n^t)/\partial s_n$, $p^a \geq \partial^- \hat{c}_n(s_n^a)/\partial s_n$, $p^a \leq \partial^+ \hat{c}_n(s_n^a)/\partial s_n$, at the domain where the left or right derivative is defined, and $p^t = \frac{\alpha}{N\delta}(y^t + (N-1)\delta), p^a = \frac{\alpha}{N\delta}(y^a + (N-1)\delta)$. If $y^t > y^a$, then $p^t > p^a$. Also, because the total energy reduction δ is constant, we have $\sum_n s_n^t < \sum_n s_n^a$.

Hence there exist $s_r > 0$ such that $s_r^a > s_r^t$ for some $r \in \{1, \dots, N\}$. Therefore, by strict convexity of c_n (Assumption 1):

$$p^t \leq \frac{\partial^+ c_r(s_r^t)}{\partial s_r} < \frac{\partial^- c_r(s_r^a)}{\partial s_r}. \tag{B.8}$$

However, by Lemma 8 we have $\frac{\partial^- \hat{c}_r(s_r)}{\partial s_r} \geq \frac{\partial^- c_r(s_r)}{\partial s_r}$. Hence, we have

$$p^a \geq \frac{\partial^- \hat{c}_r(s_r^a)}{\partial s_r} \geq \frac{\partial^- c_r(s_r^a)}{\partial s_r}. \tag{B.9}$$

Combining (B.8) and (B.9), we have $p^t < p^a$, contradiction. Hence we have $y^t \leq y^a$, and $p^t \leq p^a$.

Next we show the other side of the inequality $p^a \leq p^t + \frac{\alpha}{2N}, y^a \leq y^t + \frac{\delta}{2}$, by the previous part, we have $\sum_n s_n^a \leq \sum_n s_n^t$. Let $n = \arg\max_m (s_m^t - s_m^a)$, clearly $s_n^t \geq s_n^a$, otherwise $\sum_n s_n^t < \sum_n s_n^a$, contradiction.

If $s_n^t = s_n^a$, then $\forall m, s_m^t = s_m^a$, and $y^t = y^a$, then $p^t = p^a$.

If $s_n^t > s_n^a$, then by strict convexity of c_n (assumption 1), and the fact that $s_n^a \geq 0, s_n^t > 0$, we have $\frac{\partial^+ \hat{c}_n(s_n^a)}{s_n} < \frac{\partial^- c_n(s_n^t)}{s_n} \leq p^t$. Also, by Lemma 8, we have $\frac{\partial^+ \hat{c}_n(s_n)}{\partial s_n} \leq \frac{\partial^+ c_n(s_n)}{\partial s_n} + \frac{\alpha}{2N}$, this gives us $p^a \leq \frac{\partial^+ \hat{c}_n(s_n^a)}{\partial s_n} \leq \frac{\partial^+ c_n(s_n^a)}{\partial s_n} + \frac{\alpha}{2N}$. Combining the two previous inequalities about p^t and p^a, we have $p^a < p^t + \frac{\alpha}{2N}$. Hence we have

$$\frac{\alpha}{N\delta}(y^a + (N-1)\delta) < \frac{\alpha}{N\delta}(y^t + (N-1)\delta) + \frac{\alpha}{2N} \Rightarrow y^a < y^t + \frac{\delta}{2}.$$

Appendix B.4. Proof of Theorem 11

As $(\mathbf{s}^*, y^*)$ is a feasible solution to (20), by Theorem 7, we have

$$\sum_n \hat{c}_n(s_n^a) + \frac{\alpha}{2N\delta}(y^a + (N-1)\delta)^2 \leq \sum_n \hat{c}_n(s_n^*) + \frac{\alpha}{2N\delta}(y^* + (N-1)\delta)^2. \tag{B.10}$$

Rearranging, we have $\sum_n \hat{c}_n(s_n^a) + \alpha y^a - (\sum_n \hat{c}_n(s_n^*) + \alpha y^*) \leq \frac{\alpha}{N}\left((y^a - y^*)(1 - \frac{y^a + y^*}{2\delta})\right)$. By Corollary 10 and the fact that $y^* \leq \delta, y^a \leq \delta$, both terms in the brackets are positive, hence right-hand-side expression is maximized when $y^* \to 0^+$ and $y^a = \delta$, hence

$$\left(\sum_n \hat{c}_n(s_n^a) + \alpha y^a\right) - \left(\sum_n \hat{c}_n(s_n^*) + \alpha y^*\right) \leq \frac{\alpha\delta}{2N}. \tag{B.11}$$

However, by Lemma 8, we have $\sum_n \hat{c}_n(s_n^*) \leq \sum_n c_n(s_n^*) + \frac{\alpha}{2N}(\sum_n s_n) \leq \sum_n c_n(s_n^*) + \frac{\alpha\delta}{2N}$; and $\sum_n \hat{c}_n(s_n^a) \geq \sum_n c_n(s_n^a)$. Substituting the above relations into (B.11) and rearranging, we have the desired result.

Appendix B.5. Proof of Theorem 12

First, we compare the cost by operator between the price-taking and price anticipating cases, by definition (15) and rearranging, we have $\text{cost}_o(p^a, y^a) - \text{cost}_o(p^t, y^t) = (p^a - p^t)(\delta - y^t) + (\alpha - p^a)(y^a - y^t)$. By the fact that $p^a = \frac{\alpha}{N\delta}(y^a + (N-1)\delta)$ (shown in Theorem 9) and the fact that $0 \leq y^a \leq \delta$, we have

$$\alpha\left(\frac{N-1}{N}\right) \leq p^a \leq \alpha. \tag{B.12}$$

By the upper bound of p^a in (B.12) and the upper bounds of p^t, y^t in Theorem 9, we have

$$\text{cost}_o(p^a, y^a) - \text{cost}_o(p^t, y^t) \geq 0. \tag{B.13}$$

Similarly, using the lower bound of p^a in (B.12) and the upper bounds of p^a, y^a in Theorem 9, we have

$$\text{cost}_o(p^a, y^a) - \text{cost}_o(p^t, y^t) \leq \left(\frac{\alpha}{2N}\right) \cdot (\delta) + \left(\alpha \cdot \frac{1}{N}\right)\left(\frac{\delta}{2}\right) = \frac{\alpha\delta}{N}.$$

Second, we compare the cost by the operator to the social optimal. Since the energy reduction goal δ is the same, by Proposition 4 and Corollary 10, we have $p^t \leq p^*$ and $p^a \leq p^*$. Hence we have $\text{cost}_o(p^t, y^t) \leq \text{cost}_o(p^a, y^a) \leq \text{cost}_o(p^*, y^*)$. Furthermore,

$$\begin{aligned} &\text{cost}_o(p^*, y^*) - \text{cost}_o(p^t, y^t) = \alpha\delta - (p^t(\delta - y^t) + \alpha y^t) \\ =&(\alpha - p^t)(\delta - y^t) = \alpha\left(\frac{\delta - y^t}{N\delta}\right)(\delta - y^t) \leq \frac{\alpha\delta}{N}. \end{aligned} \tag{B.14}$$

Lastly by (B.13) and (B.14), we have $\text{cost}(p^*, y^*) - \text{cost}(p^a, y^a) \leq \text{cost}(p^*, y^*) - \text{cost}(p^t, y^t) \leq \frac{\alpha\delta}{N}$.

Appendix B.6. Proof of Theorem 13

Given any $\varepsilon > 0$, let $\varepsilon' = \frac{1}{2}\varepsilon$. Consider the following set of cost function:

$$c_1(s_1) = \begin{cases} \frac{\alpha}{2N} s_1, & \text{if } s_1 < \varepsilon'; \\ \alpha(1 - \frac{3\varepsilon'}{2N\delta})s_1 + C_1, & \varepsilon' \leq s_1 \leq \delta - \varepsilon'; \\ 2\alpha s_1 + C_2, & s_1 > \delta - \varepsilon' \end{cases}$$

where C_1, C_2 are constants that make c_1 continuous[6], then c_1 is piece-wise linear and convex. Also, $\forall m \neq 1, c_m(s_m) = 2\alpha s_m$. It is easy to see that $s_1^* = \delta - \varepsilon'$ and $y^* = \varepsilon'$ is the optimal allocation.

Let $s_1^a = \varepsilon', y^a = \delta - \varepsilon'$, and $\forall m \neq 1, s_m^a = 0$, we claim that $(\mathbf{s}^a, y^a)$ is the unique optimal solution to (20a)-(20c). To see this, let $\rho = \alpha(1 - \varepsilon/(N\delta))$, then,

$$\frac{\alpha}{N\delta}(y^a + (N-1)\delta) = \rho; \quad \sum_n s_n^a = \delta - y^a; \tag{B.15a}$$

$$\frac{\partial^- \hat{c}_1(s_1^a)}{\partial s_1} \leq \rho; \quad \frac{\partial^+ \hat{c}_1(s_1^a)}{\partial s_1} \geq \rho; \quad \frac{\partial^+ \hat{c}_m(0)}{\partial s_m} \geq \rho, \quad \forall m \neq 1. \tag{B.15b}$$

where the second inequality is because if we let H_n be the term under square root for $\frac{\partial^+ \hat{c}_n(s_n)}{\partial s_n}$, then

$$\begin{aligned} H_n =& \sqrt{\left(\frac{\partial^+ c_n(s_n)}{\partial s_n} - (\frac{\alpha}{2N} - \frac{\alpha}{N}\frac{s_n}{\delta})\right)^2 + (\frac{\alpha^2}{N^2}\frac{(\delta + s_n)(\delta - s_n)}{\delta^2})} \\ \geq& \frac{\partial^+ c_n(s_n)}{\partial s_n} - (\frac{\alpha}{2N} - \frac{\alpha}{N}\frac{s_n}{\delta}). \end{aligned}$$

Note that $\frac{\partial^+ \hat{c}_n(s_n)}{\partial s_n} = \frac{1}{2}(\frac{\partial^+ c_n(s_n)}{\partial s_n} + \frac{\alpha}{2N}) + \frac{1}{2}H_n$. Hence we have $\frac{\partial^+ \hat{c}_1(s_1^a)}{\partial s_1} \geq \frac{\partial^+ c_1(s_1^a)}{\partial s_1} + \frac{\alpha s_1}{2N\delta} = \rho$. These conditions correspond to (B.5a)-(B.5d), so we conclude that $(\mathbf{s^a}, y^a)$ is the unique optimal solution to (20a)-(20c). Hence $y^a - y^* = \delta - 2\varepsilon' = \delta - \varepsilon$.

[6] $C_1 = -\alpha\varepsilon'(\frac{(2N-1)\delta - 3\varepsilon'}{2N\delta})$, and $C_2 = -\frac{\alpha}{N\delta}(N\delta^2 + \delta\varepsilon' - 3\varepsilon')$