2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS): 2016 IEEE Infocom Green and Sustainable Networking and Computing (GSNC) Workshop

Incentive-aligned Mechanism for Emergency Demand Response in Multi-tenant Mixed-Use Buildings

Nguyen H. Tran*, Chuan Pham*, Minh N.H Nguyen*, Shaolei Ren‡, Choong Seon Hong*
*Department of Computer Engineering, Kyung Hee University
‡Department of Electrical & Computer Engineering, University of California at Riverside

***Abstract*—Buildings and multi-tenant datacenters (MDCs) have been identified as crucial participants for emergency demand response (EDR) which is the last line of defense to avoid cascading failures during emergency events. One important overlooked fact is that the majority of MDCs are physically located in mixed-use buildings (MUBs) and share the electricity supply with other operations (e.g., office spaces). However, the existing studies on EDR have not considered this fact and EDR for buildings treats MDC operation as "miscellaneous loads", which nullifies the flexible electricity consumption of a MDC. Furthermore, even when both building offices and MDCs are jointly considered for EDR, tenants will incur cost to shed energy for EDR, which raises serious incentive questions for their participation. To overcome this *uncoordinated energy shedding and mis-aligned incentives*, we propose a *first-of-its-kind* incentive mechanism for EDR in MUBs, such that the total incurred loss (e.g., latency performance degradation for tenants of MDC, thermal discomfort for office spaces) is minimized for energy shedding during EDR. We also design a distributed algorithm to implement the incentive mechanism that can optimally (a) control indoor temperature for non-MDC space, (b) perform server provisioning in MDC, and (c) manage on-site electricity generation. Simulation results show that our algorithm has better performance in terms of MUB total cost compared to the current non-coordinated approaches.**

I. Introduction

Demand response has proven to be an effective technique to modulate consumer's power demand via market-based approaches for the time-varying electricity generation. Due to the aging infrastructure, frequent extreme weather and/or wide adoption of renewables, the power grid is becoming increasingly fragile. Hence, emergency demand response (EDR) has become one of the most widely-adopted demand response programs in the U.S., representing 87% of demand reduction capabilities across all reliability regions [1]. EDR protects the power grid against cascading blackouts by coordinating multiple energy consumers to shed their loads during emergency events (e.g., extreme weather).

Ideal participants in EDR programs are large energy consumers which include buildings and datacenters [2]. In the U.S., buildings consume approximately 40% of total generated electricity [3] while datacenters have large yet flexible power demands [4]. However, *mixed-use buildings* (MUBs) where there are both datacenter operations and non-datacenter operations (e.g., office spaces) have been largely overlooked in EDR programs. In fact, according to a report by Green Grid [5], "the majority of datacenters are located within mixed-use facilities". A recent study [6] presented that large dedicated datacenters (e.g., Google) only account for 4% of the total datacenter energy consumption, whereas the remaining 96% is used by other types of datacenters (e.g., scientific computing cluster, multi-tenant datacenters, server room) that are mostly located in MUBs. Especically, multi-tenant datacenters (MDCs), which consume up to 40% of the datacenter energy in the U.S. [7], play a vital role in datacenter industry [8], [9].

At present, buildings and MDCs participate in EDR by means of their on-site generators (usually diesel) which are economically costly and ecologically unfriendly. Consequently, many have proposed to explore green alternatives for EDR. One approach proposed to reduce the electricity consumption by cutting the non-critical energy usage (e.g. heating, ventilation and air conditioning, or HVAC, switching of unused lighting) instead of on-site diesel generation [10]. However, these studies are not applicable to MUBs with MDC operation since MDCs are identified as "miscellaneous loads", and thus, the flexible electricity consumption of a MDC is wasted. One studies state that datacenters operation may account for 50% or more of total energy consumption of a MUB [11].

Similarly, EDR for datacenter has been extensively researched [4], [8], [12]. These works take advantage of widely available IT control knobs (e.g., server turning on/off and workload migration). Further, a field study by Lawrence Berkeley National Laboratory has demonstrated that datacenters can reduce energy consumption by 10-25% for EDR, without significantly impacting their normal operations [13]. Nonetheless, the existing research is targeted towards both owner-operated datacenters and MDCs where all the spaces and supporting infrastructure (e.g., cooling) are directly associated with datacenters [5]. Hence, despite sharing main electrical power line in MUBs, the existing studies on demand response for buildings and MDCs have been isolated from each other. This results in *uncoordinated* energy management and leading to inefficient EDR for MUBs.

