On the Performance of Distributed and Cloud-Based Demand Response in Smart Grid

Mohammad Hossein Yaghmaee, *Senior Member, IEEE*, Alberto Leon-Garcia, *Fellow, IEEE*, and Morteza Moghaddassian, *Student Member, IEEE*

***Abstract*—By locally solving an optimization problem and broadcasting an update message over the underlying communication infrastructure, demand response program based on the distributed optimization model encourage all users to participate in the program. However, some challenging issues present themselves, such as the existence of an ideal communication network, especially, when utilizing wireless communication, and the effects of communication channel properties, like the bit error rate, on the overall performance of the demand response program. To address the issues, this paper first defines a cloud-based demand response (CDR) model, which is implemented as a two-tier cloud computing platform. Then a communication model is proposed to evaluate the communication performance of both the CDR and distributed demand response models. This paper shows that when users are finely clustered, the channel bit error rate is high and the user datagram protocol (UDP) is leveraged to broadcast the update messages, making the optimal solution unachievable. Contradictory to UDP, the transmission control protocol will be caught up with a higher bandwidth and increase the delay in the convergence time. Finally, this paper presents a cost-effectiveness analysis which confirms that achieving higher demand response performance incurs a higher communication cost.**

***Index Terms*—Demand side management, smart grid communication, cloud computing, microgrids, optimization.**

NOMENCLATURE

Indices and Sets

$A_{i,r}$	Set of shiftable appliances of customer i in region r
a	Appliances number
g	Microgrid index
i	Customer index
k	Cluster index
r	Region index
t	Time index.

Manuscript received July 22, 2016; revised November 19, 2016, February 1, 2017, and March 11, 2017; accepted March 21, 2017. Date of publication March 28, 2017; date of current version August 21, 2018. Paper no. TSG-00968-2016. *(Corresponding author: Mohammad Hossein Yaghmaee.)*

M. H. Yaghmaee is with the Department of Computer Engineering, Ferdowsi University of Mashhad (FUM), Mashhad 91779-11111, Iran (e-mail: hyaghmae@um.ac.ir).

A. Leon-Garcia and M. Moghaddassian are with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 2E4, Canada (e-mail: alberto.leongarcia@utoronto.ca; m.moghaddassian@mail.utoronto.ca).

Digital Object Identifier 10.1109/TSG.2017.2688486

Parameters

$\gamma_{a,i,r}^{max}$	Maximum power levels denoted of home appliances $a \in A_{i,r}$
$\mathcal{C}_r$	Number of cache points in region r
Δ_r	Hop distance between cache points
a_t, b_t, c_t	Hourly cost coefficients
$E_{a,i,r}$	Total energy needed for the operation of the shiftable appliances a of customer i in region r
P_t	Hourly energy price determined by the utility company
$\gamma_{a,i,r}^{min}$	Minimum power levels denoted of home appliances $a \in A_{i,r}$
B_1	First power consumption block in the piecewise linear cost function
$C^l(L_r^t)$	Cost of providing L_r^t energy by utility companies in region r using linear cost function
$C^q(L_r^t)$	Cost of providing L_r^t energy by utility companies in region r using quadratic cost function
H_r	Number of Data Aggregation Points (DAP) in region r
N_c	Number of clusters in the power system
N_k	Number of customers in cluster k
N_r	Number of customers in region r
pr_t^1, pr_t^2	Hourly price of the first and second steps for piecewise linear function
s_m	Message size in bits
ϵ_k	Channel bit error rate in region k
$[\alpha_{i,r}^a, \beta_{i,r}^a]$	Desired operation time interval of appliances a of customer i in region r
∂	Profit factor of the utility company
T	Scheduling horizon (usually $T = 24$)
γ	A constant factor
G	Number of microgrids in the system
R	Number of regions
RTT	Round trip time for sending a TCP message and receiving its acknowledgment
$d(X, \hat{X})$	Euclidean distance between X and $\hat{X}$.

Power System Variables

$l_{a,i,r}^t$	Energy consumption for the operation of the shiftable appliances a of customer i in region r at time slot t

1949-3053 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

L^{*t}_{CDR}	Total hourly optimized load in the power system for the CDR model
L^{*t}_{DDR}	Total optimal hourly load in the power system for the DDR model
$l^{*t}_{n,k,r}$	Optimal schedule of user n in cluster k of region r
$l\_NS^t_{i,r}$	Power consumption of the non-shiftable appliances of customer i in region r at time slot t
$l\_S^t_{i,r}$	Power consumption of the shiftable appliances of customer i in region r at time slot t
$L\_S_{i,r}$	Total shiftable daily load of customer i in region r
PAR	Peak to Average Ratio
B^t_g	Predicted remaining battery energy in microgrid g at time t
$C^u_{i,r}$	Daily billing of user i in region r
L^{*t}	Total hourly demand load from all regions
L^{*t}_r	Optimized total load in region r at time t
L^t_r	Un-optimized total load in region r at time t
L^t	Total hourly load in the power system
P^t_g	Predicted power generation by microgrid g at time t
Y^{*t}	Optimized consumption vector from microgrid storages
Y^t_g	Power consumption level of microgrid g at time t
Y^t	Total un-optimized hourly power consumption of all microgrids in the system
$l^{*t}_{i,r}$	Optimized load of customer i in region r at hour t
$l^t_{i,r}$	Un-optimized load of customer i in region r at hour t
$l^t_{n,k,r}$	Hourly un-optimized load of user n in cluster k of region r
$l^t_{-n,k,r}$	Hourly un-optimized load of all other users (except user n) in cluster k of region r.

Communication Network Variables

$p_k(s_m)$	Packet error probability for a packet with size s_m in region k
B_{DDR}	Total bandwidth required by the DDR model to find the optimum load in all regions in the power system
D_{CDR}	Convergence time of the CDR model
D_{DDR}	Convergence time of the DDR model
N^k_{DDR}	Total number of required update messages to reach the optimum point for the DDR model in cluster k
$N_{HBH}(i,H,s_m)$, $N_{e2e}(i,H,s_m)$, $N_{IC}(i,H,s_m)$	Total number of transmissions of a packet with size s_m from source node i through H intermediate hops by hop by hop, end-to -end, and intermediate caching methods, respectively
N_{it}	Number of required convergence iterations
N^k_{it}	Total number of iterations to find the Nash equilibrium point in cluster k
N_{ret}	Number of update messages
T_{com}	End-to-end delay required to transfer information between customers and the edge cloud
T_{opt}	Central optimization process time
T_{trans}, T_{prop}, T_{proc} and T_{queue}	Total delay required for the transmission, propagation, processing, and queuing of all messages in the path, respectively
t_{opt}	Time required to run the optimization process in each iteration.

I. Introduction

THE SMART grid has been recently developed to provide different services for both customers and utility companies. It utilizes the benefits of Information and Communication Technology (ICT) to control and monitor different parts of a power system. As the amount of demand in the power grid is not fixed and electrical energy cannot be easily stored, the difference between demand and supply is crucial in reducing energy costs. Demand Response (DR) is utilized to balance total demand with the amount of supply. In demand response approaches, customers change their power consumption depending on the energy price. This also helps power supply authorities to better control the grid. By utilizing demand response approaches, it is possible to reduce or shift energy consumption from peak hours to period of less demand. To achieve this, customers can decide to disconnect non-essential loads at peak hours.

The rest of this section explains different demand response programs and existing mathematical approaches in detail.

A. Demand Response Programs

Demand response programs are mainly divided into the following two classes, load-response and price-based programs.

1) Load-Response Programs: In load-response programs, utilities offer customers credit for reducing electricity consumption for specified periods of time. These programs can take a number of different forms, including Direct Load Control (DLC), curtailable load, interruptible load, and scheduled load. Direct load control [1] is a demand response technique that enables utility companies to turn on/off some specific appliances during peak demand periods and critical events by sending direct communication signals. In load curtailment, customers or utility companies agree to reduce part of their consumption at peak hours to prevent any power outages. Customers participating in interruptible-load programs agree to switch off major portions or even their entire load for specified periods of time. In the load scheduling approach, some non-critical loads are shifted from peak hours to non-peak hours. Note that, in this case, the total power consumption before and after scheduling is the same. In other words, load scheduling just moves some loads to non-peak periods.

2) Price-Based Programs: These programs allow customers to voluntarily reduce their demand in response to price signals. Such programs are categorized into time-based pricing and demand bidding. In time-based pricing, the cost of using energy is calculated based on the current demand and supply in the power network [2], [3]. This can be determined in advance (such as Time of Use (ToU) pricing) or changed in real time (Real Time Pricing (RTP)). In most demand response programs, the primary goal is to develop a robust model in which the total demand of customers is lowered at peak-time hours so that the cost of power generation is reduced [4]. For demand bidding [5], there are some incentive-based demand response programs for the smart grid which credit customers to reduce their electric load during peak hours. Customers can submit their load reduction bids on an hourly basis for any event without a financial penalty.

B. Demand Response Mathematical Approaches

The main objective of most demand response programs is to optimize the cost function. As investigated in [6], there are different optimization techniques, such as convex optimization, game theory, dynamic programming, the Markov decision process, stochastic programming, and Particle Swarm Optimization (PSO). Most optimization problems in demand response are converted to distributed optimization. Distributed Demand Response (DDR) models, such as the ones given in [7], require that the utility company send a price signal to all users. After that, each user tries to individually solve an optimization problem. As the optimization problem is mostly defined based on the aggregate load, users can participate in a non-corporative game by sending their current optimal load to all users in the region.

