Aggregating a Large Number of Residential Appliances for Demand Response Applications

Fadi Elghitani, and Weihua Zhuang, *Fellow, IEEE*

***Abstract*—Current demand response (DR) programs focus on industrial consumers as they can provide a large magnitude of demand modification. In order to extend DR programs to the residential sector, aggregating service demands from a large number of residential consumers is necessary in order to achieve a sensible benefit to the power network. In this paper, we propose a methodology for residential demand aggregation, based on a multi-class queueing system. Each class represents demand blocks of a specific power level, time duration, and a service delay requirement. We use this model to minimize the cost of the appliances' aggregated power consumption under day-ahead pricing (DAP). Using realistic appliances' data, we show that the proposed framework achieves a cost reduction that is close to the best achievable one.**

***Index Terms*—Demand response, Residential appliances, Aggregation Model, Cutting-plane method.**

I. Introduction

THE operation of electric power systems is traditionally based on demand following. The role of supply-demand matching of the electric energy is assigned to the generation-side of a power network, while the demand-side is assumed to be non-controllable and has to be satisfied regardless of its cost. Demand following can be very expensive, especially during periods of peak demand, when the least efficient generators have to be activated to satisfy the increase in demand [1]. Therefore, utility companies began to realize the potential benefits of controlling consumers' demands, referred to as demand-side management (DSM), which contrasts traditional generation-side decisions. A specific type of DSM is demand response (DR), which seeks demand modification via different financial incentives for the consumers.

The residential sector represents a significant part in electric energy demand (e.g., 32.9% of total demand in Ontario [2]). As a result, it is desired to incorporate the residential sector to various DR applications. Residential consumers can contribute to DR via two categories of appliances: deferrable loads and thermostatically-controlled loads (TCLs). Deferrable loads can be controlled to defer their energy consumption to a future time, but should be satisfied within a certain deadline to avoid consumer's inconvenience. Examples of deferrable loads include washing machines and dish washers. On the other hand, TCLs represent appliances which control the temperature of some compartments. TCL examples include refrigerators, air-conditioners, and electric water heaters.

Many studies on residential DR are related to Home Energy Management (HEM) applications, where appliances within a single household are scheduled to minimize the resident's electricity bill [3]–[8]. Since on average a residential consumer has a relatively few number of controllable appliances, individual demand models for appliances can be used in the optimization problem. However, applying algorithms developed for HEM to schedule the appliances of several consumers is difficult due to increased computational complexity. Hence, aggregation demand models are needed to represent the demands of different groups of appliances, in order to reduce the DR problem size. In existing studies, separate aggregation models are used for both deferrable and TCL appliances. For deferrable appliances, aggregation model is based on only the energy requirements of different appliances, without taking account of power consumption profiles [9]–[14]. The aggregation is done by simply summing up all the energy requirements from all appliances. The limitations of this model come from two assumptions: 1) all appliances require the same class of delay performance, which is not accurate due to the heterogeneity of the appliance types and consumer preferences; and 2) the demand is perfectly elastic, which can be satisfied within a small time or over a long time as desired. The model is not suitable for scheduling TCLs as there is no information about the controlled temperature. On the other hand, aggregation models for TCLs in existing studies are based on classifying TCLs according to their current temperature and its (ON/OFF) operation status [15]–[17]. The system state is represented by the probability mass function (PMF) over these different bins. An advantage is that the accuracy of the model depends on the resolution of the PMF, but not on the number of aggregated TCLs. The control signal represents the switching probabilities (turning ON or OFF) that are submitted to each group of TCLs residing in the same bin. Since each bin represents a state variable, the system state-space can be huge for a given temperature range. Therefore, it is difficult to use the aforementioned aggregation model for optimization. Nevertheless, the model is used for evaluating the performance of a large number of TCLs.

Different from the existing studies, our proposed methodology works for both deferrable and TCL appliances. Demands from a heterogeneous population of appliances is aggregated into a set of pre-specified classes of demand. This multi-class approach allows an efficient utilization of DR resources while ensuring consumers' convenience, in contrast to restricting demand aggregation of controllable appliances into a single class. Also, demands can be differentiated according to their energy consumption durations, which allows us to deduce with high accuracy the controllable power consumption profile resulting from DR decisions.

The authors are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. (e-mail: felghita@uwaterloo.ca; wzhuang@uwaterloo.ca)

In this paper, we focus on the optimal scheduling of a large number of residential appliances. Our contributions can be summarized as follows: 1) We present a demand aggregation methodology for residential appliances, which overcomes the aforementioned limitations. The model balances the tradeoff between accuracy and computational complexity. The methodology is based on introducing new units of electric energy demand, referred to as "demand blocks". Each demand block is characterized by three parameters: power level, energy consumption duration, and a delay requirement for demand satisfaction, i.e. a constraint on the waiting time of a demand block before being scheduled. We only allow a limited set of types (classes) of demand blocks. The demand requirement of each controllable appliance is approximately represented by a demand block of the closest characteristics; 2) We apply the methodology to minimize the energy consumption cost under day-ahead pricing. The DR operator will have to decide whether to schedule a demand block immediately, or to defer the scheduling to future times. The system behavior resembles that of a non-stationary, multi-class queueing system. We develop an optimization algorithm based on the cutting-plane method to achieve the optimal demand block scheduling.

The remainder of this paper is organized as follows. We describe the system model in Section II. Our proposed methodology and problem formulation are presented in Section III. Numerical results are given in Section IV to demonstrate the performance of our proposed methodology. Finally, Section V concludes this research.

II. System Model

Consider a residential community as shown in Fig. 1. The energy need is supplied by a retailer who purchases electric energy from the wholesale market according to day-ahead pricing, where the price profile for one day is given in advance to the retailer. The price is updated every fixed time period Δ. The retailer seeks to maximize his profit by minimizing the cost of the purchased energy. To achieve this goal, the retailer is allowed to directly control some of the appliances of residential consumers, who will be compensated using suitable financial incentives. The communication between the retailer and the corresponding consumers is established using a device referred to as "demand controller" in the retailer side and using a smart meter (SM) in the consumer side. The communication is two-ways: the SM sends power consumption information and the demand requests of the controllable appliances to the demand controller, which in turn sends control decisions back to the SMs. The SMs resend control decisions to a unit referred to as Energy Management Center (EMC), which is responsible for appliance scheduling. For simplicity (and in consistence with existing studies in the area), we model our system as a discrete time system, where information is gathered and decisions are made in pre-specified time slots corresponding to price update times. The scheduling time horizon is finite, and consequently the set of time slots is also finite, denoted by $\mathcal{T} = \{1, 2, \ldots, T\}$.