Especially, even when careful designers attempt to coordinate both building offices and MDCs, challenges further escalate. Since chiller is shared among different tenants and owned by the MUB manager and each individual tenant with small load may not qualify for EDR, the MUB operator acts as an aggregator for EDR. However, while the operator only

978-1-4673-9955-5/16/$31.00 ©2016 IEEE

provides facilities (e.g., power, cooling), MDC tenants which can flexibly vary energy consumption with several IT-knobs have individual costs and objectives to shed energy for EDR, inducing a *mis-aligned incentive* that further complicate the coordinated approach.

To overcome these challenges, we propose a mechanism that not only coordinates both building offices and MDCs but also aligns their incentives to enable EDR in MUBs, such that the total incurred loss (e.g., latency performance degradation for tenants of MDCs, thermal discomfort for office spaces) is minimized for energy shedding. During EDR, the MUB operator has multiple options: shedding MDC energy (via, e.g., turning off unused servers, scaling down CPU frequency), shedding non-MDC energy (via, e.g., increasing temperature set-point), and turning on on-site backup generation (usually, diesel generator). A *coordinated and incentive-aligned* approach is required because all three options mentioned have their limitations and drawbacks: shedding MDC energy can possibly degrade application performance, tuning HVAC temperature set-point for non-MDC space results in human discomfort, while using diesel generation contaminates the environment. Thus, the facility manage must carefully control these three energy reduction knobs to minimize the overall negative impact while still satisfying the total energy reduction requirement for EDR. Towards this end, we formulate a cost minimization problem and propose an incentive mechanism implemented by a distributed algorithm that can coordinately control optimal HVAC temperature for non-MDC space, server provisioning and load balancing in MDC, and usage of diesel generation. Then, we performed a case study to validate our approach, which shows that the proposed algorithm outperforms the existing non-coordinated approach in terms of total incurred cost.

II. Preliminaries

A. Mixed-Use Building

In general, a MUB refers to a building with a combination of multiple distinct uses, such as residential, office, lab, industrial, among others. For this paper, we explicitly focus on MUB that includes MDC operation and a *significant* space for *non-MDC* functions. In a MUB with MDC operation, the IT load may usually take up to 50% or more of a building's overall energy consumption [14]. Unless otherwise stated, we assume that the default function for the large non-MDC space in a MUB is for office[1]. Henceforth, "MUB" mentioned in this work refers to MUB with MDC operations which shares the building space and main electrical power supply with the office operation. Servers often have dedicated backup power infrastructure (e.g., uninterrupted power supply, or UPS) for emergency. In some cases, cooling system (e.g., chiller and cooling tower) may be shared between the MDC and the office [5]. Note that MDCs in a MUB can be of various types, ranging from state-of-the-art commercial MDC to scientific computing cluster and to small-/medium-size server rooms.

[1]Dedicated data-center, like Google, also has office space, but it is negligible compared to the data-center part.

B. Emergency Demand Response

EDR requires a mandatory power demand reduction response (with a significant penalty for non-compliance) for the participants, who are mostly paid for their availability for shedding loads even when no emergent signal is triggered [15]. At present, such programs are employed by many Independent System Operators, e.g., New England, where the customers' contracts can be established three years in advanced [16]. In particular, if there are some reliability issues in the grid (e.g., forecast capacity shortages or extreme weather event), the load serving entity (LSE) will trigger a signal to customers from at least 10 minutes to one day in advance, and customers must comply with the notified reduction volume. The problem is that at present, customers often participate in EDR using on-site backup diesel generators which is the least favorable choice economically and ecologically. Hence, in this work, we present alternatives for the MUB operator for EDR, only using the on-site diesel generation as a last resort.

III. System Modelling

We consider a typical MUB: a datacenter physically co-located with non-datacenter (office) operation inside a MUB managed by the same operator. We consider a one-period demand response, as in [8], [17]–[19], where its duration Δt is controlled by an LSE, e.g., 15 minutes or 1 hour.

A. HVAC Energy and User Comfort

We consider a set $\mathcal{N}$ of N offices in the building. Even though reducing the building office energy has attracted much attention, there exists a critical trade-off issue between building energy reduction and user comfort satisfaction that dictates the performance of indoor environments. Generally, building energy is mainly contributed by: (a) HVAC system, and (b) lighting and electrical equipments, which constitute 43% and 30% of building energy usage, respectively [3].

Energy reduction by controlling HVAC. Adopted from the energy-temperature correlation model in [20], the energy consumed by HVAC during a period Δt of an office n, which depends on the difference between the mean outdoor temperature[2] T^o and indoor temperature T_n^i of office n controlled by HVAC system, is given as follows

$$q_n(T_n^i) = \frac{m_n}{M} \left| T_n^i - T^o \right| \Delta t, \quad \forall n \in \mathcal{N}. \tag{1}$$

where m_n is the conductivity of the office n and M is energy transformation which indicates the energy efficiency of the HVAC system.

