Cloud-Based Charging Management of Heterogeneous Electric Vehicles in a Network of Charging Stations: Price Incentive vs. Capacity Expansion

Cuiyu Kong, *Student Member, IEEE,* Bhaskar P. Rimal, *Senior Member, IEEE,* Martin Reisslein, *Fellow, IEEE,* Martin Maier, *Senior Member, IEEE,* Islam Safak Bayram, *Senior Member, IEEE,* and Michael Devetsikiotis, *Fellow, IEEE*

***Abstract*—This paper presents a novel cloud-based charging management system for electric vehicles (EVs). Two levels of cloud computing, i.e., local and remote cloud, are employed to meet the different latency requirements of the heterogeneous EVs while exploiting the lower-cost computing in remote clouds. Specifically, we consider time-sensitive EVs at highway exit charging stations and EVs with relaxed timing constraints at parking lot charging stations. We propose algorithms for the interplay among EVs, charging stations, system operator, and clouds. Considering the contention-based random access for EVs to a 4G Long-Term Evolution network, and the quality of service metrics (average waiting time and blocking probability), the model is composed of: queuing-based cloud server planning, capacity planning in charging stations, delay analysis, and profit maximization. We propose and analyze a *price-incentive method* that shifts heavy load from peak to off-peak hours, a *capacity expansion method* that accommodates the peak demand by purchasing additional electricity, and a hybrid method of prince-incentive and capacity expansion that balances the immediate charging needs of customers with the alleviation of the peak power grid load through price-incentive based demand control. Numerical results demonstrate the effectiveness of the proposed methods and elucidate the tradeoffs between the methods.**

***Index Terms*—Cloud Computing, Electric Vehicles, Charging Management, Quality of Service, 4G Long-Term Evolution Network**

I. Introduction

A. Motivation

Electric vehicles (EVs) have emerged as an attractive and viable solution to decrease greenhouse gas (GHG) emissions and reduce reliance on fossil fuels. Over the last few years we have witness a renewed push toward EVs as the governments in the United Kingdom and France have announced to ban the sale and use of gasoline and diesel cars by 2030 and more countries, including China, India, Netherlands, and Norway, have committed to phase out such vehicles in the near future and to promote the deployment of EVs. In addition, major auto manufacturers are investing billions of US dollars into the EV business to improve energy storage technologies and to introduce a wider variety of vehicle models [2]. In parallel, the adoption of EVs has surged to a record: the aggregate sales have exceeded four Million since the first introduction of EVs in 2010 [3]. It is also projected that the EV stock will reach at least 40 Million by 2025 [4].

Such high EV penetration rates necessitate the widespread presence of public charging facilities, such as level 2 chargers (6–7 kW) at parking lots and fast DC chargers (40–50 kW) at places similar to gas stations. It has been well-documented that uncontrolled EV charging, especially during peak hours, threatens the stability of the power grids. Hence, the *economic* operation of power grids requires a careful coordination among EVs and charging stations [5]. This coordination requires massive data exchanges and processing [6]. Also, it is worth noting that the EVs typically do not have the computational power, memory, and storage capabilities, to solve the large-scale optimization required for the careful coordination of the charging.

Cloud computing has gained increasing attention for realizing cooperation in the smart grid [7], [8], [9]. The stability, flexibility, security, and on-demand performance of cloud computing provide an efficient platform for charging management [10]. The high computing and storage capacities in cloud infrastructures can support large-scale EV deployments with low latency [11]. Cloud computing can readily handle the computational and communication complexities for large-scale EV deployments of EVs, while offering flexible shared networked computing services that EVs can access from anywhere at anytime [11]. On the other hand, from a business point of view, instead of investing and maintaining data centers and additional reliable and efficient communication infrastructures, the system operator (SO) can make use of existing cloud computing services and communication infrastructures

Cuiyu Kong was with the Dept. of Electr. and Computer Eng., North Carolina State Univ., NC 27695 USA, and is currently with the Dept. of Computer Science and Computer Information Science, Highline College, Des Moines, WA 98198 USA (e-mail: cuiyu.kong@gmail.com).

Bhaskar P. Rimal was with the Dept. of Electr. and Computer Eng., Univ. of New Mexico, Albuquerque, NM 87131 USA and is now with the Beacom College of Computer and Cyber Sciences, Dakota State University, Madison, SD 57042 USA (e-mail: bhaskar.rimal@ieee.org).

Martin Reisslein is with the School of Electr., Comp., and Energy Eng., Arizona State Univ. (ASU), Tempe, AZ 85287 USA (e-mail: reisslein@asu.edu).

Martin Maier is with the Institut National de la Recherche Scientifique, Montreal, QC H5A 1K6, Canada (e-mail: maier@emt.inrs.ca)

Islam Safak Bayram is with the Dept. of Electronic and Electrical Eng., Univ. of Strathclyde, Glasgow, UK (e-mail: safak.bayram@strath.ac.uk).

Michael Devetsikiotis is with the Dept. of Electrical and Computer Eng., Univ. of New Mexico, Albuquerque, NM 87131 USA (e-mail: mdevets@unm.edu).

A shorter version of this paper appeared in [1].

to lower both capital expenditures (CAPEX) and operating expenses (OPEX) [12]. Consequently, industries are striving to incorporate cloud computing services into smart charging system [11], [13].

B. Related Work

To manage the charging of a large-scale fleet of EVs, demand-side management (DSM) of EVs in the smart grid has been studied in the literature with queuing models, without cloud computing, and with cloud computing. We have grouped the prior studies into these three groups. The following three subsections briefly review the prior related work and explain how our present study fundamentally advances the DSM of EV charging.

1) DSM of EVs with Queuing Models: Queuing models have been widely applied to evaluate system performance and characterize customers' demands in the DSM of EVs. For instance, EV charging and discharging was characterized with $M/M/c$ models and priorities in [14]. In [15], a similar queuing model was applied in a spatial and temporal system to characterize the charging demands at highway exits. The charging model in [16] was an $M/M/c/c$ queue in fast-charging stations. In [17], a model was proposed using a queuing model followed by a neural network to represent the total charging load at an EV charging station in terms of the number of EVs being charged, the total charging current, arrival rate, and time. However, the queueing model studies did not consider the customers' willingness to delay charging, nor the concept of local cloud-based management.

2) DSM of EVs without Cloud-based Management: The studies [18], [19], [20], [21], [22], [23] focused on DSM without cloud computing-based management. In [18], a centralized recharge scheduling system used traffic data to maximize the total parking lot revenue and the number of charged EVs. A decentralized charging mechanism with EVs' charging status controlled via probability automation was proposed in [19]. A distributed EV charging strategy integrated with the grid was proposed in [24]. An EVs frequency regulation service and an optimal capacity scheduling algorithm were proposed in [20]. A network capacity optimization model considering the EV charging preferences was studied in [22]. However, these preceding studies have not considered load shifting, nor capacity expansion, nor profit maximization strategies.

A stochastic-based optimal charging strategy for an aggregator that incorporates incentive and regulatory policies in terms of voltage profile and power loss cost of the network has been proposed in [25]; however no load shifting nor capacity expansion were considered. On the other hand, the study [26] proposed smart charging strategies that aimed to minimize total daily cost and peak-to-average ratio (PAR) but does not maximize the profit. In [27], to minimize the EVs trip duration, a charging station-selection scheme was proposed. However, the scheme does not account for the power loss of the distribution system, capacity planning, peak-load shifting, and any communications model between the CS and EVs.

An EV coordinated discrete charging model with grid capacity constraints was studied in [28] that aims to optimize the total load variations and total number of on-off switching in the charging process. However, only a single charging rate was considered and other important factors, such as power loss, profit maximization, and capacity expansion were not considered. An EV route selection optimization and a charge navigation based on crowd sensing that aims to reduce travel costs and improve the load level of the distribution system was proposed in [29]. A distributed algorithm to jointly optimize the routing selection and the charging scheduling was proposed in [30]. An EV charging scheduling with the objective to minimize the total overhead of recharging, considering charging availability and electricity price fluctuation was developed in [31].

We note that reviewed studies on DSM of EVs without cloud computing are representative of the literature on DSM of EVs without cloud computing, which we cannot review comprehensively due to space constraints. However, we emphasize that—to the best of our knowledge—none of the existing DSM of EVs studies without cloud computing has considered the profit maximization of the charging of heterogeneous EVs (with heterogeneous charging levels and delay requirements) with capacity planning and capacity expansion. Generally, cloud-based charging management is widely viewed as highly promising to be commercialized in the future. Therefore, we focus on cloud-based management in this study.

3) DSM of EVs with Cloud-based Management: Cloud-based DSM models were presented in [32], [33], [14]. A demand response algorithm for the smart grid, considering a cloud architecture was proposed in [32]. The study [33] proposed a cloud-based energy forecasting model in micro-grids to reduce the message overhead and energy consumption. The study [14] modeled the cloud-based charging and discharging management in public charging stations for demand response. In [34], a stochastic convex optimization problem was formulated to minimize the average overall trip time for all customers relative to their actual trip time without in-route charging, whereby cloud technology was used to enable the real-time collection of EV state-of-charge information, departure rates from passenger stations, energy supplies at different stations, and trip times without charging. A mobile edge computing-based system enabled by big data analytics for the EV charging use case was explored in [35]. A four-layer framework was proposed in [36] to coordinate charging and discharging requests of EVs.

However, the existing studies are limited to homogeneous EVs. In contrast, we consider the practical scenario of heterogeneous EVs with different time sensitivities (highly time-sensitive EVs at highways and parked EVs with relaxed timing requirements) as well as heterogeneous charging levels (modeling for instance DCFast charging and level 2 three phase charging). Only multiple charging levels have been considered in some prior studies, e.g., [18]; however, without considering the communication and computation delays or cost for cloud management. We address these heterogeneous EV requirements through two levels of cloud computing, a local cloud (with relatively high rental fee, but short communication delay and fast computing service) to manage the time-sensitive highway EVs and a remote cloud (with relatively low rental but long communication delay and slow computing service) to manage the parked EVs. Moreover, the existing cloud-based studies have not jointly considered price incentives and capacity expansion. In contrast, we comprehensively model and compare price incentives and capacity expansion and introduce a hybrid strategy that combines price incentives and capacity expansion.

C. Contributions

Facing the issues of limited energy and the management complexity of very large scale EV deployments, we present a cloud-based charging management framework in a network of charging stations and propose a profit maximization model. We take into account price incentives and different quality of service (QoS) provided to customers, whereby two levels of cloud computing are considered. The contributions of this paper are summarized as follows:

- From the infrastructure's perspective, we present for the first time two levels of cloud computing infrastructures, i.e., remote cloud and local cloud, to satisfy EVs' different latency requirements, thereby providing high scalability and fast response times. We develop novel algorithms for the interplay among EVs, SO, charging stations, and both clouds.
- Second, different from prior studies, from the viewpoint of charging stations, we consider the charging characteristics and geography of charging stations, and therefore consider charging stations with multiple charging levels (e.g., level 2 three-phase charging and DC fast charging), different locations (i.e., charging stations at highway exits (CS_H) and in parking lots (CS_P), and multiple charging power loss rates relating to the multiple charging levels. The multiple charging settings, locations, and charging power loss rates influence the system performance and service requirements of the EVs.
- Third, from the modeling viewpoint, we present cloud-based charging models, which are composed of queuing-based cloud server planning, power capacity planning, delay analysis, and profit maximization. For the cloud server planning, we consider the QoS metric (expected waiting time) and strive to minimize the cloud server rental fee. For the power capacity planning of charging stations, we present a greedy algorithm to reduce the planning complexity. The multiple charging settings, EVs' charging preferences (time sensitivity based on charging stations locations), maximal power constraints, and QoS metric (the weighted blocking probability) are incorporated.
- Fourth, from the communication aspect, the contention-based random access in 4G Long-Term Evolution (LTE) networks is considered. The EV end-to-end delay (D_{e2e}), which includes the contention delay, is analyzed and considered in the profit model.
- Fifth, from the business aspect, the ultimate goal of this study is to maximize the system profit while providing high QoS to EVs, whereby three strategies are introduced, namely, the price-incentive method (PIM), the capacity expansion method (CEM), and a hybrid method of the PIM and CEM (HPC). In the PIM, the optimal discount is offered to encourage customers to delay charging. In the CEM, the optimal extra power that needs to be purchased, and a penalty factor are included in the model. The HPC balances the immediate charging needs of customers (through the CEM) and the alleviation of heavy power grid loads (through the PIM). The goal of these strategies is to control and coordinate the EVs, thereby ensuring that the EV charging needs are satisfied while meeting different communication latency and maximizing the SO profit. requirements.

The remainder of this paper is organized as follows. Section II describes the system model, including cloud server planning, power capacity planning, delay analysis, and formulation of the profit maximization. Section III discusses the numerical results. Finally, conclusions are drawn in Section IV.

II. System Model

In Section II-A, we present an overview of our system model. Then considering the system QoS, the optimal server planning in the cloud is introduced in Section II-B, the capacity planning in a network of charging stations is conducted in Section II-C, and the delay is analyzed in Section II-D. Section II-E formulates the profit optimization and introduces the PIM, CEM, and HPC.

A. System Overview

We consider two classes of EVs: EVs at highway exit charging stations CS_H and EVs at parking lot charging stations CS_P. EVs at a highway charging stations CS_H are offered high-priority service as they require real-time service, while EVs in parking lot charging stations CS_P are offered low-priority service as they typically have comparably relaxed latency requirements since the EVs usually park for some periods of time. The SO purchases energy from the grid to provide charging service to EVs. Also, the SO distributes the energy to its own sub-networks.

We assume that time is divided into slots $t \in \{1, 2, \ldots, T\}$, and there are f charging stations near highway exits (CS_H) and $(K - f)$ charging stations at parking lots (CS_P), where $k \in \{CS_1, CS_2, \ldots, CS_f, CS_{f+1}, \ldots, CS_K\}$. We assume that the charging stations comply with the US Society of Automotive Engineers (SAE) standard. The arrival of EVs in CS_k in time slot t is commonly modeled as an exponential distribution with rate $\lambda_{k,t}$ [16]. We consider J charging levels in these public charging stations due to the different charging requirements of the EVs. (The prior model in [1] considered only a single charging level.) Each charging station k is permanently mapped to one charging level j.

Recently, hierarchical cloud computing and network architectures with multiple levels have been developed, see e.g., [37], [38], [39], [40], [41], [42]. In accordance with the needs of the two considered heterogenous types of EVs (highway EVs and parking lot EVs), we employ a two-level cloud computing architecture. The nearby local cloud with a relatively short end-to-end delay D^L_{e2e} [43], [44] can support the low latency requirements of the EVs at highway charging stations, see Fig. 1. The local cloud (edge cloud) brings the cloud computing capabilities in close proximity of the EVs, charging stations, and BSs. The local cloud can be distributed, i.e., there could be several local cloud locations that are each physically close to their corresponding BSs. There are multiple ways to deploy the local clouds. For example, a local cloud can be deployed at an aggregation point, where multiple BSs are located close together sharing a single local cloud platform [41]. This aggregation point approach centralizes resources and thus helps reduce both CAPEX and OPEX without experiencing significant latency. Alternatively, local clouds can be deployed at the BSs, i.e., integrated with cloudlets in the form of a so-called Access Edge Cloud, or deployed one level higher in the network hierarchy, i.e., at the aggregation point of the metro/core network as a so-called Metro Edge Cloud [43]. The local cloud locations can synchronize among each other using some cloud synchronization technique, such as Openstack Swift.

The remote cloud (e.g., Amazon EC2 or Microsoft Azure) with relatively long D^R_{e2e} can serve the EVs in parking lots, which have relatively relaxed latency requirements. For a given computing service rate, the rental fee of the remote cloud is typically lower than the corresponding fee for a local cloud due to the better economies of scale in a remote centrally located cloud installation. Thus, with a two-level cloud computing structure, our approach can exploit the better economies of scale of the remote cloud computing. The local and remote clouds can synchronize data by connecting their object storage APIs, e.g., Amazon S3 or Openstack Swift. The EVs, SO, and the clouds communicate through 4G LTE base stations. The EVs and SO connect to the base stations with wireless communication, while the base stations and the clouds are connected via optical fiber links. Power flows exist between EVs and charging stations for the charging.

The EVs in CS_H and CS_P publish and subscribe to a local cloud and a remote cloud, respectively. The remote cloud obtains the information of EVs from parking lots, synchronizes the data with the local cloud, and publishes the response messages to EVs in parking lots, so as to alleviate the load on the local cloud. The local cloud deployment details are out of scope of this paper. The base stations forward the EV information to the clouds. The cloud control center in the local cloud performs processing and controlling services, specifically, the server planning, capacity planning, and profit optimization. The SO sends real-time information, e.g., real-time electricity price U_t and real-time supply S_t, to the base stations. Then, the base stations forward the information to the local cloud. The deployment of charging stations may depend on several parameters, including the traffic density and the load on the power grids in a particular area. The locations of the charging stations can, for instance, be determined based on the flow-capturing location-allocation model [45].

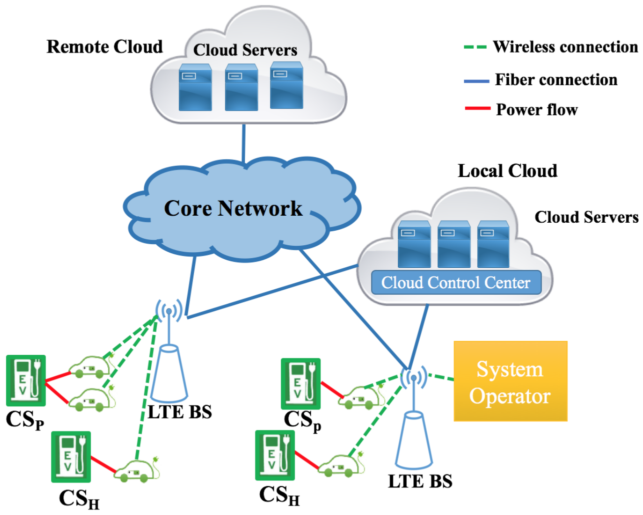


Fig. 1: Two-layered cloud computing-based EV charging system: A local cloud manages the high-priority charging of EVs at charging stations near highway exits (CS_H), while a remote cloud manages the charging stations in parking lots (CS_P) where the EVs have relaxed timing constraints.

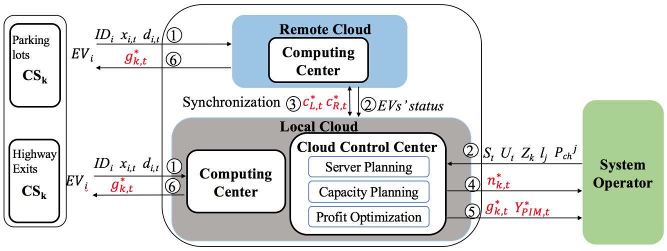


Fig. 2: System model with input and output parameters.

The detailed model framework is depicted in Fig. 2. EVs and SO communicate with both clouds via base stations. The cloud system module collects each customer's information (i.e., an EV i's ID, spatial location $x_{i,t}$ and charging demand $d_{i,t}$) on the left side of Fig. 2. The cloud system module collects the electricity supply S_t, the power capacity limits of each charging station Z_k (whereby k is the index for the charging station), the charging station power loss rate l_j (whereby the charging level j ranges from $j = 1, 2, \ldots, J_k$), the power of charging levels P^j_{ch} and the electricity price U_t from the SO, see right side of Fig. 2. After the cloud control center has completed the cloud server planning, the cloud control center sends the optimal number of local and remote cloud servers $c^*_{L,t}$ and $c^*_{R,t}$ to the local cloud and the remote cloud, respectively. The optimal number of charging outlets $n^*_{k,t}$ is published to the SO. Next, the cloud control center responds to the SO with discount $g^*_{k,t}$ and profit $Y^*_{PIM,t}$, and to customers with discount $g^*_{k,t}$. The parameters and their description are summarized in Table I.

Algorithms 1, 2, 3, and 4 summarize the model operations. The algorithm executed by the computing center in the local cloud is similar to Algorithm 2 executed in the remote cloud and is therefore omitted. The cloud control center executes the algorithms in real time, as triggered by new EV arrivals. The algorithms use the current data and estimated hourly arrival rates. The EV charging schedule is planned on a time slot (hourly) basis. