Performance Evaluation 91 (2015) 229–254

Contents lists available at ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Greening multi-tenant data center demand response[☆]

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

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

ARTICLE INFO

Article history:
Available online 4 July 2015

Keywords:
Demand response
Mechanism design
Multi-tenant data center
Supply function bidding

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).

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 US 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

☆ This work is supported in part by the US NSF CNS-1319820, EPAS-1307794, CNS-1423137, and CNS-1453491.
* Corresponding author.
E-mail addresses: ncchen@caltech.edu (N. Chen), xren@caltech.edu (X. Ren), sren@ece.ucr.edu (S. Ren), adamw@caltech.edu (A. Wierman).

http://dx.doi.org/10.1016/j.peva.2015.06.014
0166-5316/

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].

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 US, 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 US 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$/W [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., [13–17,12,2]. 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 US, 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 heterogeneous, 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

[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].