Each consumer has three types of appliances: deferrable appliances, TCLs, and non-controllable loads. We follow an approach for residential demand modeling similar to that in [18]. The general behavior of a consumer can be modeled as a Markov chain whose states represent different consumer's activity level. We consider three levels of activities: absent, inactive, and active. The activity level influence different appliances as follows.

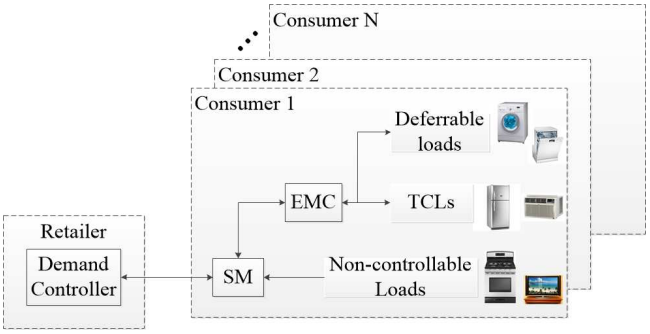


Fig. 1: System model under consideration

1) Deferrable loads: Each consumer, i, has a set $\mathcal{N}_i$ of deferrable appliances. Each appliance $j \in \{\mathcal{N}_i\}$ is characterized by an arbitrary task pattern, $g_{ij}(t)$, a probability to be turned on, p_{ij}, and a required deadline for the task completion, d_{ij}. Task requests arrive only when the consumer is in the active state.
2) TCLs: Each consumer, i, has a set $\mathcal{C}_i$ of TCL appliances. Each appliance follows a simple equivalent thermal parameter (ETP) model, where the controlled temperature evolves as follows [19]

$$\theta(t) = \theta_e + s(t)P_{TCL}R_{th} - (\theta_e + s(t)P_{TCL}R_{th} - \theta_i)e^{-t/R_{th}C_{th}}. \quad (1)$$

In (1), $\theta(t)$ is the temperature under control at time t, θ_e is the external temperature, R_{th} and C_{th} represent the thermal resistance and capacitance respectively, $s(t)$ is the switching function which takes either 0 or 1, and P_{TCL} is the power consumed by the TCL appliance when it is turned on. Different TCLs can have different physical and setting parameters. The TCLs' operation is assumed to be independent of the consumer state.
3) Uncontrollable loads: Each consumer, i, has a set $\mathcal{U}_i$ of uncontrollable appliances. Each appliance is modeled as an (on/off) Markov chain whose transition probabilities depend on the consumer state. If the consumer is in the absent state, the appliance is always turned off. Each appliance $j \in \mathcal{U}_i$ consumes constant power as long as it is turned on, while it does not consume any energy when turned off.

III. Problem Formulation

The demand controller has to schedule a large number of appliances in real time, such that the cost of aggregated power consumption is minimized. Consumer's convenience should be guaranteed to a high level, to ensure the continuity of consumers' DR participation. Consumer's convenience is achieved when deferrable loads are scheduled to consume their required energy before the deadline and the temperatures of the

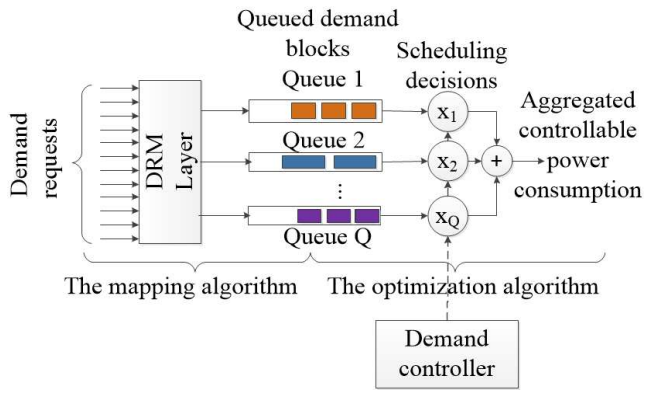


Fig. 2: The proposed demand aggregation methodology.

TCL appliances are kept within their deadbands. In addition to these constraints, the total power consumed should not exceed a certain threshold due to power network capacity limitations.

As discussed in Section I, it is difficult to use individual demand models when the number of appliances to be scheduled is large. We propose a novel demand aggregation methodology for residential appliances, which can be used for solving the aforementioned DR problem. First, we introduce a layer that acts as an interface between different consumer appliances and the demand controller, referred to as the DR management (DRM) layer. The input to the DRM layer is the demand requirements of deferrable and TCL appliances. The output of this layer is referred to as "demand blocks", new units of demand information. Each demand block ideally has a constant power consumption over a fixed duration, represented by a rectangular-shaped demand profile. A demand block can result from the demand requirement of a single appliance or from a mixture of demand requirements of several appliances. The responsibility for the generation of demand blocks is assigned to the EMC located in the consumer premises. Demand blocks are classified into different classes according to three parameters: demand block duration, power consumption, and delay requirement. When demand blocks are generated, they are placed into different queues according to their classes, where each queue represents a distinct class of demand blocks, as shown in Fig. 2. The set of all possible classes $\mathcal{Q} = \{1, 2, \ldots, Q\}$ is defined in advance, and is the same for all residential consumers.

The idea behind our approach is not to focus on an appliance identity, but on its DR capability. We are interested in how it can provide benefits to the power system. This DR capability is expressed through the three aforementioned parameters (energy consumption duration, power level, and delay requirements). If two appliances have the same values for these three parameters, they are considered as of the same type (class) from the demand controller's point of view, regardless if they have different functions or if they belong to different consumers. The demand aggregation is done by grouping appliances of closer DR capabilities together (assign them to the same queue) in order to reduce DR complexity.

The DRM layer divides the DR problem into two parts: 1) how a demand requirement of an appliance is assigned to one queue of the given queue set, and 2) determining the optimal scheduling of queued demand blocks such that the total energy consumption cost is minimized. Therefore, two algorithms are needed: the mapping algorithm for the first part and the optimization algorithm for the second part. The mapping algorithm is run by the EMC unit within each consumer premises, while the optimization algorithm is run by the demand controller.