Consider the DDR model given in [7]. As shown in Fig. 1, the distributed optimization presented in [7] is sequential, which means that optimization is done asynchronously and in a sequential manner. At each iteration, all users have to broadcast their schedules to all the other users in the system. When the communication channel is not ideal, some update messages may get lost. In this case, more iterations are needed to achieve the Nash equilibrium point which increases the required bandwidth and convergence time. To transfer update messages, Transmission Control Protocol (TCP) or User Datagram Protocol (UDP) may be used. Both TCP and UDP use the Internet Protocol (IP) at the lower layer to transfer their data. TCP guarantees data delivery by employing an Acknowledgement (ACK) mechanism to make sure that the data is received. If TCP does not receive an ACK message within a certain amount of time, it will retransmit the lost message. Finally, after some retransmissions, the update message is correctly delivered to all users, but this might increase the convergence delay due to packet retransmissions. Unlike TCP, UDP uses a simple connectionless transmission model which cannot provide a guarantee of delivery, ordering, or duplicate protection. If some UDP update messages are lost, the DDR might converge to an un-optimized solution.

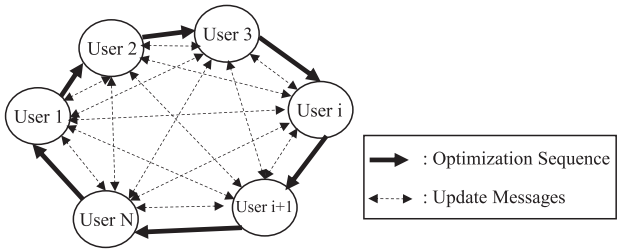


Fig. 1. The sequentially distributed optimization model [7].

To prove how cloud computing-based optimization can improve the DDR model in terms of communication and power network performance, the present study presents a Cloud-based Demand Response (CDR) which performs central optimization to achieve a global optimal solution. The CDR facilitates the sharing of information among all users once privacy issues are met. The elastic nature of cloud computing allows the CDR model to dynamically provide the required communication and computation resources as needed. The CDR architecture consists of two levels of cloud computing, edge, and core. The power consumption information of all users in each region is transmitted to the edge cloud through some data aggregation points.

As investigated in [8], there are various challenges, security limitations, and solutions related to cloud privacy. Current security methods mainly focus on authentication to protect the user's privative data from illegal access. In [9], the shared authority-based Privacy-Preserving Authentication Protocol (SAPA) was proposed which addresses privacy issues for cloud storage. In [10], different techniques for cloud computing privacy which can be used in CDR have been introduced.

C. Motivations

The main motivations of the current work are summarized as follows:

- Most demand response programs assume that there is a perfect two-way communication network between the utility company and customers. However, the real communication network is not perfect especially when there are time dependent wireless channels between customers and utility companies. Knowing that data communication networks have some inherent properties, including the channel bit error rate, one should consider the effects of these properties on the required bandwidth and the convergence time of CDR and DDR. The packet loss ratio and limited bandwidth of the wireless channels are important issues which should be considered in designing efficient demand response programs.

- The distributed demand response model classifies all customers in some clusters with different members and provides peer-to-peer data networking as well as a communications protocol between all of them. In this case, it is a huge challenge to provide a local area network among all users located in the different geographic regions. The effect of cluster size on the bandwidth and delay performance of the demand response program should be investigated.

D. Contributions

The main contribution of the present paper is to present a communication model to evaluate the communication performance of DDR and CDR models. It should also be noted that the CDR model is application-agnostic and any future or current demand response model can be re-designed. However, one should recall that design choice is a function of the computational model of the demand response. Therefore, although there are many concerns in re-designing an existing demand response model, it should be generally considered as feasible. The present study considers the wireless mesh tree network with different packet delivery models, including Hop by Hop (HBH), end-to-end (e2e), and Intermediate Caching (IC). Also investigated is the effect of communication channel characteristics on the performance of distributed and cloud-based demand response. Most related work in demand response evaluates the performance in terms of power network performance metrics and do not study the effect of the communication channel on demand response performance. To the best of our knowledge, the current work is the first study to consider the effect of the communication channel on the performance of distributed and cloud-based demand response programs. In the present research, the distributed and cloud-based demand responses are evaluated in terms of both power network and communication network performance metrics. An analytical model is developed for measuring bandwidth requirements and the convergence time of both DDR and CDR models. It is shown that, under the non-ideal communication channel, the demand response performance is degraded. A cost-effectiveness analysis is presented which indicates a direct relation between demand response performance and communication cost. Furthermore, the current paper introduce a cloud-based demand response architecture for the future smart grid which utilizes Network Function Virtualization (NFV) concepts and presents virtual Networked Home Energy Management (vNHEM). The vNHEM is a virtual implementation of network home management systems which utilizes the benefits of cloud computing virtualization, such as flexibility, cost, scalability, and security.

E. Paper Organization

The rest of the present paper is organized as follows. Section II discusses some related work in this field. Section III explains the proposed model in detail. Section IV presents simulation results that confirm the superior performance of the proposed model. Finally, Section V concludes the paper.

II. Related Work

By considering two different classes of customer appliances, a distributed energy scheduling algorithm is proposed in [11]. Each user knows the price and runs an iterative greedy algorithm to find the optimal energy consumption schedule. In each iteration, all users communicate their energy consumption schedule to the utility company. The price will then be adjusted depending on the overall system load and then be broadcast to all users. Users will then adjust their energy consumption based on the new price. These iterations continue until convergence. Tan *et al.* [12] consider a smart grid with plug-in electric vehicles (PEVs) and on-site renewable distributed generators. They conceive a scenario in which users and PEVs can sell back the energy they generate to the grid. The optimization algorithm is computed in parallel and all users need to report their usage curve only to the utility company. In [13], a distributed demand response management model for residential customers is proposed. Each customer receives information from the utility about the electricity costs and total load profile of the network. This is updated by customers and the daily schedule of their loads is sent back. The data exchange and load profile updates continue until no further improvement in the total load profile is achieved. Reference [14] offers a distributed algorithm for real-time demand response in the future multi seller-multi buyer smart distribution grid. Each utility company and user locally solve subproblems to perform energy allocation. Each utility company sets a clearing price to balance supply and demand, while each user coordinates demand from different companies to meet the baseline demand. Safdarian *et al.* [15] present a decentralized demand response model based on bi-level optimization, where the upper subproblem modifies the system load profile and the lower subproblem minimizes individual customer energy costs. Using a relaxation method, the problem is relaxed to a single-level optimization problem. Similar to [7], a Home Load Management (HLM) module is embedded in the customer smart meter to locally solve the optimization problem. The utility exchanges load information with the customer HLM modules to achieve the optimal load profile.

With the goal of maximizing the total welfare of all users, [16] proposes a distributed demand response program in which each user is independently responsible for finding the optimal power consumption schedule. Zheng and Cai [17] select a heating ventilation and air conditioning (HVAC) system as an elastic interruptible load and develop a queueing model for the HVAC systems thermal dynamics. Furthermore, an optimization problem is suggested which minimizes the average variation of the nonrenewable power demand by turning on/off the HVAC systems. A suboptimal distributed control algorithm is also implemented that reduces complexity and communication costs. Reference [18] presents a direct residential demand response program for the AC power distribution network of a large number of residential households. In this model, customers have a contract with the Load Service Entity (LSE) and allow the LSE to control some of their appliances through customer Home Energy Management systems (HEMs). This demand response program is designed so that the social welfare is maximized and the total demand is minimized during peak hours. Reference [19] offers a decentralized and scalable demand-side energy management approach which not only optimizes the demand levels of each household, but also explores computing the demand states of each household and the feasible transitions between these states. By the clustering of each household's historical consumption data, a data-driven methodology is proposed which captures the temporal dynamics of demand and identifies

target consumers for DR programs. Communication networks play a significant role in real-time demand response and efficient use of energy. Reference [20] considers a wireless Neighbor Area Network (NAN) with a large number of sensors and some concentrators. Due to the limited bandwidth of the wireless channel, the sensors are divided into groups which compete to access a shared wireless channel in a time-division-multiplexing manner. The main objective is to provide Quality of Service (QoS) for supporting demand response service. Different QoS parameters are considered such as packet delay, packet error probability, and node outage probability. To determine the density of the concentrator, the number of required concentrators and the location of these concentrators, an analytical model is developed which quantifies the QoS metrics.

Pruckner *et al.* [21] study the influence of packet loss and latency on the performance of the balance between electricity supply and demand. By a zero-mean random variable with normal distribution, [22] models the impact of unreliable communications on the demand response management performance of real-time pricing in the microgrid system. Reference [23] models and analyzes the reliability of wireless communication network for demand response control accuracy. For this purpose, the impact of wireless communication errors on demand response control strategy is evaluated. To analyze the reliability of communication services, the outage probability as the performance metric is then considered. Analytical and simulation results confirm the accuracy of the presented model.

III. System Model

Fig. 2 depicts a power network with R different regions. In each region r, N_r, customers are connected to the grid and consume energy. There are G distinct microgrids in the system. Each microgrid is equipped with a Distributed Generation (DG) and Distributed Storage (DS) unit. Without loss of generality, the presented study considers Photovoltaic (PV) energy generation in each microgrid. This involves two-tier cloud computing consisting of edge and core cloud. Edge clouds are located close to end users to decrease end-to-end latency.

The consumption information of all users in each region is transferred to the edge cloud. It runs the optimization problem and finds the optimal power consumption schedule for users so that the total energy consumption cost is minimized. The optimal consumption schedule is transferred to each customer via existing communication networks. After the edge cloud calculates the optimal power consumption schedule, the total scheduled load of each region is transferred to the core cloud. Based on the power system's total hourly load, the core cloud schedules the power consumption in each microgrid so that the total cost for the utility company is minimized.

Two different demand response approaches can be compared: the proposed Cloud-based Demand Response (CDR) and the existing Distributed Demand Response (DDR) [7] as shown in Fig. 3. Generality, any other DDR model can be considered. Let a_t, represents the energy price at time t which,

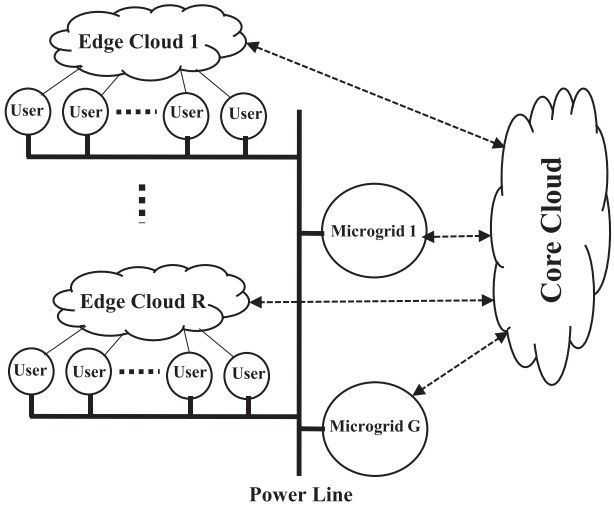


Fig. 2. System model.