In practice, the offices often put high priority on their user comfort. Therefore, the indoor temperature is usually set to comfort temperature $T^c(T^o)$, which is an affine function of the mean outdoor temperature [20]. Henceforth, for brevity we will use T^c without its argument. With this comfort temperature, the HVAC energy consumption is $q_n(T^c)$. When

[2]We assume that all offices in the building have the same mean outdoor temperature. Furthermore, the temperature unit is implicitly used as Celsius.

the office operator adjusts the indoor temperature to T_n^i such that $q_n(T_n^i) \leq q_n(T^c)$, the HVAC energy is reduced as follows

$$\begin{aligned} e_n &:= q_n(T^c) - q_n(T_n^i) \\ &= \kappa_n \, |T_n^i - T^c|, \quad \forall n \in \mathcal{N}, \end{aligned} \tag{2}$$

where $\kappa_n := \frac{m_n}{M} \, \Delta t$.

W.l.o.g., we consider the EDR time slot in summer (e.g., T^o = 32°C). Therefore, in order to have a non-negative reduced energy (i.e., $e_n \geq 0$) of the HVAC cooling system, we have

$$\bar{T}_n := T_n^i - T^c \geq 0, \tag{3}$$

and we can rewrite in (2) as an energy-temperature difference correlation in the following

$$e_n = \kappa_n \, \bar{T}_n, \quad \forall n \in \mathcal{N}. \tag{4}$$

In this work, we mainly consider the impact of HVAC system on user comfort.

User comfort model. Even though user comfort is an abstract concept and heavily depends on individual tastes, we consider a comfort model as in [20]. In this model, the user comfort consists of two components: (a) heat gain $G_n(t)$ which depends on the mean time t an office user spend in the room, and (b) heat loss $L_n(\bar{T}_n)$ which depends on the difference between indoor and comfort temperature.

Since $\bar{T}_n \geq 0$ in our model, the heat loss model from [20] can be presented as follows

$$L_n(\bar{T}_n) = \begin{cases} 3, & \bar{T}_n > R; \\ k \, \bar{T}_n, & 0 \leq \bar{T}_n \leq R, \end{cases} \tag{5}$$

where $k = 2/7$ and $R = 21/2$.

The user comfort, which is the sum of heat gain and loss and normalized in the range of $[-3, 3]$ to encode the feeling from cold to hot according to AHSAH model [21], is as follows

$$C_n^{cf} = G_n(t) + L_n(\bar{T}_n). \tag{6}$$

Then we have the comfort of all office users as follows

$$C^{cf}(\{\bar{T}_n\}) = \omega^{cf} \sum\nolimits_{n \in \mathcal{N}} C_n^{cf}(\bar{T}_n), \tag{7}$$

where ω^{cf} is the weight represents the unit discomfort cost of the office users due to the difference of indoor temperature and comfort temperature.

B. Datacenter

We consider a colo-datacenter in which a set of $\mathcal{I} = \{1, \ldots, I\}$ tenants house their servers. Tenant i has S_i homogeneous servers. A tenant with heterogeneous servers can be viewed as multiple virtual tenants, each having homogeneous servers.

Energy reduction of datacenters. Even though tenants may use various control knobs (e.g., scaling down CPU frequencies, migrating loads to other places) for energy saving, the simple yet widely-studied approach that our study adopts as an example is turning off idle servers [8], [19], [22]. If tenant i has no intention to participate in demand response, all of its servers are active, and the workload will be evenly distributed to all servers to optimize performance [22]; hence, the total power consumption (i.e., cooling and IT) of this case is [8]

$$q_i(S_i) = S_i \Big(p_{i,s} + p_{i,a} \frac{\lambda_i}{S_i \mu_i} \Big) \, \Delta t \, PUE, \tag{8}$$

where $p_{i,s}$ and $p_{i,a}$ are the static and active powers of each server, respectively, λ_i is the workload arrival rate, μ_i is a server's service rate measured in terms of the amount of workload processed per unit time, $\frac{\lambda_i}{S_i \mu_i}$ is the server utilization with S_i active servers, and PUE is the power usage effectiveness of a datacenter, which is measured by total power consumption divided by IT power consumption. When performing demand response by turning off s_i servers, the total power consumption of tenant i is

$$q_i(s_i) = (S_i - s_i) \Big(p_{i,s} + p_{i,a} \frac{\lambda_i}{(S_i - s_i)\mu_i} \Big) \, \Delta t \, PUE. \tag{9}$$

Therefore, the total energy reduction by tenant i is

$$e_i := q_i(S_i) - q_i(s_i) = \gamma_i s_i, \quad \forall i \in \mathcal{I}, \tag{10}$$

where $\gamma_i := p_{i,s} \, PUE \, \Delta t$ is a constant value of tenant i.