A. The Mapping Algorithm

The goal of the mapping algorithm is to assign different appliance demand requirements to different given classes of demand. The mapping is done such that the total distortion resulting from the assignment process is minimized. The distortion resulting from assigning demand requirement profile $g(t)$ to class i is defined as

$$\zeta_i(g(t)) = \frac{1}{P_i^2 K_i} \int_0^\infty [g(t) - P_i[U(K_i) - U(0)]]^2 dt \quad (2)$$

where P_i and K_i is the power level and duration of a demand block of class i, and $U(.)$ is the unit step function. The distortion represents the amount of deviation of the task profile from the rectangular shape of an ideal demand block in the assigned queue. Equation (2) is applied only to the subset of classes whose deadline is lower than or equal to that of the appliance deadline. If this subset is empty, the appliance demand is treated as a non-controllable demand.

Knowing the demand requirement of a deferrable appliance is straight-forward, as the task pattern is given in advance. However, in TCL appliances, there is no predefined task pattern. Furthermore, the consumer's convenience from TCL appliances is defined by temperature constraints instead of delay constraints. Therefore, we need to deduce a method for representing demand requirements of TCL appliances similar to that of deferrable appliances. Our proposed method is described in the following.

To achieve consumer's convenience, the controlled temperature should be kept within the deadband:

$$\theta_{min} \leqslant \theta \leqslant \theta_{max} \quad (3)$$

where θ_{min} and θ_{max} represent the minimum and maximum temperatures beyond which the operation of the TCL becomes inconvenient for the consumer. Let the appliance send a request for energy demand whenever the temperature falls below a certain threshold $\theta_s \in [\theta_{min}, \theta_{max}]$, which can be, for example, the temperature setting point. The appliance requires a non-interruptible power consumption for a duration K and a deadline D for satisfying the demand. Parameters K and D are selected such that the TCL's temperature stays within the deadband between two successive requests, regardless of demand controller's decision. From the DR perspective, it is desirable for an appliance to have a short power consumption duration and a long deadline for demand satisfaction. However, as we shall see, there is a trade-off between both quantities.

After the demand request is sent, the appliance is not turned-on until it receives a permission from the EMC. The maximum

time that the appliance can wait without causing a discomfort for the consumer is the period that θ drops from θ_s to θ_{min}. Therefore, the maximum achievable deadline D_{max} can be obtained from (1) as

$$D_{max} = R_{th}C_{th}\ \ln\Big(\frac{\theta_s - \theta_e}{\theta_{min} - \theta_e}\Big). \tag{4}$$

Since the TCL appliance does not have a prior information when it will receive the permission for power consumption, the duration of the power consumption, K, should be specified under a worst-case condition. The TCL power consumption duration should be sufficiently long for the controlled temperature to rise above θ_s. The worst-case happens when the appliance receives the permission exactly at the end of the deadline, when θ is at its minimum value. Therefore, the minimum power consumption duration, K_{min}, is determined as the time required for θ to rise from the aforementioned minimum value to θ_s. The duration K_{min} can be derived from (1), given by

$$K_{min} = R_{th}C_{th}\ \ln\left[\frac{P_{TCL}R_{th} - (\theta_s - \theta_e)e^{-D/R_{th}C_{th}}}{P_{TCL}R_{th} - (\theta_s - \theta_e)}\right]. \tag{5}$$

With the TCL demand representation, the input to the mapping algorithm, at each time slot, is a set of tasks, $\mathcal{Y}$, representing different demand requirements from deferrable and TCL appliances. Each task $i \in \mathcal{Y}$ is characterized by a power consumption pattern $g_i(t)$ and a deadline D_i. Considering that a residential consumer usually has a small number of appliances contributing to DR, which do not have to request energy demand at the same time, we simply use enumeration for the best assignment of these tasks to available classes of demands.

B. The Optimization Algorithm

The goal of the optimization algorithm is to schedule different demand blocks such that the total consumer energy consumption cost is minimized. The problem has two types of constraints: 1) power capacity constraints due to the limited power rating of the distribution transformer, and 2) the delay QoS constraints for different types of demand blocks. The power capacity constraints can be written as

$$L_t^c \leqslant L_t^{max}, \qquad \forall t \in \mathcal{T} \tag{6}$$

where L_t^c is the total controllable energy consumption of residential consumers in time slot t. For convenience, we simply refer to the energy consumption in a unit time slot as the power consumption. Therefore, L_t^c also represents the total controllable consumer power consumption at time slot t. Similarly, L_t^{max} represents the difference between the power rating of the distribution transformer and the total non-controllable power consumption.

On the other hand, the delay QoS constraints are given in terms of the probability that the demand blocks are scheduled before their deadlines. The constraints are global, i.e., not associated with demand blocks generated at a specific time period. The QoS constraints are given by

$$\Pr\{W_i \leqslant D_i\} \geqslant \delta, \qquad \forall i \in \mathcal{Q} \tag{7}$$

where W_i and D_i represent respectively the waiting time and the deadline of a demand block of class i, and δ represents the minimum level of QoS. From a practical point of view, the value of δ should be defined in advance in the contract between the retailer and its consumers.

The only source of randomness in the problem is the arrival process of different demand blocks, which evolves according to some general and non-stationary stochastic process. All other parameters are deterministic from the controller's viewpoint. Therefore, the system under consideration can be described as a multi-class $G_t/D/m_t$ queueing system, where G_t refers to a general, non-stationary arrival process of demand blocks, D refers to the deterministic duration for a demand block, whereas m_t denotes a time-varying capacity of the number of demand blocks that can be scheduled. Our proposed optimization algorithm works by dividing the decision-making process into two phases: 1) capacity planning, and 2) real-time scheduling. In the first phase, we set a limit for power consumption profiles for each queue such that all constraints are satisfied. In the second phase, demand blocks are scheduled in a greedy way regardless of the cost of energy consumption, but they should not exceed power consumption profiles defined in the first phase.