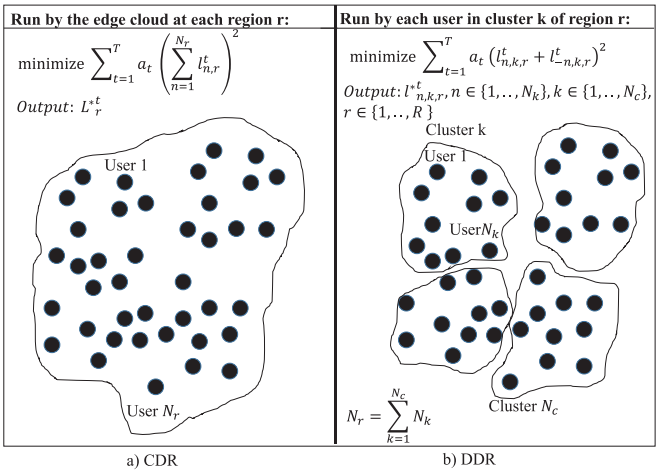


Fig. 3. CDR and DDR model.

in the DDR model is determined by the utility company and announced to all users. In CDR, all consumption information is transferred to a central controller (edge cloud) and then the central optimization problem is solved. The results are then forwarded to the customers.

A. Model Formulation

Suppose $l_{i,r}^t$, represents the un-optimized load of user i in region r at hour t. The following equation is always satisfied:

$$l_{i,r}^t = l\_S_{i,r}^t + l\_NS_{i,r}^t \tag{1}$$

where $l\_S_{i,r}^t$ and $l\_NS_{i,r}^t$ are the power consumption of the shiftable and non-shiftable appliances of customer i in region r at time slot t, respectively. Suppose $L\_S_{i,r}$ represents the total shiftable daily load of customer i in region r. Note that $L\_S_{i,r}$ before and after optimization is a constant value. Let $A_{i,r}$ denote the set of shiftable appliances a of customer i in region r. For any shiftable appliances $a \in A_{i,r}$, the desired operation time is determined by $[\alpha_{i,r}^a, \beta_{i,r}^a]$. The operation time depends on the amount of energy required. Let $l_{a,i,r}^t$ and $E_{a,i,r}$ denote the hourly consumption and total energy needed for the

operation of the shiftable appliances a of customer i in region r at time slot t, respectively. Note that $\sum_{a\in A_{i,r}} l^t_{a,i,r} = l\_S^t_{i,r}$. For any shiftable appliances $a \in A_{i,r}$, $\gamma^{min}_{a,i,r}$ and $\gamma^{max}_{a,i,r}$ are the minimum and maximum power levels.

Let $L^t_r = \sum_{i=1}^{N_r} l^t_{i,r}$ is hourly un-optimized total load in region r. The following quadratic cost function has been widely used to model the cost of energy provided by utility companies in region r, $C^q(L^t_r)$ [7], [24]:

$$C^q\left(L^t_r\right) = a_t {L^t_r}^2 + b_t L^t_r + c_t \tag{2}$$

where a_t, b_t and c_t are hourly cost coefficients determined by some elements, such as operating costs, facility construction, and ownership cost. In most cases, coefficients b_t and c_t are set to 0. Thus the above function is simplified as follows:

$$C^q\left(L^t_r\right) = a_t {L^t_r}^2 \tag{3}$$

The following day ahead minimization problem is run by the utility companies to minimize the total daily cost:

$$\begin{aligned} \textit{minimize} \quad & \sum_{t=1}^{T} C^q\left(L^t_r\right) \\ & = \sum_{t=1}^{T} a_t \left(\sum_{i=1}^{N_r} \left(\sum_{a\in A_{i,r}} l^t_{a,i,r} + l\_NS^t_{i,r} \right) \right)^2 \end{aligned} \tag{4}$$

$$\text{Subject to:} \quad \sum_{t=\alpha^a_{i,r}}^{\beta^a_{i,r}} l^t_{a,i,r} = E_{a,i,r} \tag{5}$$

$$\sum_{t=1}^{T} \sum_{a\in A_{i,r}} l^t_{a,i,r} = L\_S_{i,r} \tag{6}$$

$$\gamma^{min}_{a,i,r} \le l^t_{a,i,r} \le \gamma^{max}_{a,i,r} \tag{7}$$

At the end of the optimization process at each region r, the total optimal demand load vector $L^{*t}_r = \sum_{i=1}^{N_r} l^{*t}_{i,r}$ is transferred to the core cloud. At the core cloud, the aggregate optimized load in all regions is computed as: $L^{*t} = \sum_{r=1}^{R} L^{*t}_r$. Each microgrid g, also sends to the core cloud three vectors, P^t_g, B^t_g, and Y^t_g which represent the prediction of the hourly power generation by the microgrid, the hourly remaining energy in the microgid storage and the hourly power consumption level of each microgrid g, respectively. Let $L^t = (L^{*t} - Y^t)$ denote the total hourly load in the power system and $Y^t = \sum_{g=1}^{G} Y^t_g$ the total un-optimized hourly power consumption of all microgrids in the system.

The following optimization problem is defined at the core cloud:

$$\textit{minimize} \quad \sum_{t=1}^{T} a_t \left(L^{*t} - Y^t\right)^2 \tag{8}$$

$$\text{Subject to:} \quad Y^1_g = P^1_g \tag{9}$$

$$B^t_g = B^{t-1}_g + P^t_g - Y^t_g, t = 2, \ldots, T \tag{10}$$

As the battery of each microgrid is charged by solar energy (P^t_g) and discharged by local consumption (Y^t_g), the Constraints (10) should be satisfied meaning that, at each hour t, the remaining energy of microgrid storage (B^t_g) is the remaining energy from the previous hour (B^{t-1}_g) plus the generated energy at the current time slot (P^t_g) minus the amount of microgrid consumption at the current time slot (Y^t_g). It is supposed that, at the first time slot $(t = 1)$, the battery is completely discharged and all energy generation has been consumed (Constraint (9)).

When the core cloud finishes the running of the central optimization process, the optimized consumption vector from the microgrid storage, Y^{*t} is obtained. Then $L^{*t}_{CDR} = (L^{*t} - Y^{*t})$ is calculated which indicates the total hourly optimized load in the power system.

The Peak to Average Ratio (PAR) is an important power network metric which is defined as the maximum daily load divided by the average load, as follows:

$$PAR = \frac{max\{L^t\}_{t\in\{1,\ldots,T\}}}{avg\{L^t\}_{t\in\{1,\ldots,T\}}} = \frac{Tmax\left\{L^t\right\}_{t\in\{1,\ldots,T\}}}{\sum_{t=1}^{T} L^t} \tag{11}$$

The demand response programs and load shifting approaches are widely used to shift parts of power consumption from high peak hours to off-peak hours so as to reduce the peak-to-average ratio.

Consider the Distributed Demand Response (DDR) model given in [7]. To extend the DDR to a large number of users, it is assumed that all customers in each region r construct N_c different local clusters. At each cluster k, there are N_k different users where $N_r = \sum_{k=1}^{N_c} N_k$. In the DDR model shown in the right side of Fig. 3, $l^t_{n,k,r}$ and $l^t_{-n,k,r}$ represent the hourly un-optimized load of user n and the total load of all the other users (except user n) in cluster k of region r, respectively [7].

There is a local broadcast data network between all users in each cluster. Each user in a cluster solves a local optimization problem and then participates in a game with the other users in the cluster. As users do not have any prior information about each other, they start with some random initial conditions. Then, each user individually solves the local optimization problem. The result is announced to the other users by the broadcasting of a control message over the existing local area network. Each user also updates its local memory whenever it receives a control message from other users. This procedure is repeated until the users converge at the local optimum point and reach the Nash equilibrium point. The output of the DDR optimization process, $l^{*t}_{n,k,r}$, represents the optimal schedule of user n in cluster k of region r. As there are R different regions consisting of N_c different clusters, each with N_r customers, the total optimal hourly load in the power system (L^{*t}_{DDR}) is calculated as follows:

$$L^{*t}_{DDR} = \sum_{r=1}^{R} \sum_{k=1}^{N_c} \sum_{n=1}^{N_k} l^{*t}_{n,k,r}. \tag{12}$$

B. Customer Billing

As mentioned earlier, $C^q(L^t_r)$ given in equ. (2), represents the hourly utility energy cost at each region r. For each user i, in region r let $C^u_{i,r}$ denote the daily customer billing. For the profitability of utility companies, energy sales should be higher than the cost. Therefore, the following equation should

always be satisfied:

$$\sum_{i=1}^{N_r} C_{i,r}^{u} \geq \sum_{t=1}^{T} C^{q}\left(L_r^t\right) \tag{13}$$

Suppose $\partial = \frac{\sum_{i=1}^{N_r} C_{i,r}^{u}}{\sum_{t=1}^{T} C^{q}(L_r^t)}$ represents the profit factor of the utility company. When $\partial > 1$, the utility company profits. When $\partial = 1$, the billing system is budget-balance and the utility company only charges the users its cost of providing energy. Note that users are billed for their total daily energy consumption. The exact amount of user's billing depends on the hourly energy price and the user demand. Thus, the following equation is used to calculate the daily energy cost of each user:

$$C_{i,r}^{u} = \sum_{t=1}^{T} P_t \; l_{i,r}^{t} \tag{14}$$

where $l_{i,r}^{t}$ and P_t are the hourly power consumption of user i, in region r, and the hourly energy price determined by the utility company. Note that P_t is different with a_t used in equ. (2). P_t is the hourly energy sales price (per \$/KW), while a_t is the hourly energy purchase cost coefficient (per \$/(KW)^2). If the utility company wants to make profits, based on the purchase price of energy, the energy sale price (P_t) will be determined. By applying the proposed demand response program, not only the total cost for the utility company decreases, but the cost for customers participating in the demand response program also lowers.