Turning servers off can have negative effects on tenant performance, inducing tenant costs. We rely on two typical costs that are widely used for tenants: the wear-and-tear cost and Service Level Agreement (SLA) cost [8], [22].

SLA cost. Since many Internet services hosted in datacenters are sensitive to response/delay time, the SLA cost can be viewed proportionally to tenant average response time. Using the M/M/1 queue, the average response time of each tenant i's workload is

$$d_i(s_i) := \frac{1}{\mu_i - \frac{\lambda_i}{S_i - s_i}}. \tag{11}$$

We note that the queueing model has been widely used as a reasonable approximation for the actual service process [23], [24]. When s_i increases, the workload distributed to the remaining active servers (i.e. $\frac{\lambda_i}{S_i - s_i}$) increases due to the added migrating load, which leads to the increase of $d_i(s_i)$.

Wear-and-tear cost. This cost occurs when tenants switch/toggle servers between active and idle states in every period and is linear with the number of turned-off servers [22].

Therefore, tenant i's total cost when turning off s_i servers is

$$C_i^{dc}(s_i) = \omega^{wat} \, s_i + \omega^{sla} \, \lambda_i d_i(s_i), \tag{12}$$

where ω^{wat} and ω^{sla} are the weights represent the unit cost of wear-and-tear and SLA. Then we have the total cost of the multi-tenant datacenter as follows

$$C^{dc}(\{s_i\}) = \sum\nolimits_{i \in \mathcal{I}} C_i^{dc}(s_i), \tag{13}$$

C. Backup Generator

As HVAC and datacenter altogether may not shed enough energy as required by EDR, the MUB operator needs to resort to other knobs, typically on-site diesel generator, to make up the energy reduction shortage. Thus, the backup generator cost is expresses as

$$C^{bg}(e_z) = \omega^{bg} \, e_z, \tag{14}$$

where ω^{bg} is the unit cost of backup generator (e.g., diesel price).

D. EDR Problem Formulation

It is critical for the MUB to satisfy the EDR signals without causing much negative impact on the SLA performance of datacenter's jobs as well as the user comfort in the building. Consequently, we consider the MUB's cost minimization problem for EDR as follows

$$\mathbf{P}_{mub}: \min. \quad C^{cf}(\{\bar{T}_n\}) + C^{dc}(\{s_i\}) + C^{bg}(e_z) \tag{15}$$

$$\text{s.t.} \quad e_i = \gamma_i s_i, \ \forall i \in \mathcal{I}, \tag{16}$$

$$e_n = \kappa_n \bar{T}_n, \ \forall n \in \mathcal{N}, \tag{17}$$

$$\sum_{n\in\mathcal{N}} e_n + \sum_{i\in\mathcal{I}} e_i + e_z = Q, \tag{18}$$

$$\text{var.} \quad e_z \geq 0, s_i \geq 0, i \in \mathcal{I}, \bar{T}_n \geq 0, \ \forall n \in \mathcal{N}. \tag{19}$$

In $\mathbf{P}_{mub}$, the objective is to minimize the MUB's total "cost" incurred for shedding energy for EDR. The EDR is reflected in constraint (18) such that the MUB's response is equal to an an power reduction target Q required by the LSE [19]. With thousands of servers in a datacenter, we can further relax the integer $\boldsymbol{s}$ as continuous variables such that this problem is tractable.

We see that the MUB operator needs to solve a joint optimization problem by judicially optimizing server allocation, backup generator, and indoor temperature controlling such that the total cost (which represents the overall negative impact of EDR) is minimized. Furthermore, it is straightforward that $\mathbf{P}_{mub}$ is a convex problem, which can be solved efficiently using interior-point method. However, such centralized approach requires global access to all information of office users as well as datacenter users (tenants), which may not be possible since all users are not willing to share their private data.

In practice, moreover, the three sub-systems logically coupled (due to the shared total energy reduction required for EDR) but are physically self-managed in a separate manner. That is, they have their own controllers that are independent of each other. Therefore, in order to utilize these independent controllers, a distributed implementation is essential. Furthermore, these individual and independent sub-systems will raise the question on whether to participate in this EDR with their own cost.

To address those challenges, we will propose a market-based mechanism to incentivize the MUB's sub-systems to shed energy in a distributed manner such that the solutions of $\mathbf{P}_{mub}$ can be obtained in the next section.

IV. Incentive Mechanism for MUB Energy Shedding

We first propose the EDR mechanism for MUB operator to incentivize the sub-systems:

1) The **MUB operator** receives the EDR signal Q, then determines two components of the mechanism:
   - A *reward rate*, denoted by $\boldsymbol{\delta} := (\{\delta_n\}, \{\delta_i\}, \delta_z)$, is the compensation price paid for every unit of energy shedding by sub-systems
   - A *commitment*, characterized by a vector $\hat{\boldsymbol{e}} = (\{\hat{e}_n\}, \{\hat{e}_i\}, \hat{e}_z)$ such that
   $$\mathbf{1}^T \hat{\boldsymbol{e}} = Q \tag{28}$$
   and a penalty $\rho(\boldsymbol{e} - \hat{\boldsymbol{e}})^2$ for commitment deviation where $\boldsymbol{e} = (\{e_n\}, \{e_i\}, e_z)$ is energy shedding vector of MUB sub-systems.
2) **MUB sub-systems**, given the mechanism package, decide the energy shedding as follows:
   - Each office n, $\forall n \in \mathcal{N}$, solves
   $$\max_{\bar{T}_n \geq 0} \quad \delta_n e_n - C_n^{cf}(\bar{T}_n) - \frac{\rho}{2}(e_n - \hat{e}_n)^2, \tag{29}$$
   $$\text{s.t.} \quad e_n = \kappa_n \bar{T}_n, \tag{30}$$
   - Each tenant i, $\forall i \in \mathcal{I}$, solves
   $$\max_{s_i \geq 0} \quad \delta_i e_i - C_i^{dc}(s_i) - \frac{\rho}{2}(e_i - \hat{e}_i)^2, \tag{31}$$
   $$\text{s.t.} \quad e_i = \gamma_i s_i, \tag{32}$$
   - Backup generator solves
   $$\max_{e_z \geq 0} \quad \delta_z e_z - C^{bg}(e_z) - \frac{\rho}{2}(e_z - \hat{e}_z)^2. \tag{33}$$
3) **Mechanism equilibrium** is a vector $(\boldsymbol{\delta}^*, \hat{\boldsymbol{e}}^*, \boldsymbol{e}^*)$ satisfying (28), (29), (31) and (33) such that $\hat{\boldsymbol{e}}^* = \boldsymbol{e}^*$.

In this mechanism, while the operator designs its reward to incentivize the sub-system follows the commitment that satisfies the EDR, each sub-system individually decides energy shedding to maximize their reward and minimize their costs, including their own cost and the commitment deviation cost.

Even though the rationale behind the mechanism design is intuitive, its efficiency can be raised by many questions: Is there an existence of a mechanism equilibrium? If yes, then is it an $\mathbf{P}_{mub}$'s optimal or sub-optimal solution? And how to implement it distributively?

Algorithm 1 Distributed Algorithm for MUB Energy Shedding Mechanism (DAMESH)

1: $k = 0$, the MUB operator broadcasts a random $(\boldsymbol{\delta}^{(0)}, \hat{\boldsymbol{e}}^{(0)})$ to all sub-system;
2: **repeat**
3: **MUB sub-system:** decides its reduced energy as the following and then submit it to the operator

$$e_n^{(k+1)} = \left[\hat{e}_n^{(k)} + \frac{\delta_n^{(k)}}{\rho} - \frac{\omega^{cf}}{\rho}\frac{k}{\kappa_n}\right]^+, \tag{20}$$

$$e_i^{(k+1)} = \left[\hat{e}_i^{(k)} + \frac{\delta_i^{(k)}}{\rho} - \frac{C_i^{dc\prime}(e_i^{(k+1)}/\gamma_i)}{\gamma_i \rho}\right]^+, \tag{21}$$

$$e_z^{(k+1)} = \left[\hat{e}_z^{(k)} + \frac{\delta_z^{(k)}}{\rho} - \frac{\omega^{bg}}{\rho}\right]^+, \tag{22}$$

$\forall n \in \mathcal{N}, i \in \mathcal{I}$, where $[x]^+ = \max(0, x)$ and

$$C_i^{dc\prime}(s_i) = \omega^{wat} + \omega^{sla}\frac{\lambda_i^2}{((S_i - s_i)\mu_i - \lambda_i)^2} \tag{23}$$

is the first derivative of $C_i^{dc}(s_i)$.
4: **MUB operator:** decides its commitment by solving the following problem

$$\text{min.} \quad \hat{\boldsymbol{e}}^T \boldsymbol{\delta}^{(k)} + \frac{\rho}{2}\|\hat{\boldsymbol{e}} - \boldsymbol{e}^{(k+1)}\|_2^2 \tag{24}$$

$$\text{s.t.} \quad \mathbf{1}^T \hat{\boldsymbol{e}} = Q, \tag{25}$$

$$\hat{\boldsymbol{e}} \geq 0. \tag{26}$$

and updates the reward rate as follows

$$\boldsymbol{\delta}^{(k+1)} = \boldsymbol{\delta}^{(k)} + \rho(\hat{\boldsymbol{e}}^{(k+1)} - \boldsymbol{e}^{(k+1)}), \tag{27}$$

5: $k = k + 1$;
6: **until** $\|\boldsymbol{\delta}^{(k+1)} - \boldsymbol{\delta}^{(k)}\|_2 < \epsilon$.

To answer all the questions, we first propose a distributed algorithm for this mechanism, namely DAMESH, in Alg. 1. In each iteration of DAMESH, each MUB sub-system updates its optimal energy shedding in line 3, whereas the operator decides its reward rate and commitment by solving a strictly convex problem in line 4. Then, we have the following results, which is proved in Appendix A.

Theorem 1. DAMESH *converges to a mechanism equilibrium, which is also a solution of* $\mathbf{P}_{mub}$.

V. Simulation Results

In this section, we provide simulation settings and results for solving EDR MUB's cost minimization.

A. Settings

For the building offices, the temperature outdoor is set to $T^o = 32^o$C. We use the same model of [20] such that $T^c = a_2 + b_2 \cdot T^o$, where a_2 is set to 17.8 and b_2 is set to 0.31. Hence, the comfort temperature T^c is 27.7^oC. The heat gain $G_n(t)$ of each office is set in a range from 1.2 to 1.7, and the its conductivity m_n is in range from 200 to 300. These values are conducted to emulate an approximate number of 100 occupants in an office [20].

For MDCs, tenants have $S_i = 2000$, $\forall i$, and their corresponding service rates μ_i are increased from 2 to 6. We set $p_{i,a}$ and $p_{i,s}$ to 200 and 400W, respectively, for all servers. The PUE is set to 1.2. In addition, the SLA weight ω_{sla} and wear-and-tear weight ω_{wat} of tenant datacenters are set to 5×10^{-3}. In terms of tenant workload, we use the traces of Facebook [25]. Due to unavailability of access to the datacenter information, we randomly collect contiguous 20 time slots from Facebook trace for each tenant workload. During the considered time slots, the workloads are normalized with respect to the total maximum capacity of the tenants.

Fore the remaining parameters, the backup unit cost is set comparably to a diesel price where $\omega^{bg} = 0.3$ \$/kW, $\rho = 0.03$, and $Q = 1$ MWh.

B. Results

In order to illustrate the efficiency of our mechanism, we compare DAMESH with the optimal value and the following two baselines:

1) Baseline 1 only relies on energy shedding of the multi-tenant datacenter control and on-site backup generator for the EDR. That is, the constraint (18) is changed to $\sum_{i\in\mathcal{I}} e_i + e_z = Q$, and $T_n = 0$, $\forall n$.
2) Baseline 2 only uses the HVAC control and the backup generator for the EDR, i.e., the constraint (18) is changed to $\sum_{n\in\mathcal{N}} e_n + e_z = Q$ and $s_i = S_i$, $\forall i$.

Convergence. To illustrate to convergence of DAMESH, we first consider a small MUB system with three offices and three MDCs in Fig. 1a. Comparing with other baselines, with a given stopping condition, we see that Fig. 1a shows the smallest gap between DAMESH and the optimal value. The detailed energy shedding convergence of each sub-system is depicted in Fig. 1b.

MUB cost comparison. We next compare the MUB cost performance between DAMESH and other baselines with the same workload trace and an MUB setting with ten offices and five MDCs in Fig. 2. In this figure, we see that DAMESH has the smallest MUB cost because DAMESH has more control knobs to balance the individual sub-system cost. On the other hand, while the MUB cost of Baseline 1 fluctuates with tenant workload pattern and is lower than that of Baseline 2, which is mainly contributed by the backup generator cost (inducing higher cost of Baseline 1) since there is less energy shedding from offices due to limited temperature control.

VI. Conclusions

In this paper, we study MUB's energy shedding for EDR. In view that the existing studies on demand response by buildings and datacenters have been largely isolated to date and resulted in uncoordinated energy management in MUBs, we propose a *first-of-its-kind* coordinated and incentive-aligned approach to enabling EDR in MUBs, such that the total incurred loss (i.e., latency performance degradation for datacenter, thermal discomfort for office, and diesel generation) is minimized for energy shedding during EDR. We also propose a distributed algorithm that can optimally control HVAC temperature for non-datacenter space, server provisioning and load balancing in datacenter, and usage of diesel generation. We also conduct a case study to validate our coordinated approach, which outperform other un-coordinated approaches in terms of total incurred cost.

Acknowledgment

This work was supported in part by the U.S. NSF under grants CNS-1551661 (CAREER) and CNS-1565474. This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (R0126-15-1009, Development of Smart Mediator for Mashup Service and Information Sharing among ICBMS Platform). Dr. CS Hong is the corresponding author.

Appendix A
Proof of Theorem 1

We first define

$$f(\boldsymbol{e}) = \sum_{n\in\mathcal{N}} C_n^{cf}(e_n/\kappa_n) + \sum_{i\in\mathcal{I}} C_i^{dc}(e_i/\gamma_i) + C^{bg}(e_z). \quad (34)$$

Then $\mathbf{P}_{mub}$ can be presented in vector form as follows

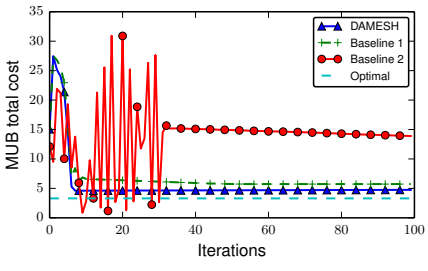
$$\text{min.} \quad f(\boldsymbol{e}) \quad (35)$$

$$\text{s.t.} \quad \mathbf{1}^T\boldsymbol{e} = Q, \quad (36)$$

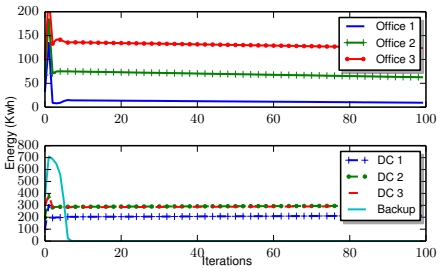
$$\boldsymbol{e} \geq 0. \quad (37)$$

Defining a set $\mathcal{E} = \{\boldsymbol{e} : \boldsymbol{e} \geq 0, \mathbf{1}^T\boldsymbol{e} = Q\}$ and a following indicator function

$$I_{\mathcal{E}}(\boldsymbol{e}) = \begin{cases} 0, & \boldsymbol{e} \in \mathcal{E}; \\ \infty, & \text{otherwise}, \end{cases} \quad (38)$$



(a) Convergence of MUB cost.



(b) Convergence of each sub-system.

Fig. 1: Convergence of DAMESH.

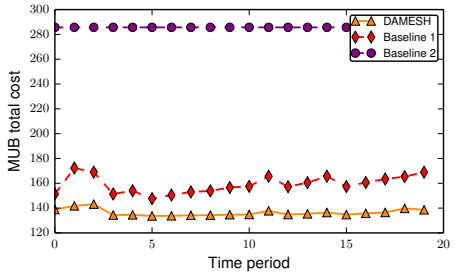


Fig. 2: Comparison of total cost between DAMESH, Baseline 1 and Baseline 2.

then we can rewrite $\mathbf{P}_{mub}$ as

$$\text{min.} \qquad f(\boldsymbol{e}) + I_{\mathcal{E}}(\boldsymbol{e}) \tag{39}$$

By introducing auxiliary variable vector $\hat{\boldsymbol{e}}$, we have an equivalent problem of $\mathbf{P}_{mub}$ as follows

$$\mathbf{P}'_{mub}: \text{min.} \qquad f(\boldsymbol{e}) + I_{\mathcal{E}}(\hat{\boldsymbol{e}}) \tag{40}$$

$$\text{s.t.} \qquad \boldsymbol{e} = \hat{\boldsymbol{e}}, \tag{41}$$

with its augmented Lagrangian

$$L_\rho(\boldsymbol{e}, \hat{\boldsymbol{e}}, \boldsymbol{\delta}) = f(\boldsymbol{e}) + I_{\mathcal{E}}(\hat{\boldsymbol{e}}) + \boldsymbol{\delta}^T(\hat{\boldsymbol{e}} - \boldsymbol{e}) + \frac{\rho}{2}\|\hat{\boldsymbol{e}} - \boldsymbol{e}\|_2^2,$$

where ρ is a chosen parameter. Next, we obtain the following sequential decomposable updates

$$\boldsymbol{e}^{(k+1)} = \arg \min_{0 \le \boldsymbol{e} \le \boldsymbol{e}^{max}} \left(f(\boldsymbol{e}) - \boldsymbol{e}^T \boldsymbol{\delta}^{(k)} + \frac{\rho}{2}\|\boldsymbol{e} - \hat{\boldsymbol{e}}^{(k)}\|_2^2 \right), \tag{42}$$

$$\hat{\boldsymbol{e}}^{(k+1)} = \arg \min \left(I_{\mathcal{E}}(\hat{\boldsymbol{e}}) + \hat{\boldsymbol{e}}^T \boldsymbol{\delta}^{(k)} + \frac{\rho}{2}\|\hat{\boldsymbol{e}} - \boldsymbol{e}^{(k+1)}\|_2^2 \right), \tag{43}$$

$$\boldsymbol{\delta}^{(k+1)} = \boldsymbol{\delta}^{(k)} + \rho(\hat{\boldsymbol{e}}^{(k+1)} - \boldsymbol{e}^{(k+1)}). \tag{44}$$

The first update (42) is the solution of the MUB sub-systems problems (29), (31), and (33). Since all of these problems are convex, using the first-order condition, we can solve them to have the results (20), (21), and (22) of DAMESH. The second update (43) is the solution of the equivalent problem in line 4 of DAMESH. Finally, because $\mathbf{P}_{mub}$ is a convex problem, the ADMM method can guarantee the convergence [26].

REFERENCES

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

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

[3] F. Jazizadeh, A. Ghahramani, B. Becerik-Gerber, T. Kichkaylo, and M. Orosz, "Human-Building Interaction Framework for Personalized Thermal Comfort-Driven Systems in Office Buildings," *Journal of Computing in Civil Engineering*, vol. 28, no. 1, pp. 2–16, jan 2014.

[4] A. Wierman, Z. Liu, and H. Mohsenian-Rad, "Opportunities and challenges for data center demand response," in *IEEE IGCC*, Dallas, TX, jun 2014.

[5] The Green Grid, "Pue: A comprehensive examination of the metric," 2012.

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

[7] NRDC, "Scaling up energy efficiency across the data center industry: Evaluating key drivers and barriers," in *Issue Paper*, 2014.

[8] S. Ren and M. A. Islam, "Colocation demand response: Why do I turn off my servers?" in *USENIX ICAC*, Philadelphia, PA, jun 2014, pp. 201–208.

[9] N. H. Tran, C. T. Do, S. Ren, Z. Han, and C. S. Hong, "Incentive Mechanisms for Economic and Emergency Demand Responses of Colocation Datacenters," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 12, pp. 2892–2905, dec 2015.

[10] T. Wei, T. Kim, S. Park, Q. Zhu, S. X.-D. Tan, N. Chang, S. Ula, and M. Maasoumy, "Battery management and application for energy-efficient buildings," in *DAC*, 2014.

[11] Y. Agarwal, T. Weng, and R. K. Gupta, "The energy dashboard: Improving the visibility of energy consumption at a campus-wide scale," in *BuildSys*, 2009.

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

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

[14] S. K. Aggarwal, L. M. Saini, and A. Kumar, "Electricity price forecasting in deregulated markets: A review and evaluation," *International Journal of Electrical Power & Energy Systems*, vol. 31, no. 1, pp. 13–22, 2009.

[15] EnerNOC, "Demand response: A multi-purpose resource for utilities and grid operations," 2009. [Online]. Available: http://www.enernoc.com/our-resources/white-papers/demand-response-a-multi-purpose-resource-for-utilities-and-grid-operators

[16] PJM, "Emergency demand response (load management) performance report-2012/2013."

[17] M. Ghamkhari and H. Mohsenian-Rad, "Profit maximization and power management of green data centers supporting multiple slas," in *IEEE ICNC*, 2013, pp. 465–469.

[18] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. H. Andrew, "Greening geographical load balancing," in *Proc. ACM SIGMETRICS*, San Jose, CA, 2011, pp. 233–244.

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

[20] A. H.-y. Lam, Y. Yuan, and D. Wang, "An occupant-participatory approach for thermal comfort enhancement and energy conservation in buildings," in *Proc. ACM e-Energy*, New York, New York, USA, jun 2014, pp. 133–143.

[21] ASHRAE, "ASHRAE standard 55-2010:Thermal Environmental Conditions for Human Occupancy," 2010.

[22] M. Lin, A. Wierman, L. L. H. Andrew, and E. Thereska, "Dynamic Right-sizing for Power-proportional Data Centers," in *Proceedings IEEE INFOCOM*, Shanghai, China, apr 2011.

[23] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. H. Andrew, "Geographical load balancing with renewables," *ACM SIGMETRICS Performance Evaluation Review*, vol. 39, no. 3, pp. 62–66, dec 2011.

[24] A. Gandhi, M. Harchol-Balter, R. Das, and C. Lefurgy, "Optimal power allocation in server farms," *ACM SIGMETRICS Performance Evaluation Review*, vol. 37, no. 1, pp. 157–168, 2009.

[25] "Facebook Dashboard: PUE & WUE." [Online]. Available: https://PrinevilleDataCenter/app_399244020173259

[26] S. Boyd, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, jan 2010.