Let x_{it} denote the maximum number of demand blocks of class i that can be scheduled at time slot t, and $\mathbf{x} = \{x_{it} : i \in \mathcal{Q}, \quad t \in \mathcal{T}\}$ denote the vector of allocated capacity of demand blocks for each class at each time slot. We denote the cost of scheduling a demand block of class i at time t by c_{it}, given by

$$c_{it} = P_i \sum_{s=t}^{t+K_i-1} \gamma(s) \tag{8}$$

where $\gamma(s)$ is the given electricity price at time slot s. The capacity allocation vector, $\mathbf{x}$, is the decision variable for the first phase. Since arrivals of the demand blocks are random, there is no single choice of $\mathbf{x}$ that minimizes the energy consumption cost under all possible demand arrival scenarios. Instead, we seek to minimize the upper bound of the energy consumption cost which happens if the allocated capacity $\mathbf{x}$ is fully utilized. The optimization problem is given by

$$\begin{aligned} \mathbf{P1}: \quad & \min_{x} \left\{\sum_i \sum_t c_{it}x_{it}\right\} \\ & \text{s.t. } \mathbf{Ax} \leqslant \mathbf{L^{max}} \\ & \quad h_i(\mathbf{x}) \geqslant 0 \qquad \forall i \in \mathcal{Q} \\ & \quad \mathbf{x} \in \mathbb{N}^{T\times Q} \end{aligned} \tag{9}$$

The first set of constraints represents the power capacity constraints in (6), where $\mathbf{A}$ is the matrix that transforms the capacity allocation vector $\mathbf{x}$ into the total power consumption profile. The second set of constraints represents the QoS constraints in (7), where $h_i(\mathbf{x}) = \Pr\{W_i \leqslant D_i\} - \delta$, representing the difference between the actual QoS level and the target QoS level.

The difficulty of solving **P1** lies on how to evaluate $h_i(\mathbf{x})$ over its domain. We adopt an algorithm used for non-linear programming, called *cutting plane method* (CPM) [20], [21]. The approach is based on an assumption that $h_i(\mathbf{x})$ is a concave function. The assumption is justified in our problem as follows. It is intuitive that the delay performance improves as we add more power (demand block) capacity. Therefore, $h_i(\mathbf{x})$ is a non-decreasing function. However, the maximum possible value for $h_i(\mathbf{x})$ cannot exceed $1-\delta$, which means that the function saturates at sufficiently high values of $\mathbf{x}$. Thus, $h_i(\mathbf{x})$ should be concave at least at high QoS levels, which are the main region of interest.

The CPM principle is close to that of the popular branch-and-bound method. The original problem **P1** is relaxed by removing the second set of constraints, $h_i(\mathbf{x}) \geqslant 0$. Thus, **P1** is transformed into a pure integer programming (IP) problem, that can be solved by its regular techniques. The relaxed problem is

$$\begin{aligned} \textbf{P2}: \quad & \min_{x} \left\{ \sum_i \sum_t c_{it} x_{it} \right\} \\ & \text{s.t. } \mathbf{Ax} \leqslant \mathbf{L^{max}} \\ & \mathbf{x} \in \mathbb{N}^{T \times Q} \end{aligned} \tag{10}$$

After obtaining the optimal solution of **P2**, $h_i(\mathbf{x})$ is evaluated at that particular solution, $\mathbf{x}_r^*$, using a simulation or analytical method. If all the constraints are satisfied, $h_i(\mathbf{x}_r^*) \geqslant 0, \forall i \in \mathcal{Q}$, the relaxed solution is the optimal solution of the original problem ($\mathbf{x}^* = \mathbf{x}_r^*$). If one or more constraints are violated, then for each violated constraint a linear constraint is created such that the "failed" relaxed solution is removed *without* changing the feasible region of **P1**. However, adding these constraints will reduce the feasible region of the relaxed problem, **P2**. The new solution of the relaxed problem will again be used to test if the constraints are satisfied or violated. The process is repeated and, with each iteration, the feasible region of the relaxed problem shrinks until a relaxed solution is reached that satisfies all the constraints of **P1**. The benefit of the approach is that we do not need to characterize $h_i(\mathbf{x})$ over its domain, but rather evaluate it at a specific finite number of points.

Next, we describe how linear constraints are added whenever one or more of the QoS constraints are violated. Since $h_i(\mathbf{x})$ is assumed concave, then:

$$h_i(\mathbf{x}) - h_i(\mathbf{x}_r) \leqslant \mathbf{q}_i(\mathbf{x}_r) \cdot (\mathbf{x} - \mathbf{x}_r) \tag{11}$$

where $\mathbf{q}_i(\mathbf{x}_r)$ is the gradient of $h_i(\mathbf{x})$ at point $\mathbf{x}_r$. If we add the following linear constraint to the relaxed problem:

$$h_i(\mathbf{x}_r) + \mathbf{q}_i(\mathbf{x}_r) \cdot (\mathbf{x} - \mathbf{x}_r) \geqslant 0 \tag{12}$$

then we guarantee that the infeasible solution $\mathbf{x}_r$ is excluded, since $h_i(\mathbf{x}_r) < 0$. Also, this constraint will not affect the feasible region of the original problem **P1**, because at any point $\mathbf{x} : h_i(\mathbf{x}) \geqslant 0$, the constraint is not violated (using inequality (11)). Therefore, the constraint defined in 12 can be used to cut the feasible region of the relaxed problem **P2**.

Before evaluating the QoS performance with either analytical or simulation methods, the rules for real-time scheduling of different demand blocks must be defined. In the capacity planning phase, we define the allocated capacity of demand blocks for each class at each time slot. However, the allocated capacity for some classes may not be fully utilized in real-time. Therefore, we need to develop a rule for utilizing the leftover capacity to serve other classes of demand blocks. First, the duration of the accepted demand blocks should not exceed the duration of the allocated capacity. Given this condition, we use the following heuristic. We prioritize demand blocks according to the shortest deadline. If two demand blocks have the same deadline, they are prioritized according to the longer duration.

Sample Average Approximation

It is difficult to evaluate the performance of a multi-class, non-stationary queueing system using analytical methods. Therefore, we use simulation to generate samples of demand block arrivals and use the sample average approximation (SAA) to approximate the values of $h_i(\mathbf{x})$. For each simulation run (day) j, a total number, N_j^i, of demand blocks of type i are generated. Let S_j^i denote the fraction of N_j^i that are satisfied within their deadlines. Both S^i and N^i are random variables. The approximation is done by replacing $h_i(\mathbf{x})$ by $\bar{h_i}(\mathbf{x}, m)$ as follows:

$$h_i(\mathbf{x}) \approx \frac{\sum_{j=1}^{m} S_j^i}{\sum_{j=1}^{m} N_j^i} - \delta = \bar{h_i}(\mathbf{x}, m) \tag{13}$$

Notice that $\bar{h_i}(\mathbf{x}, m)$ is also a random variable. The optimization problem **P1** will be updated to **P3** as follows:

$$\begin{aligned} \textbf{P3}: \quad & \min_{x} \left\{ \sum_i \sum_t c_{it} x_{it} \right\} \\ & \text{s.t. } \mathbf{Ax} \leqslant \mathbf{L^{max}} \\ & \quad \bar{h_i}(\mathbf{x}, m) \geqslant 0 \qquad \forall i \in \mathcal{Q} \\ & \quad \mathbf{x} \in \mathbb{N}^{T \times Q} \end{aligned} \tag{14}$$

Let X^* and X_m^* denote the set of optimal solutions to **P1** and **P3** respectively. Assume $\mathbf{x_m} \in X_m^*$. $\mathbf{x_m}$ will have the following properties: 1) it satisfies the original linear set of constraints (power capacity constraints), 2) it is "possible" that $\mathbf{x_m}$ violates the QoS constraints, i.e. $\exists i \in \mathcal{Q} : h_i(\mathbf{x_m}) < 0$, and 3) it is "possible" that $\mathbf{x_m}$ to be a suboptimal solution. For SAA to be justified, we need to prove that $X_m^* \subseteq X^*$ almost surely as $m \to \infty$. Our proof is composed of the following steps.

Lemma 1.

$$\bar{h_i}(\mathbf{x}, \infty) = h_i(\mathbf{x}), \ \textit{almost surely} \tag{15}$$

Proof. Since both sequences S_j^i and N_j^i where index $j = 1, 2, \ldots, m$ are i.i.d., then using SLLN:

$$\lim_{m\to\infty}\frac{1}{m}\sum_{j=1}^{m}S_j^i=\mathbb{E}[S^i],\ \text{almost surely}$$
$$\lim_{m\to\infty}\frac{1}{m}\sum_{j=1}^{m}N_j^i=\mathbb{E}[N^i],\ \text{almost surely} \tag{16}$$

From the definition of S^i, it can be written as

$$S^i=\sum_{k=1}^{N^i}\mathbb{1}_{\{W_i\leqslant D_i\}} \tag{17}$$

where $\mathbb{1}_{\{.\}}$ is the indicator random variable which takes the value of one if the condition between the braces is satisfied and zero otherwise. The expectation of S^i can be given by:

$$\begin{aligned}\mathbb{E}[S^i]&=\sum_{n^i=1}^{\infty}\mathbb{E}\left\{\sum_{k=1}^{N^i}\mathbb{1}_{\{W_i\leqslant D_i\}}|N^i{=}n^i\right\}\mathbf{Pr}(N^i{=}n^i)\\&=\sum_{n^i=1}^{\infty}n^i\mathbf{Pr}(W_i\leqslant D_i)\mathbf{Pr}(N^i{=}n^i)\\&=\mathbf{Pr}(W_i\leqslant D_i)\mathbb{E}[N^i]\\&=[h_i(\mathbf{x})+\delta]\,\mathbb{E}[N^i]\end{aligned} \tag{18}$$

We attempt to find the probability that $\bar{h_i}(\mathbf{x},m)=h_i(\mathbf{x})$ as m goes to infinity:

$$\begin{aligned}&\mathbf{Pr}\left\{\bar{h_i}(\mathbf{x},m)=h_i(\mathbf{x})\right\}\\&=\mathbf{Pr}\left\{\frac{\sum_{j=1}^{m}S_j^i}{\sum_{j=1}^{m}N_j^i}{=}\frac{\mathbb{E}[S^i]}{\mathbb{E}[N^i]}\right\}\geqslant\\&\mathbf{Pr}\left\{\sum_{j=1}^{m}S_j^i{=}\mathbb{E}[S^i]\ \text{and}\ \sum_{j=1}^{m}N_j^i{=}\mathbb{E}[N^i]\right\}\\&=\mathbf{Pr}\left\{\sum_{j=1}^{m}S_j^i{=}\mathbb{E}[S^i]\right\}+\mathbf{Pr}\left\{\sum_{j=1}^{m}N_j^i{=}\mathbb{E}[N^i]\right\}\\&\quad-\mathbf{Pr}\left\{\sum_{j=1}^{m}S_j^i{=}\mathbb{E}[S^i]\ \text{or}\ \sum_{j=1}^{m}N_j^i{=}\mathbb{E}[N^i]\right\}\end{aligned} \tag{19}$$

By combining equations (16) and (19) and taking the limit as $m\to\infty$ we can deduce:

$$\mathbf{Pr}\left\{\bar{h_i}(\mathbf{x},\infty)=h_i(\mathbf{x})\right\}=1 \tag{20}$$

which proves the lemma. □

Lemma 2. *Given Γ a finite subset of the feasible solution set X, then:*

$$\boldsymbol{Pr}\left\{\bigcap_{i,\mathbf{x}}\bar{h_i}(\mathbf{x},\infty)=h_i(\mathbf{x}),\forall i\in\mathcal{Q},\mathbf{x}\in\Gamma\right\}=1 \tag{21}$$

Proof. From De Morgan's rule:

$$\begin{aligned}&\mathbf{Pr}\left\{\bigcap_{i,\mathbf{x}}\bar{h_i}(\mathbf{x},\infty)=h_i(\mathbf{x}),\forall i\in\mathcal{Q},\mathbf{x}\in\Gamma\right\}\\&=1-\mathbf{Pr}\left\{\bigcup_{i,\mathbf{x}}\bar{h_i}(\mathbf{x},\infty)\neq h_i(\mathbf{x}),\forall i\in\mathcal{Q},\mathbf{x}\in\Gamma\right\}\\&\geqslant 1-\sum_{i,\mathbf{x}}\mathbf{Pr}\left\{\bar{h_i}(\mathbf{x},\infty)\neq h_i(\mathbf{x})\right\}=1\end{aligned} \tag{22}$$

□

Theorem 1. *If there exist at least one feasible solution to problem **P1** (i.e. X is non-empty), then $\mathbf{x}_m$ is feasible almost surely as $m\to\infty$ (i.e. $X_m^*\subseteq X$). Furthermore, if there exist at least one optimal solution such that $h_i(\mathbf{x})>0,\forall i\in\mathcal{Q}$, then $\mathbf{x}_m$ is also optimal almost surely as $m\to\infty$.*

Proof. We define the set $X_1=\{\mathbf{x}\in X:h(\mathbf{x})\ngtr 0\}$. X_1 represent the set of solutions violating the QoS constraints. If we guarantee that $\mathbf{x}_m\notin X_1$, then $\mathbf{x}_m$ is a feasible solution. By applying lemma 2 for each point $\mathbf{x}\in X_1$, then $\bar{h}(\mathbf{x},\infty)=h(\mathbf{x}),\forall\mathbf{x}\in X_1$, almost surely. Let $\epsilon=\min\limits_{x\in X_1}\max\limits_{i}\{-h_i(\mathbf{x})\}$, then $\epsilon>0$. If we define:

$$M_1=\inf\left\{m_1:\max_{x}\ \max_{i}|\bar{h_i}(\mathbf{x},m)-h_i(\mathbf{x})|<\epsilon,\forall m\geqslant m_1\right\}$$

then $M_1<\infty$, since ϵ is positive and $\bar{h}(\mathbf{x},\infty)-h(\mathbf{x})=0$, almost surely. As a result, we guarantee that after a finite number of simulation runs, $\bar{h_i}(\mathbf{x},m)$ has to be greater than zero $\forall i$. In other words, $X_m^*\nsubseteq X_1$ as $m\to\infty$ almost surely, which proves the first part of the theorem.

Similarly, we define the set $X_2=\{\mathbf{x}\in X^*:h(\mathbf{x})>0\}$. By applying lemma 2 for each point $\mathbf{x}\in X_2$, then $\bar{h}(\mathbf{x},\infty)=h(\mathbf{x}),\forall\mathbf{x}\in X_2$, almost surely. Let $\epsilon=\min\limits_{x\in X_2}\min\limits_{i}\{h_i(\mathbf{x})\}$, then $\epsilon>0$. If we define:

$$M_2=\inf\left\{m_2:\max_{x}\ \max_{i}|\bar{h_i}(\mathbf{x},m)-h_i(\mathbf{x})|<\epsilon,\forall m\geqslant m_2\right\}$$

then $M_2<\infty$, since ϵ is positive and $\bar{h}(\mathbf{x},\infty)-h(\mathbf{x})=0$, almost surely. As a result, we guarantee that after a finite number of simulation runs, $\bar{h_i}(\mathbf{x},m)$ has to be greater than zero $\forall i$. In other words, $X_2\subseteq X_m^*$ as $m\to\infty$ almost surely. If X_2 is non-empty, then $\mathbf{x}_m$ is guaranteed to be optimal, as all the members of X_m^* corresponds to the same value of the objective function. □

The condition that X_2 to be non-empty is easy to be satisfied practically, since it requires a good coincidence for an integer solution to hit exactly one of the pre-specified QoS levels. Even if this coincidence happens, the QoS levels can be perturbed by an infinitesimal amount to make sure that no integer solution can perfectly hit them, while practically not changing the original problem.

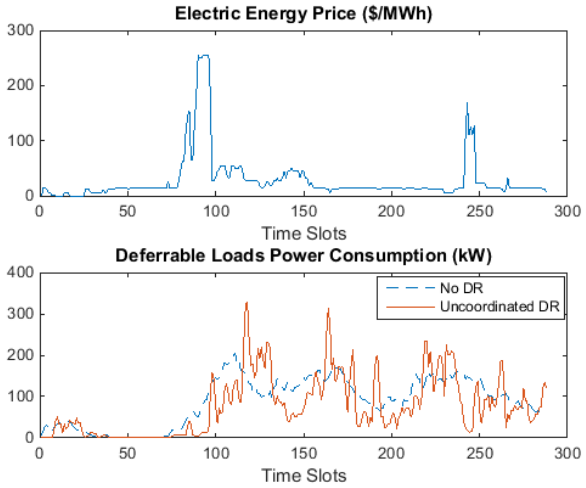


Fig. 3: Top: Ontatio's electricity market clearance profile for 12th of June, 2015. The profile is considered in our study as the profile of the day-ahead electricity price. Bottom: The effect of uncoordinated DR on deferrable load profile.

IV. Numerical Results

A. Raw Data

We randomly select an Ontario's electricity market clearance profile of day (12th of June, 2015) as our DAP profile, as shown in Fig. 3. The price is updated every 5 minutes, hence, the time horizon is composed of different 288 time slots. The data collected for the residential demand should be an appliance-level with a temporal resolution of at least 5 minutes. The power drawn by an appliance when it is turned-on can usually be deduced or given in the appliance's data sheet. However, the consumer's behavior towards using this appliance cannot be easily deduced. The same appliance can be used differently by different consumers. Therefore, a statistical, appliance-level demand data should be collected by monitoring each appliance for a large number of consumers. Unfortunately, such data is not abundant, since the concern of the operators and planners of power networks is the aggregated demand of many consumers, not the appliance-level demand. Nevertheless, there exists a study [22] which recorded the appliance-level residential demand for a period of almost two years (2012-2014) for a small number of consumers (5 consumers only). Our approach is to use the demand data of one of these consumer to represent the demand behavior of all the consumers in our system. We diversify the demand for 1000 consumers by altering these data randomly from a consumer to another. TCL parameters are extracted from different studies (given in Table I), since it is difficult to deduce them parameters from the power measurements alone.

B. Simulation Results

We first study effects of the uncoordinated cost minimization, i.e. when each consumer minimizes its own energy consumption cost, disregarding the capacity of distribution transformer (such as 600 kW). In the uncoordinated DR approach, the price profile is sent to each individual appliance.

TABLE I: TCL Appliances' Parameters

Appliance	Power level (kW)	R_{th} (°C/kW)	C_{th} (kJ/°C)	θ_e (°C)	Setting point (°C)	Dead band (°C)
Fridge [16]	0.2	100	2880	20	4	0.25
A/C [23]	5	2.5	9000	30	20	1

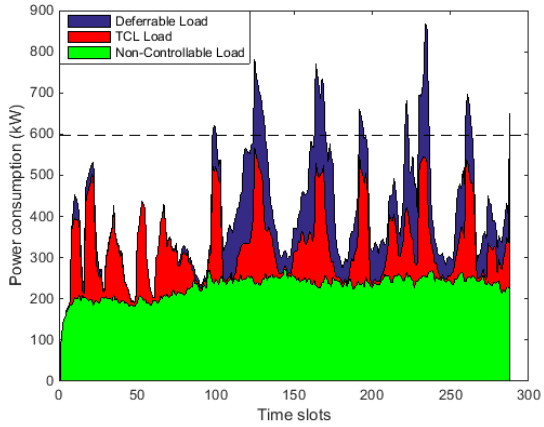


Fig. 4: Aggregated demand profiles for controllable and non-contrallable loads after implementing an uncoordinated demand response.

Whenever an appliance is turned on, it schedules its demand in the first time slot before the deadline, which corresponds to the minimum power consumption cost of its demand only, regardless of other appliances' decisions. The effect of uncoordinated DR on deferrable loads is shown in Fig. 3, where demands are shifted from time slots of high price to be concentrated around low price time slots. Aggregated load profiles for all consumers after implementing DR is shown in Fig. 4, respectively. It is clear from Fig. 4 that the power consumption of the controllable loads is concentrated at specific time periods, where the electricity price is minimal. However, since there is no coordination among consumers, the power consumption exceeds from time to time the capacity of the distribution transformer. Therefore, it is beneficial to use our optimization framework to avoid transformer overloading.

Before applying our proposed methodology, we need to define the set of demand classes such that their parameters are close to the characteristics of the input appliances. To achieve this target, we utilize the popular algorithm (used for vector quantization) described in [24]. We applied this algorithm when the number of classes are set to 4; The CPM is then used for optimal allocation of different demand block capacities. The total allocated power profile is illustrated in Fig. 5 and the actual power consumption profile after implementing DR is illustrated in Fig. 6. Finally, we evaluated the performance of the proposed demand management framework by comparing the percentage of cost reduction to the case of the uncoordinated DR, which is the best achievable performance. As illustrated in Fig. 7, it is observed that as the cost reduction becomes closer to the best achievable one (43%) as we employ more queues in our demand management framework.

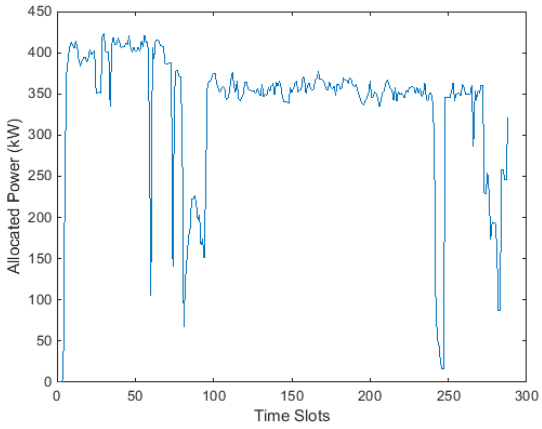


Fig. 5: Allocated Power Profile .

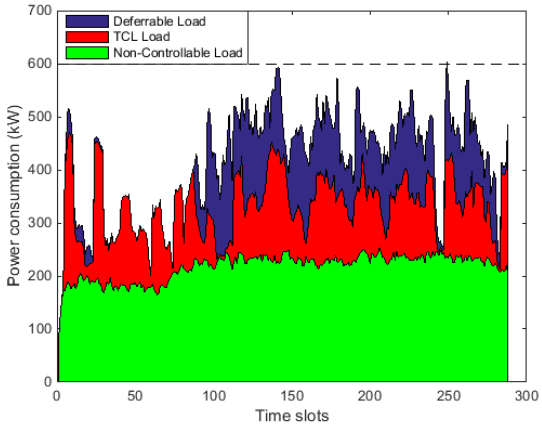


Fig. 6: Aggregated demand profiles for controllable and non-contrallable loads after implementing an coordinated demand response. Notice that the aggregated power consumtption does not exceed the capacity of the distribution transformer.

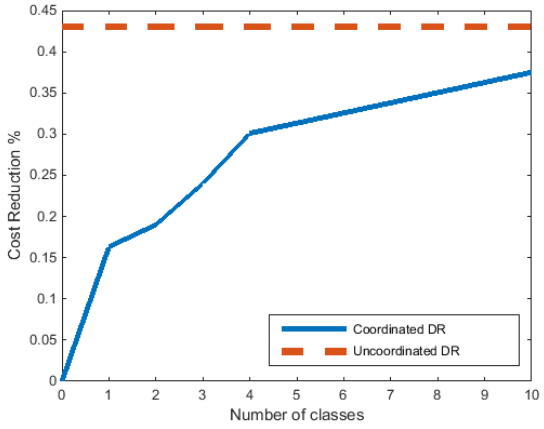


Fig. 7: Energy consumption cost reduction as a percentage of the energy cost versus the number of demand classes employed in the demand management framework.

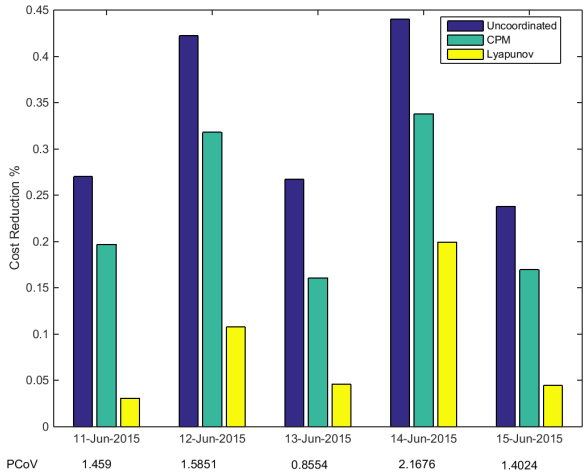


Fig. 8: A comparison between cost reduction percentage of three different DR schemes: uncoordinated, cutting-plane method (CPM) used in our work, and Lyapunov optimization used in [10]. The results are depicted for five different days i.e. under different price profiles. The price coefficient of variation (PCoV)for each price profile is written under the date of the profile's day.

It must be noted that the amount of cost reduced highly depends on the price profile. No DR benefit can be obtained from a flat price profile regardless of the efficiency of the optimization algorithm used. Fig.8 shows different cost reduction values under different price profiles. In this figure, we compare the results of our proposed DR methodology (under 4 classes of demand) to that of uncoordinated DR and the popular DR methodology in [10]. The improvement of the our results over that in [10] is attributed to: 1) our work uses a multi-class queueing system instead of single-class system used in [10], and 2) the work in [10] assumes a perfectly elastic demand i.e. an appliance energy demand can be satisfied in one time slot. In our model, an appliance demand can extend over several time slots, which makes our methodology more efficient in handling residential demand, especially for short time scale systems.

We select price coefficient of variation (PCoV) as a measure of the magnitude of price variations. However, from Fig. 8, higher cost reduction does not always correspond to higher PCoV as indicated from the third and fifth days. This is because cost reduction does not only depend on the total variations magnitude, but also the location of these variations.

V. Conclusion

In this paper, we propose a methodology for residential demand management, which is composed of two steps. The first step is to classify the demand requirements of different residential appliances to fixed classes of demands. Each class of demands is defined by three parameters, power consumption level, energy duration, and a deadline for satisfying its requirement. Following this approach, new units of demand are defined, referred to as "demand blocks", where each unit represents a certain class of demands. The second step of the proposed framework is to schedule these demand blocks in an optimal way such that the DR objective is achieved.

We develop a computationally efficient algorithm that aims to minimize the upper bound of the energy consumption cost, while taking the random nature of demand block arrivals into consideration.

REFERENCES

[1] S. Braithwait, "Behavior modification," *IEEE Power and Energy Magazine*, vol. 8, no. 3, pp. 36–45, Apr. 2010.

[2] Canada's Ministry of Industry, "Report on energy supply and demand in canada: 2014 preliminary," 2016.

[3] Z. Chen, L. Wu, and Y. Fu, "Real-time price-based demand response management for residential appliances via stochastic optimization and robust optimization," *IEEE Transactions on Smart grid*, vol. 3, no. 4, pp. 1822–1831, Sept. 2012.

[4] Y. Li, B. L. Ng, M. Trayer, and L. Liu, "Automated residential demand response: Algorithmic implications of pricing models," *IEEE Transactions on Smart Grid*, vol. 3, no. 4, pp. 1712–1721, Sept. 2012.

[5] Z. Zhao, W. C. Lee, Y. Shin, and K.-B. Song, "An optimal power scheduling method for demand response in home energy management system," *IEEE Transactions on Smart Grid*, vol. 4, no. 3, pp. 1391–1400, June 2013.

[6] P. Chavali, P. Yang, and A. Nehorai, "A distributed algorithm of appliance scheduling for home energy management system," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 282–290, Jan. 2014.

[7] P. Yi, X. Dong, A. Iwayemi, C. Zhou, and S. Li, "Real-time opportunistic scheduling for residential demand response," *IEEE Transactions on Smart Grid*, vol. 4, no. 1, pp. 227–234, Feb. 2013.

[8] T.-H. Chang, M. Alizadeh, and A. Scaglione, "Real-time power balancing via decentralized coordinated home energy scheduling," *IEEE Transactions on Smart Grid*, vol. 4, no. 3, pp. 1490–1504, Aug. 2013.

[9] R. Deng, Z. Yang, J. Chen, N. R. Asr, and M.-Y. Chow, "Residential energy consumption scheduling: A coupled-constraint game approach," *IEEE Transactions on Smart Grid*, vol. 5, no. 3, pp. 1340–1350, Apr. 2014.

[10] M. J. Neely, A. S. Tehrani, and A. G. Dimakis, "Efficient algorithms for renewable energy allocation to delay tolerant consumers," in *Proceedings of First IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Oct. 2010.

[11] Y. Guo, M. Pan, and Y. Fang, "Optimal power management of residential customers in the smart grid," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 9, pp. 1593–1606, Jan. 2012.

[12] Y. Guo, M. Pan, Y. Fang, and P. P. Khargonekar, "Decentralized coordination of energy utilization for residential households in the smart grid," *IEEE Transactions on Smart Grid*, vol. 4, no. 3, pp. 1341–1350, Aug. 2013.

[13] Y. Huang, S. Mao, and R. M. Nelms, "Adaptive electricity scheduling in microgrids," in *Proceedings of IEEE INFOCOM'13*, Apr. 2013.

[14] D. O. Neill, M. Levorato, A. Goldsmith, and U. Mitra, "Residential demand response using reinforcement learning," in *Proceedings of First IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Oct. 2010.

[15] N. Lu, D. P. Chassin, and S. E. Widergren, "Modeling uncertainties in aggregated thermostatically controlled loads using a state queueing model," *IEEE Transactions on Power Systems*, vol. 20, no. 2, pp. 725–733, May 2005.

[16] E. Kara, M. Berges, and G. Hug, "Impact of disturbances on modeling of thermostatically controlled loads for demand response," *IEEE Transactions on Smart Grid*, vol. 6, no. 5, pp. 2560–2568, Sept. 2015.

[17] J. L. Mathieu, S. Koch, and D. S. Callaway, "State estimation and control of electric loads to manage real-time energy imbalance," *IEEE Transactions on Power Systems*, vol. 28, no. 1, pp. 430–440, Jan. 2013.

[18] J. Widén and E. Wäckelgård, "A high-resolution stochastic model of domestic activity patterns and electricity demand," *Applied Energy*, vol. 87, no. 6, pp. 1880–1892, June 2010.

[19] P. Du and N. Lu, "Appliance commitment for household load scheduling," *IEEE Transactions on Smart Grid*, vol. 2, no. 2, pp. 411–419, May 2011.

[20] J. Atlason, M. A. Epelman, and S. G. Henderson, "Call center staffing with simulation and cutting plane methods," *Annals of Operations Research*, vol. 127, no. 1-4, pp. 333–358, 2004.

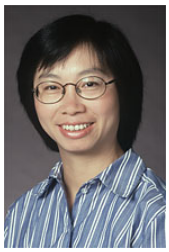
[21] P. E. Fishback, *Linear and nonlinear programming with Maple: an interactive, applications-based approach*. CRC Press, 2009.

[22] J. Kelly and W. Knottenbelt, "The uk-dale dataset, domestic appliance-level electricity demand and whole-house demand from five uk homes," *Scientific data*, vol. 2, Mar. 2015.

[23] H. Hao, B. M. Sanandaji, K. Poolla, and T. L. Vincent, "A generalized battery model of a collection of thermostatically controlled loads for providing ancillary service," in *Proceedings of 51st IEEE Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Oct. 2013.

[24] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on communications*, vol. 28, no. 1, pp. 84–95, Jan 1980.

Fadi Elghitani received the B.Sc. and M.Sc. degrees from Ain-Shams University, Cairo, Egypt. Currently, Mr. Fadi is working on his Ph.D. in the University of Waterloo, Waterloo, ON, Canada. His research interests include Smart Grid and Wireless Communications.



Weihua Zhuang has been with the Department of Electrical and Computer Engineering, University of Waterloo, since 1993, where she is a Professor and a Tier I Canada Research Chair in wireless communication networks. Her current research focuses on resource allocation and QoS provisioning in wireless networks, and on smart grid. She is a co-recipient of several best paper awards from IEEE conferences. Dr. Zhuang is/was the Technical Program Chair for IEEE VTC Fall 2017 and Fall 2016, and the Editor-in-Chief of IEEE Transactions on Vehicular Technology (2007-2013). She is a Fellow of the IEEE, a Fellow of the Canadian Academy of Engineering, a Fellow of the Engineering Institute of Canada, and an elected member in the Board of Governors and VP Publications of the IEEE Vehicular Technology Society.