C. CDR Network Architecture

Fig. 4 shows the network architecture of the CDR. This architecture consists of three different parts: the application layer, the control layer, and the data plane layer. The demand response service is an application layer protocol. This centric design helps to develop a central optimal demand response program. In the control layer, the controller utilizes different network functions and the network operating system. The controller consists of the following functions:

- Home Appliances Discovery: It is critical to transfer all appliance information to the edge cloud server in order to design an efficient demand response program. For this purpose, the home appliances should be discovered first. The function of home discovery is discovering and maintaining the status of all home appliances via a specific discovery protocol. Each home appliance informs its gateway of its identity and power consumption information. After this, all home gateways send their gathered information to the cloud server to be stored in the proper database for further processing. Note that this process is performed periodically. Whenever a new home appliance joins or leaves a home network, its information is forwarded to the edge cloud server.

- Virtual Networked Home Energy Management (vNHEM): The physical home energy management systems of the distributed demand response model are connected to the smart meter or combined with a Home Area Network (HAN). They receive the electricity pricing signals from the utility company and perform home automation functions. In the proposed

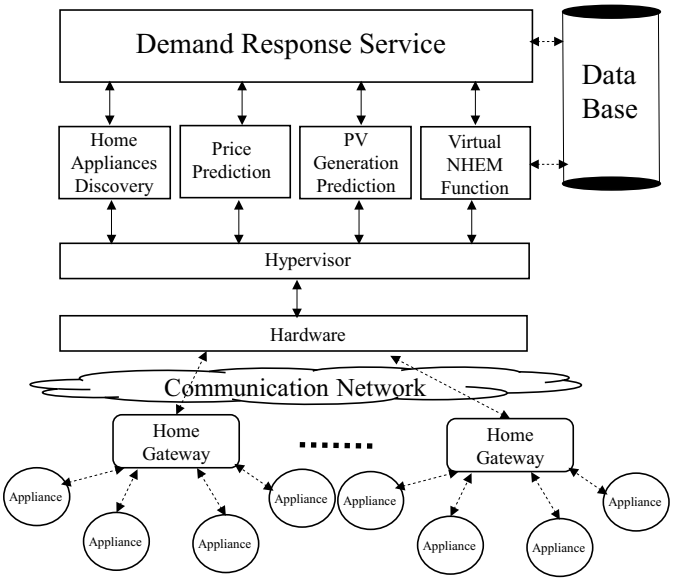


Fig. 4. Edge cloud components.

CDR, these functions can be virtualized at the cloud servers by virtualization techniques. The vNHEM is a virtual implementation of network home management systems. It utilizes the benefits of cloud computing virtualization and provides a combined energy monitoring and management functionality which increases efficiency. Compared to the physical NHEMs, this system is more cost efficient and more flexible.

- Price Prediction: Power demand is highly variable and depends on different parameters, such as time of day (morning, noon or evening), work days or weekends, season and the wholesale market. Therefore, dynamic or real-time pricing is used to set the cost for power. This module is utilized to predict the price of energy in response to market demands.

- PV Energy Generation Prediction: Since the CDR is a day-ahead power consumption optimization, one must know the amount of energy generation by the PV systems used in the microgrids. The most dominant prediction technique is time-series analysis which can be employed to find repeating patterns in the historical data to forecast future values. Consequently, the scaling action is done in advance. The Normalized Least Mean Square (NLMS) predictor is used to predict the PV energy generation.

Fig. 5 shows the time sequence of message transmission between the user, microgrid, and cloud server.

D. Communication Model of DDR

Each user in the DDR model presented in [7] simply plays its best response and locally runs an optimization problem. The new schedule is announced to the other users through a broadcasting of the control message. If the updates of the individual energy consumption scheduling vectors are asynchronous among the users, meaning that users sequentially run the optimization problem, then the algorithm converges at the Nash equilibrium point. It is clear that, by increasing the number of users, the number of required iterations needed to reach the Nash equilibrium also increases. Furthermore,

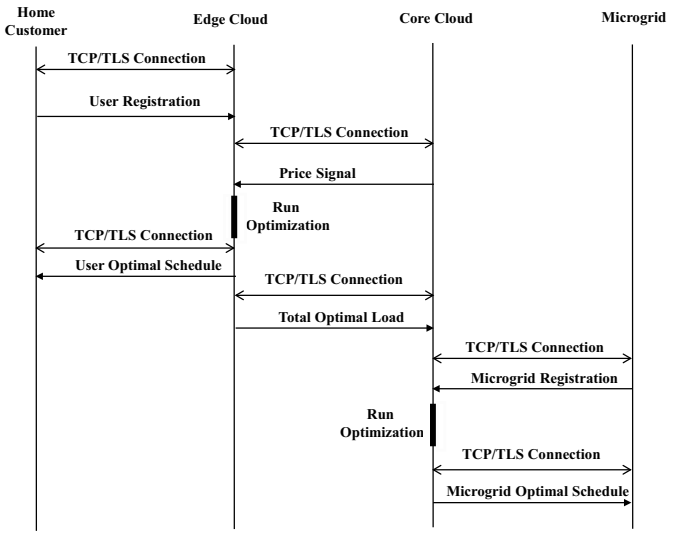


Fig. 5. The time sequence of message transfer in the CDR model.

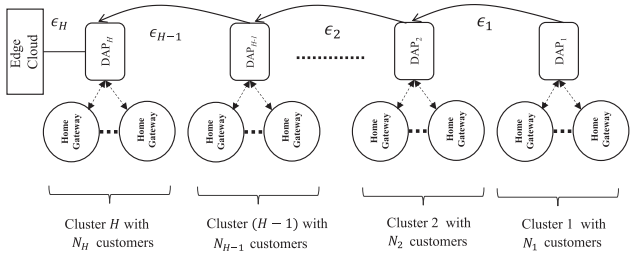


Fig. 6. Wireless mesh communication model for CDR.

if some update messages are lost, not only does the convergence time increase but more network bandwidth is also wasted. On the other hand, when a large number of geographically distributed users participate in the DDR program, providing a local area network connection among all users becomes a challenging task. Therefore, as mentioned earlier, some clusters should be built and users placed in different clusters. Suppose there are N_c different clusters. At each cluster k in region r, there are N_k different users. At each iteration of the algorithm, each user broadcasts his/her locally optimized load to all other users in the cluster with broadcasting a message the size of s_m bits. The total number of required update messages to reaching the optimum point is calculated as follows:

$$N_{DDR}^{k} = N_{it}^{k} N_k (N_k - 1) \tag{15}$$

where N_{it}^{k} is the total number of iterations needed to find the Nash equilibrium point in cluster k. The present study's experimental results confirm that, when the TCP packets transfer the update messages, the average number of iterations is almost equal to $\gamma.N_k$, where $\gamma \in [1.4, 1.9]$.

Note that, when the UDP protocol is used, the required number of iterations increases especially when the bit error rate is high. The total bandwidth required by DDR to find the optimum load in all of the power system's regions is equal to:

$$B_{DDR} = \sum_{r=1}^{R} \sum_{k=1}^{N_c} s_m \; N_{DDR}^{k}. \tag{16}$$

E. Communication Model of CDR

In the CDR model, all customer information is transferred to the edge cloud. The current study assumes that users are grouped in some clusters. A clustered wireless mesh network is considered in which, inside each cluster, some intermediate nodes called Data Aggregation Points (DAPs) aggregate customer data and then forward this data to the edge cloud using hop by hop communication, as shown in Fig. 6. Suppose that, in each region r, there are H_r different DAPs. DAP_j utilizes a wireless channel with a bit error rate equal to $\in_j$. As mentioned earlier, CDR utilizes a communication protocol with M different messages, each with a length of s_m bits. Suppose $\in_k$, represents the bit error rate of the communication channel used in cluster k. If the independent channel model employing the Bernoulli function is considered, then the packet error probability $p_k(s_m)$ for a packet the size of s_m is calculated as $p_k(s_m) = 1-(1- \in_k)^{s_m}$. To provide reliability in packet delivery, the current work uses the Automatic-Repeat-reQuest (ARQ) error recovery technique to transfer the packet between users in the system. When the sender transmits a packet on the channel, the receiver checks it for possible errors. If there are no errors, the receiver acknowledges the correct transmission by sending an Acknowledge (ACK) signal. If the receiver does not receive the ACK signal due to some problems in the channel, the receiver retransmits it. It can be easily seen that, at each cluster k with bit error rate $\in_k$, for a packet the length of s_m, the average number of retransmissions is equal to $\frac{p_k(s_m)}{1-p_k(s_m)}$.

To transfer data packets from the source user to the edge cloud through some intermediate DAPs, the following three different transmission strategies were considered.

- Hop by Hop (HBH): In the HBH model, all intermediate nodes on the source-to-destination path need to buffer all unacknowledged packets and retransmit lost packets upon request. This approach provides a low delay, but incurs very high costs in terms of memory usage. Also, all intermediate nodes need to process, buffer, and manage all packets they receive. Let $N_{HBH}(i, H, s_m)$ represents the total number of transmissions of a packet the size of s_m from source node i to the destination node (edge cloud) through H intermediate hops. $N_{HBH}(i, H, s_m)$ is computed as follows:

$$\begin{aligned} N_{HBH}(i, H, s_m) &= \sum_{k=i}^{i+H-1} \frac{1}{1-p_k(s_m)} \\ &= \sum_{k=i}^{i+H-1} (1- \in_k)^{-s_m} \end{aligned} \tag{17}$$

- End-to-end (e2e): In the e2e approach, the source node buffers unacknowledged packets and retransmits lost packets upon request. This approach minimizes memory usage because only the source node needs to buffer and retransmits lost packets. However, the delay can be very high due to retransmission with long waiting times. Let $N_{e2e}(i, H, s_m)$ represent the number of transmission for a packet the size of s_m from

source node i to the destination node (edge cloud) through H intermediate hops. $N_{e2e}(\mathrm{i}, \mathrm{H}, s_m)$ is calculated as follows:

$$N_{e2e}(i, H, s_m) = \frac{H}{1 - \sum_{j=i}^{i+H} p_j(s_m) \prod_{k=i}^{j-1} \left(1 - p_k(s_m)\right)} \tag{18}$$

- Intermediate Caching (IC): This approach offers a trade-off between the e2e and HBH approaches. Only a subset of intermediate nodes on the source-to-destination path are selected as cache nodes, which when requested, buffer unacknowledged packets for future retransmissions. The system does not perform end-to-end loss recovery when using intermediate caching. Instead, one or more intermediate nodes from a source-to-destination path are selected as cache points. These cache points buffer packets they receive and retransmit lost packets requested by the destination. Critical for real-time applications, the intermediate cache points help to minimize loss recovery time and end-to-end delay. As shown in Fig. 7, we suppose that, at each region $r \in \{1, \ldots, R\}$, there are $\mathcal{C}_r$ different cache points which have a Δ_r hop distance from each other. The following equation should be satisfied:

$$H_r = (\mathcal{C}_r - 1)\Delta_r + 1 \tag{19}$$

Let $N_{IC}(i, H, s_m)$ represent the total number of packet transmissions the size of s_m from source node i to the destination (edge cloud) through H intermediate hops. $N_{IC}(i, H, s_m)$ is calculated as follows.

1) When $i < H_r$:

$$N_{IC}(\mathrm{i}, \mathrm{H}, s_m) = N_{e2e}(\mathrm{i}, \Delta_r - A, s_m) + N_{HBH}(H_r, 1, s_m) + \sum_{k=B+2}^{\mathcal{C}_r - 1} N_{e2e}((\mathrm{k} - 1)\Delta_r + 1, \Delta_r, s_m) \tag{20}$$

where A and B are defined as follows:

$$A = mod(i - 1, \Delta_r) \tag{21}$$

$$\mathrm{B} = floor\left(\frac{i - 1}{\Delta_r}\right). \tag{22}$$

2) When $i = H_r$:

$$N_{IC}(\mathrm{i}, \mathrm{H}, s_m) = N_{HBH}(H_r, 1, s_m) \tag{23}$$

It can be proven that, for the IC strategy, when all intermediate nodes are cache points ($\Delta_\mathrm{r} = 1$), the IC strategy is equal to HBH. For the CDR model using HBH, e2e or IC reliability models, the total bandwidth required to transfer M packets, each the size of s_m bits, from all sources in all regions is calculated as follows:

$$B_{CDR} = \sum_{r=1}^{R} \sum_{j=1}^{H_r} N_{rj} \sum_{m=1}^{M} s_m \, N_{method}(j, H_r - j + 1, s_m) \tag{24}$$

where N_{rj} and H_r represent the number of users in the jth DAP in region r and the number of DAPs in region r, respectively. N_{method} refers to the delivery method which can be the HBH, e2e or IC given in equations (17), (18), (20) or (23), respectively.

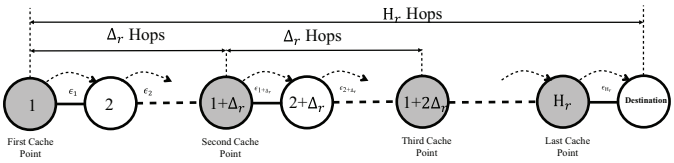


Fig. 7. Intermediate caching model.

F. Convergence Time

In this section, the convergence time of both CDR and DDR models are evaluated. As DDR with the UDP protocol is not able to find the optimal solution, especially when the bit error rate is high, only DDR with the TCP protocol is considered. Suppose N_{it}, t_{opt}, N_{ret} and RTT represent the following: the number of required convergence iterations, the time required to run the optimization process in each iteration, the number of update TCP packets, and the round trip time for sending a TCP packet and receiving its acknowledgment message, respectively. The convergence time of DDR (D_{DDR}) can be evaluated as follows:

$$D_{DDR} = N_{it}\left(t_{opt} + N_{ret} RTT\right) \tag{25}$$

The current study's experimental results confirm that N_{ret} can be estimated as $N_{ret} = \gamma \mathrm{N}_k HBH(1, 1, s_m)$ where $\gamma \in [1.4, 1.9]$.

For the CDR model, the convergence time (D_{CDR}) is computed as follows:

$$D_{CDR} = T_{opt} + 2T_{com} \tag{26}$$

where T_{opt} and T_{com} are the central optimization process time and the end-to-end delay required to transfer all user information to the edge cloud or from edge cloud to all users, respectively. T_{com} is computed as follows:

$$T_{com} = T_{trans} + T_{prop} + T_{proc} + T_{queue} \tag{27}$$

where $T_{trans}, T_{prop}, T_{proc}$ and T_{queue} are the total delay required for the transmission, propagation, processing, and queuing of all messages in the path, respectively. It is assumed that the processing and queuing delay are negligible. As HBH strategy has the best bandwidth performance, T_{com} of the HBH for the worst case is computed as follows:

$$T_{com} = \sum_{m=1}^{M} N_{HBH}(1, H_r, s_m)\left(t_{trans} + t_{prop}\right) \tag{28}$$

where $N_{HBH}(1, H_r, s_m)$ is the number of required retransmissions of a packet the length of s_m from source cluster 1 through a path with H_r hops. $t_{trans} = \frac{s_m}{Link\ Bandwidth}$ and $t_{prop} = \frac{length\ of\ link}{propagation\ speed}$ are the packet transmission delay and per hop propagation delay, respectively.

IV. Simulation Results

In this section, computer simulation is employed to evaluate and compare the performance of the DDR model given in [7] and the proposed CDR, in terms of both the power system and communication network performance. Without going into detail, a power system is considered with only one

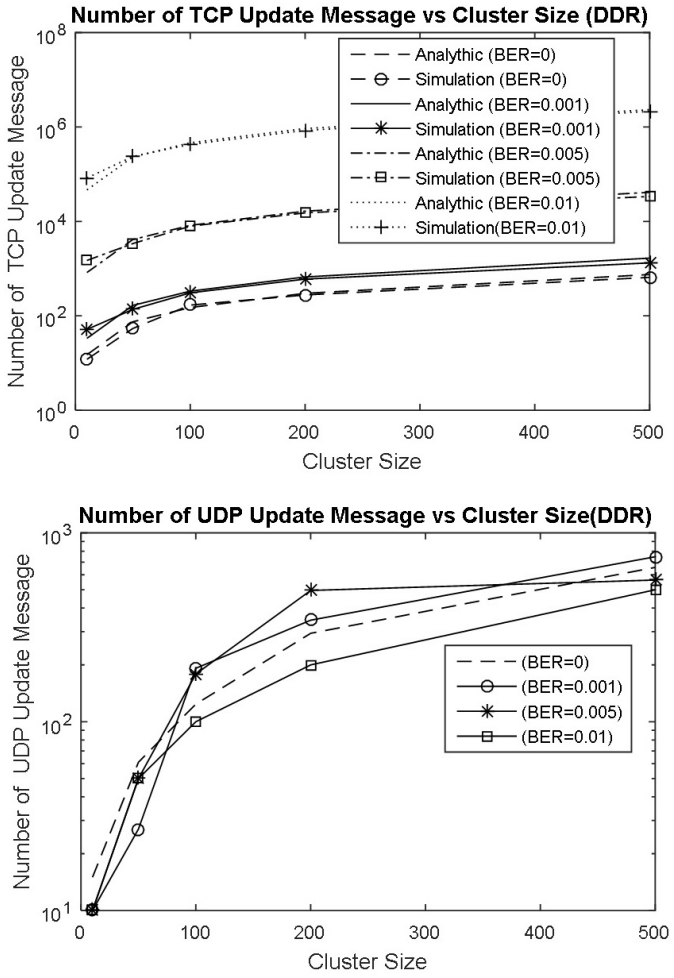


Fig. 8. Number of update messages versus cluster size at different BER.

region with 1000 customers and one PV system equipped with 2000 square meters of solar cells. Random loads for all users in the region are generated. The packet size (including all underlying headers) for DDR and CDR models are set to 100 and 150 bytes, respectively. The maximum distance between users in each cluster is 100 meters and the distance between DAPs is equal to 200 m. The wireless link bandwidth is assumed to be 1Mb/s. All simulation trials were repeated ten times and depicted the average results in the figures.

A. Power System Performance

In the first scenario for the DDR model with TCP and UDP protocols, the number of required update messages at different values of the Bit Error Rate (BER) and cluster size is evaluated. The results are given in Fig. 8. As seen, by increasing the cluster size and bit error rate, the number of required update messages also increases. When UDP is used, as there is not any retransmission, the number of update messages is less than that of the TCP. When the bit error rate is high (BER=0.01), most UDP packets are lost. In this case, as the users do not receive any update messages from others, after one cycle (which is equal to the cluster size), the optimization iteration is stopped without having reach the optimal solution.

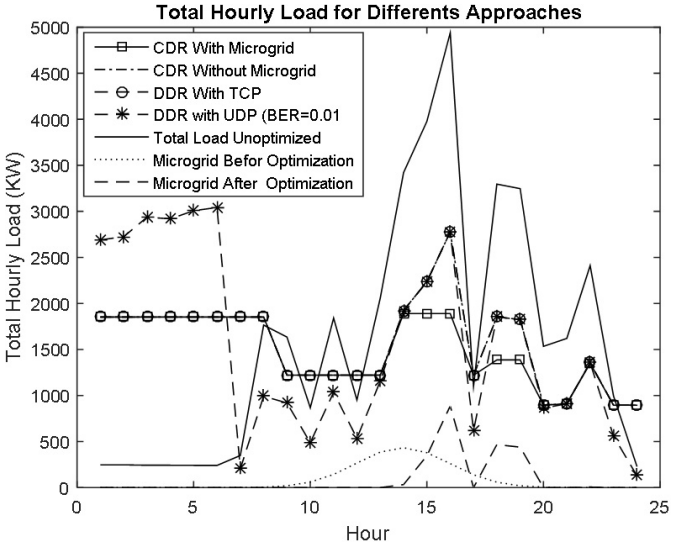


Fig. 9. The total power grid load for different approaches.

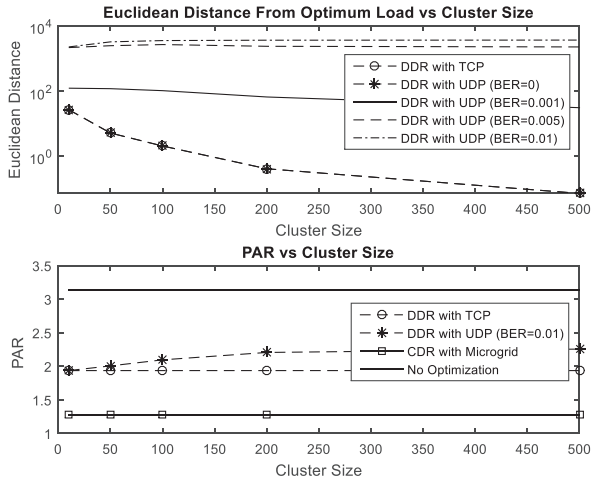


Fig. 10. Cost reduction and PAR improvement.

The next scenario evaluates the power system performance of both DDR and CDR in terms of PAR and Euclidean distance from the optimal load. The Euclidean distance is defined as $d(X, \hat{X}) = \sqrt[2]{\sum_{i=1}^{T} (X_i - \hat{X}_i)^2}$ where X and $\hat{X}$ are the output of the DDR model and the optimal load, respectively. In Fig. 9, for both DDR and CDR, the optimized load is plotted versus time. As shown in this figure, both CDR and DDR with TCP can effectively find the optimal load and shift the power consumption from a high price time to a low price time. However, the DDR with UDP protocol is not able to find the optimal load because, when the bit error rate is high, most update messages are lost and the algorithm can not converge at the optimal solution. In Fig. 10, for DDR, the PAR and Euclidean distance from the optimum load are given. The results shown in this figure confirm that, by increasing cluster size (decreasing the number of clusters), DDR is approaching the optimal solution. Furthermore, as seen, CDR with a microgrid can significantly lower the PAR of the power system.

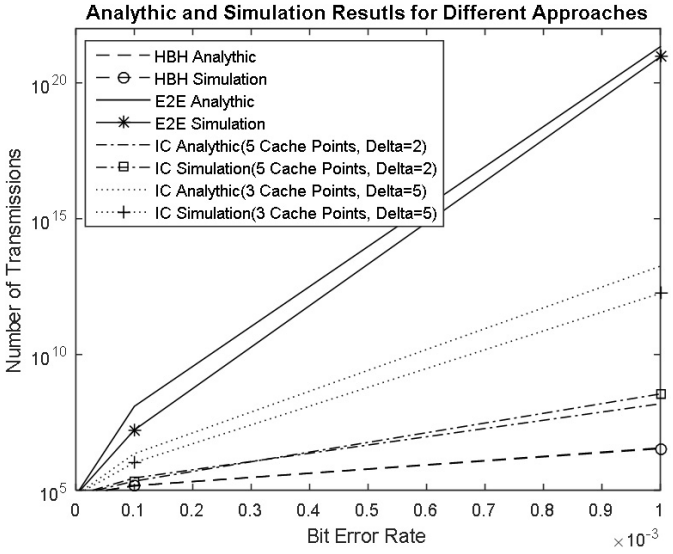


Fig. 11. The bandwidth performance of HBH, e2e, and IC at different values of the channel bit error.

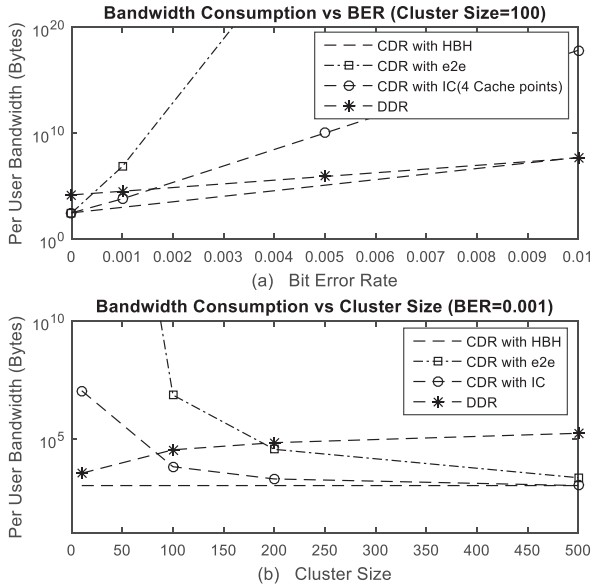


Fig. 12. The bandwidth performance of CDR and DDR at different values of the cluster size and channel bit error rate.

B. Communication Performance

This section investigates the communication performance of DDR and CDR in terms of the required bandwidth. By comparing the analytical and the simulation results, the first scenario investigates the accuracy of the HBH, e2e, and IC analytical models for a region with 110 users in 11 equal clusters, each with ten users and with a packet equal to 100 bytes. The results are shown in Fig. 11. In this figure, for all models, the number of required packet transmissions is plotted versus the channel bit error rate. As seen, the simulation results validate the accuracy of the analytic model.

Furthermore, as expected, it can be observed that, for all models, by increasing the channel bit error rate, the number of packet transmissions also increases. The performance of the IC model is always between that of the HBH and e2e methods. It can be seen that, with an increase in the number of intermediate caching nodes, the required transmissions of the IC model decreases. The HBH model, in which all intermediate nodes cache incoming packets, has the best bandwidth performance.

The next trial for both DDR with TCP and CDR evaluates the bandwidth required to converge at the optimum solution. The results are depicted in Fig. 12. As presented in Fig. 12 (a), for all models, by increasing the channel bit error rate, the required bandwidth also increases. It is clear that the CDR with the HBH model always shows the best performance. When the channel bit error rate is low, the e2e and IC mechanisms outperform DDR. As seen, by increasing the bit error rate, the required bandwidth increases. When the bit error rate is high, DDR has a better bandwidth performance than that of e2e and IC models. In Fig. 12 (b), for a communication channel with a bit error rate equal to 0.001, the required bandwidth is plotted versus cluster size. The results confirm that, by increasing the cluster size, the required DDR bandwidth also increases. Because the total number of users is fixed, when the cluster size increases, the number of intermediate DAPs decreases. This is the reason why the required bandwidth of CDR decreases.

C. Effect of Packet Size

The main difference between DDR and CDR is related to optimization problem. In DDR, this is performed locally at each customer smart meter. However, in CDR all customer information should first be transferred to the edge cloud server and then the central optimization problem is performed. Therefore, the DDR and CDR packet size are not the same. According to line 7 of Algorithm 1 in the reference model [7], after each iteration, each customer should broadcast the value of the locally optimized load to all other users in the same cluster. As shown in Fig. 13, the DR messages are encapsulated in the TCP, IP, and data link layers. Each layer adds a minimum header to the message which increases the message size. The minimum and maximum header size of TCP and IP packets are 20 and 60 bytes, respectively. For the IEEE 802.11 wireless LAN protocol, the header frame size (including preamble) for PLCP (Physical Layer Convergence Protocol) and MAC (Medium Access Control) are 22 and 34 bytes, respectively. The present study, allocates 4 bytes to $l^t_{-n,k,r}$. Thus, the minimum packet size for the DDR update message is equal to 4 ($l^t_{-n,k,r}$) +20 (TCP header) +20 (IP header) +34 (IEEE 802.11 MAC frame header) +22 (PLCP header) = 100 bytes. Consequently, the DDR update message is set to 100 bytes.

As the optimization problem in CDR is run on the cloud server, it is necessary to transfer the number of active appliances and also their energy profile consumption to the cloud server. Each new appliance connecting to the home network should be registered in the server. When the registration phase is completed, the central controller at the server has complete information about the number and type of appliances currently used in each customer's home. This data is stored in a particular database. Note that CDR is based on active appliances

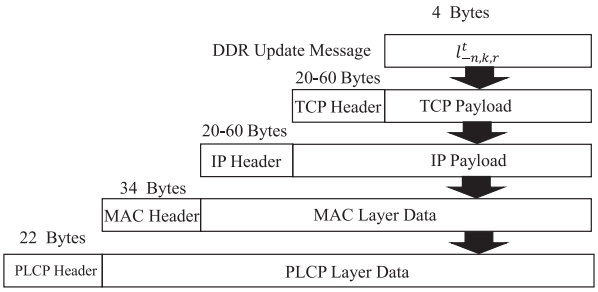


Fig. 13. Packet encapsulation.

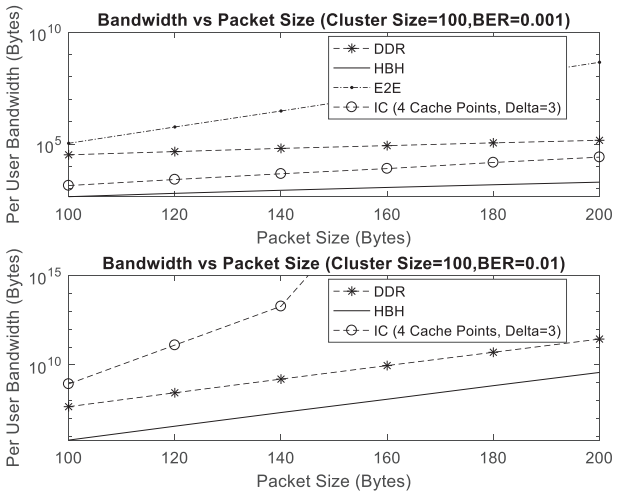


Fig. 14. Bandwidth requirement at different packet size and BER.

which are currently used in the home network. The current work sets the CDR message to 150 bytes. After subtracting the TCP, IP, and MAC header, only 52 bytes remains for informing the cloud server of each customer load information. Thus, almost 12 different shiftable appliances can be supported (each appliance requires almost 4 bytes). The remaining 4 bytes is assigned to the fixed load information. When the number of shiftable appliances is less than 12, the CDR message size is less than 150 bytes, thus consuming less bandwidth. Therefore, allocating 150 bytes, to the CDR packet size is the worst case scenario and, with a smaller message size, bandwidth consumption decreases. However, the present study evaluates the bandwidth requirements of DDR and CDR at the different values of message size and bit error rate. Fig. 14 presents the bandwidth performance results. As observed, by increasing the message size, the bandwidth requirement of both approaches increases as well. Note that, when the bit error rate increases, more packets are lost and so the necessary bandwidth increases as well.

D. Convergence Time

This subsection, investigates the scalability of the DDR (with TCP) and CDR models in terms of the convergence time of the optimization process. The results shown in Fig. 15 confirm the main cause of the convergence time's increase. That is, when the bit error rate increases, as more messages are lost, the number of retransmissions increases. Fig. 15 (a)'s

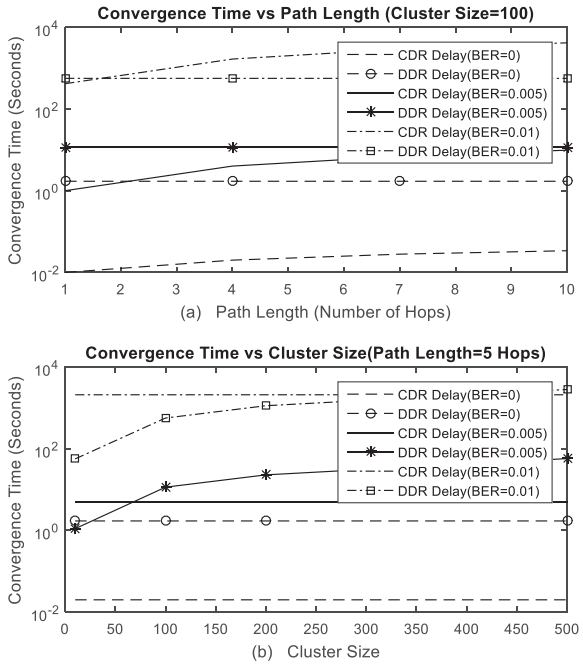


Fig. 15. Convergence time of DDR and CDR.

results indicate that the convergence time of DDR is independent of the path length. However, as Fig. 15 (b) demonstrates, by increasing the cluster size and bit error rate, the DDR convergence time increases. These results confirm that, when the bit error rate and path length are low or when the cluster size is high, the CDR has a better convergence time performance than the DDR does.

E. Cost-Effectiveness Analysis

The main difference between HBH, e2e, and IC is the cache memory requirements. For HBH, all intermediate nodes on the source-to-destination path should buffer all unacknowledged packets and retransmit lost packets upon request. This approach provides the shortest packet end-to-end delay, but incurs very high costs in terms of memory usage. In addition, all intermediate nodes need to process, buffer, and manage all packets they receive. In the e2e approach, the source buffers unacknowledged packets and retransmits lost packets upon requests. This approach minimizes memory usage because only the source needs to buffer and retransmit lost packets. However, the end-to-end packet delay can be very high due to long waiting time retransmissions. The IC offers a trade-off between the e2e and HBH approaches. Only a subset of intermediate nodes on the path is selected as cache nodes which buffer unacknowledged packets for future retransmissions when requested.

The current study considers the required number of cache points as the communication cost. The Memory Requirement Index (MRI) is also defined as the number of cache nodes divided by the total number of nodes on the path from the source to the edge cloud server, as follows:

$$MRI_r = \frac{NC_r}{H_r} \tag{29}$$

TABLE I
COST-EFFECTIVENESS ANALYSIS

BER=0.001, Number of Hops=13						
Cache Points	Cache Distance	Memory Requirement Index	Convergence Time (seconds)			
			DDR			CDR
C_r	Δ_r	MRI	Cluster Size =100	Cluster Size =300	Cluster Size =500	
1	13	0.0769	4.9	44.9	125	1.8e5
3	6	0.2308	4.9	44.9	125	41.1
4	4	0.3077	4.9	44.9	125	3.99
5	3	0.3846	4.9	44.9	125	1.28
7	2	0.5385	4.9	44.9	125	0.425
13	1	1	4.9	44.9	125	0.15

where NC_r and H_r represent the number of cache points and number of hops in region r.

Based on equ. (26), the convergence time of the demand response program is directly related to the communication delay which depends on the type of retransmission strategy. Therefore, MRI is considered as the communication cost and the convergence delay is thought of as the demand response performance. Obviously, there's a trade-off between the cost of the communication infrastructure and the performance of the demand response. Based on equ. (28), the convergence delay is directly related to the number of retransmissions. By increasing the communication cost (increasing the MRI), the number of required retransmissions decreases which improves the convergence time of the demand response program. On the other hand, by decreasing the communication cost (decreasing the MRI), the number of required retransmissions and also the convergence time increases as well.

To evaluate the cost-effectiveness of the proposed work, a network with 1300 customers and 13 different hops is considered. The bit error rate is set to 0.001. The IC model is considered with different values of cache points and cache distances. The results are given in Table I.

As seen, by increasing the communication cost (MRI), the demand response performance (convergence time) improves. Note that CDR is independent of cluster size while DDR performance lowers by increasing cluster size. More iterations are needed to find the optimum solution which increases convergence time as well. It is easy to prove that the IC model with one cache point ($\mathcal{C}_r = 1$) converts e2e and one with a cache distance equal to 1($\Delta_r = 1$) converts to the HBH mechanism.

F. Piecewise Linear Cost Function

As energy supply and demand are variable during daylight hours, the cost of energy is not fixed and changes with time. This means that the price of the same amount of energy can vary at different times of the day. As discussed in [7], the energy cost function should be increasing and strictly convex. Utility companies have two options when providing the necessary energy for their customers. They can install some power generators to produce necessary energy or buy energy from the market. The quadratic cost function has been widely used for the cost of power produced by conventional generators, such as thermal, and also for the cost of buying energy from the market [7], [14], [24]–[30]. The quadratic cost functions for thermal units are based on 'input-output' characteristics of thermal plants which have unique input-output characteristics. In any case, the amount of power generated and the total cost of this generation rise non-linearly. As mentioned earlier, customers are billed with a linear cost function. Note that, in the present study's hierarchical model, a thermal generator at each region is proposed which generates the energy required by all customers in the region.

TABLE II
CDR AND DDR PERFORMANCE WITH LINEAR AND QUADRATIC COST FUNCTIONS

Performance Metric	Cost Function	N_c		
		50	100	200
PAR DDR	Piecewise Linear	2.06	2.05	2.04
	Linear	1.95	1.87	1.83
	Quadratic	1.39	1.37	1.37
PAR CDR	Piecewise Linear	1.73	1.72	1.71
	Linear	1.78	1.74	1.72
	Quadratic	1.38	1.37	1.36
PAR Improvement	Piecewise Linear	0.33	0.33	0.33
	Linear	0.17	0.13	0.11
	Quadratic	0.01	0.00	0.01
DDR Iterations	Piecewise Linear	4517	9031	18061
	Linear	142	318	436
	Quadratic	99	196	393

However, some utility companies, such as British Columbia (BC) Hydro in Canada, uses a step piecewise linear function instead of the quadratic cost function (see [7, Fig. 4]). This subsection evaluates the performance of CDR and DDR using two steps piecewise linear function at the core and edge cloud servers. Suppose that $C^l(L_r^t)$, represents the total energy cost at each region r. $C^l(L_r^t)$ is defined using the following two steps piecewise linear cost function.

$$C^l(L_r^t) = \begin{cases} pr_t^1 L_r^t & \text{if } L_r^t \le B_1 \\ pr_t^1 B_1 + pr_t^2 (L_r^t - B_1) & \text{if } L_r^t > B_1 \end{cases} \tag{30}$$

where pr_t^1 and pr_t^2 are the hourly price of the first and second steps, respectively ($pr_t^2 > pr_t^1$). B_1 represents the first power consumption block in the load curve. By solving the optimization problem (3) for both linear ($B_1 \to \infty$), two steps piecewise linear ($B_1 < \infty$) and quadratic cost functions, the performance of CDR and DDR at different values of cluster size (N_c) is obtained. The results are shown in Table II.

The results confirm that 1) both CDR and DDR with quadratic cost functions outperform those with linear and piecewise linear cost functions in terms of PAR and the number of iterations. Also, 2) by increasing the cluster size for both approaches the PAR improves. 3) The results confirm that the quadratic cost function can arrive at the optimum solution faster than the linear function can. 4) Piecewise linear cost function needs more iterations to converge to the optimal solution than linear and quadratic cost functions.

V. Conclusion

In this paper, two different types of demand response were studied: distributed (DDR) and cloud-based (CDR). It was shown that utilizing multi-tier cloud computing based on a software-defined infrastructure can provide a high level of scalability and reliability and also improve the performance of both the power network and communication network. The present study investigated the bandwidth requirements as well as the convergence time of both approaches. DDR is based on iterations, in which, at each iteration, an update message is broadcast between all users in the same cluster. After some iterations, the algorithm converges at the optimum solution. When the communication channel is not ideal, some update messages are lost. This not only increases the convergence time but also requires more communication bandwidth. The current research demonstrated that when DDR leverages an unreliable UDP protocol, some UDP packets may become lost, thus making the optimal solution not feasible. Unlike the distributed approaches, in the cloud-based demand response, the optimization process is run centrally which is independent of the communication channel. The present study provides a cost-effective analysis and proves that, by increasing the communication cost, better demand response performance can be achieved. Also, three different communication strategies for transferring user consumption information to the cloud server were presented. Simulation results confirmed that the proposed cloud-based demand response reduces the total cost, peak-to-average load ratio, and the convergence time of the optimization process while utilizing the communication bandwidth more effectively.

References

[1] N. Ruiz, I. Cobelo, and J. Oyarzabal, "A direct load control model for virtual power plant management," *IEEE Trans. Power Syst.*, vol. 24, no. 2, pp. 959–966, May 2009.

[2] P. Centolella, "The integration of price responsive demand into regional transmission organization (RTO) wholesale power markets and system operations," *Energy*, vol. 35, no. 4, pp. 1568–1574, 2010.

[3] K. Herter, "Residential implementation of critical-peak pricing of electricity, "*Energy Policy*, vol. 35, no. 4, pp. 2121–2130, 2007.

[4] P. Palensky and D. Dietrich, "Demand side management: Demand response, intelligent energy systems, and smart loads," *IEEE Trans. Ind. Informat.*, vol. 7, no. 3, pp. 381–388, Aug. 2011.

[5] Y. Wang, X. Ai, Z. Tan, L. Yan, and S. Liu, "Interactive dispatch modes and bidding strategy of multiple virtual power plants based on demand response and game theory," *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 510–519, Jan. 2016.

[6] R. Deng, Z. Yang, M.-Y. Chow, and J. Chen, "A survey on demand response in smart grids: Mathematical models and approaches," *IEEE Trans. Ind. Informat.*, vol. 11, no. 3, pp. 570–582, Jun. 2015.

[7] A.-H. Mohsenian-Rad, V. W. S. Wong, J. Jatskevich, R. Schober, and A. Leon-Garcia, "Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid," *IEEE Trans. Smart Grid*, vol. 1, no. 3, pp. 320–331, Dec. 2010.

[8] Z. Tari, X. Yi, U. S. Premarathne, P. Bertok, and I. Khalil, "Security and privacy in cloud computing: Vision, trends, and challenges," *IEEE Cloud Comput.*, vol. 2, no. 2, pp. 30–38, Mar./Apr. 2015.

[9] H. Liu, H. Ning, Q. Xiong, and L. T. Yang, "Shared authority based privacy-preserving authentication protocol in cloud computing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 1, pp. 241–251, Jan. 2015.

[10] M. Jawurek, F. Kerschbaum, and G. Danezis, *SoK: Privacy Technologies for Smart Grids—A Survey of Option*s, Microsoft Res., Cambridge, U.K., Nov. 2012.

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

[12] Z. Tan, P. Yang, and A. Nehorai, "An optimal and distributed demand response strategy with electric vehicles in the smart grid," *IEEE Trans. Smart Grid*, vol. 5, no. 2, pp. 861–869, Mar. 2014.

[13] A. Safdarian, M. Fotuhi-Firuzabad, and M. Lehtonen, "A distributed algorithm for managing residential demand response in smart grids," *IEEE Trans. Ind. Informat.*, vol. 10, no. 4, pp. 2385–2393, Nov. 2014.

[14] R. Deng, Z. Yang, F. Hou, M.-Y. Chow, and J. Chen, "Distributed real-time demand response in multiseller–multibuyer smart distribution grid," *IEEE Trans. Power Syst.*, vol. 30, no. 5, pp. 2364–2374, Sep. 2015.

[15] A. Safdarian, M. Fotuhi-Firuzabad, and M. Lehtonen, "Optimal residential load management in smart grids: A decentralized framework," *IEEE Trans. Smart Grid*, vol. 7, no. 4, pp. 1836–1845, Jul. 2016.

[16] R. Deng, G. Xiao, R. Lu, and J. Chen, "Fast distributed demand response with spatially and temporally coupled constraints in smart grid," *IEEE Trans. Ind. Informat.*, vol. 11, no. 6, pp. 1597–1606, Dec. 2015.

[17] L. Zheng and L. Cai, "A distributed demand response control strategy using Lyapunov optimization," *IEEE Trans. Smart Grid*, vol. 5, no. 4, pp. 2075–2083, Jul. 2014.

[18] W. Shi, N. Li, X. Xie, C.-C. Chu, and R. Gadh, "Optimal residential demand response in distribution networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 7, pp. 1441–1450, Jul. 2014.

[19] S. N. A. U. Nambi, E. Pournaras, and R. V. Prasad, "Temporal self-regulation of energy demand," *IEEE Trans. Ind. Informat.*, vol. 12, no. 3, pp. 1196–1205, Jun. 2016.

[20] P.-Y. Kong, "Wireless neighborhood area networks with QoS support for demand response in smart grid," *IEEE Trans. Smart Grid*, vol. 7, no. 4, pp. 1913–1923, Jul. 2016.

[21] M. Pruckner, A. Awad, and R. German, "A study on the impact of packet loss and latency on real-time demand response in smart grid," in *Proc. IEEE Globecom Workshops*, Anaheim, CA, USA, 2012, pp. 1486–1490.

[22] C. Yang and W. Lou, "On optimizing demand response management performance for microgrids under communication unreliability constraint," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, San Diego, CA, USA, 2015, pp. 1–6.

[23] L. Zheng, N. Lu, and L. Cai, "Reliable wireless communication networks for demand response control," *IEEE Trans. Smart Grid*, vol. 4, no. 1, pp. 133–140, Mar. 2013.

[24] A. J. Wood and B. F. Wollenberg, *Power Generation, Operation, and Control*. New York, NY, USA: Wiley, 1996.

[25] A.-H. Mohsenian-Rad, V. W. S. Wong, J. Jatskevich, and R. Schober, "Optimal and autonomous incentive-based energy consumption scheduling algorithm for smart grid," in *Proc. Innov. Smart Grid Technol. (ISGT)*, Gothenburg, Sweden, Jan. 2010, pp. 1–6.

[26] B. Moradzadeh and K. Tomsovic, "Two-stage residential energy management considering network operational constraints," *IEEE Trans. Smart Grid*, vol. 4, no. 4, pp. 2339–2346, Dec. 2013.

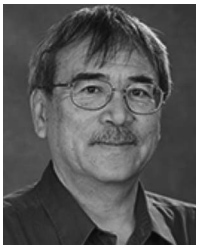
[27] R. Zhou, Z. Li, C. Wu, and M. Chen, "Demand response in smart grids: A randomized auction approach," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 12, pp. 2540–2553, Dec. 2015.

[28] S. Bahrami and A. Sheikhi, "From demand response in smart grid toward integrated demand response in smart energy hub," *IEEE Trans. Smart Grid*, vol. 7, no. 2, pp. 650–658, Mar. 2016.

[29] C. Eksin, H. Deliç, and A. Ribeiro, "Demand response management in smart grids with heterogeneous consumer preferences," *IEEE Trans. Smart Grid*, vol. 6, no. 6, pp. 3082–3094, Nov. 2015.

[30] I. Atzeni, L. G. Ordóñez, G. Scutari, D. P. Palomar, and J. R. Fonollosa, "Demand-side management via distributed energy generation and storage optimization," *IEEE Trans. Smart Grid*, vol. 4, no. 2, pp. 866–876, Jun. 2013.

Mohammad Hossein Yaghmaee (SM'10) received the B.S. degree in communication engineering from the Sharif University of Technology, Tehran, Iran, in 1993, and the M.S. degree in communication engineering and the Ph.D. degree in communication engineering from the Tehran Polytechnic (Amirkabir) University of Technology, in 1995 and 2000, respectively. He has been a Computer Network Engineer with several networking projects in Iran Telecommunication Research Center since 1992. From 1998 to 1999, he was with the Network Technology Group, C&C Media Research Laboratory, NEC Corporation, Tokyo, Japan, as a Visiting Research Scholar. From 2007 to 2008, he was with the Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, USA, as a Visiting Associate Professor. From 2015 to 2016, he was with the Electrical and Computer Engineering Department, University of Toronto, as a Visiting Professor. He is currently a Full Professor with the Computer Engineering Department, Ferdowsi University of Mashhad. He is the Head of the IP-PBX type approval laboratory with the Ferdowsi University of Mashhad. He has authored some books on Smart Grid, TCP/IP, and Smart City in Persian language. His research interests are in smart grid, computer and communication networks, quality of services, software defined networking, and network function virtualization.



Alberto Leon-Garcia received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1973, 1974, and 1976, respectively. He was the Founder and the CTO of AcceLight Networks, Ottawa, ON, Canada, from 1999 to 2002, which developed an all-optical fabric multiterabit, switch. He is currently a Professor in Electrical and Computer Engineering with the University of Toronto, ON, Canada. He holds the Canada Research Chair in Autonomic Service Architecture. He holds several patents and has published extensively in the areas of switch architecture and traffic management. His research team is currently developing a network testbed that will enable at-scale experimentation in new network protocols and distributed applications. He is recognized as an Innovator in networking education. In 1986, led the development of the University of Toronto-Northern Telecom Network Engineering Program. He has also led in 1997 the development of the Master of Engineering in Telecommunications program, and the communications and networking options in the undergraduate computer engineering program. He has authored the leading textbook entitled *Probability and Random Processes for Electrical Engineering and Communication Networks: Fundamental Concepts and Key Architecture*. His current research interests include application-oriented networking and autonomic resources management with a focus on enabling pervasive smart infrastructure. He was a recipient of the 2006 Thomas Eadie Medal from the Royal Society of Canada and the 2010 IEEE Canada A. G. L. McNaughton Gold Medal for his contributions to the area of communications. He is a fellow of the Engineering Institute of Canada.

Morteza Moghaddassian received the M.S. degree in computer engineering from the Ferdowsi University of Mashhad, Mashhad, Iran, in 2015. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Toronto, ON, Canada. He was a Research Assistant with the Center of Excellence on Soft Computing and Intelligent Information Processing, Ferdowsi University of Mashhad, Mashhad, Iran, where he successfully managed a research team to develop a social energy network application for the purpose of metering with built-in demand response management modules, called "SocioGrid." His research interests include cloud computing, edge computing, resource monitoring and management, SG, and optimization of communication networks and open networking. He was a recipient of the Edward S. Rogers Sr. Departmental Graduate Fellowship.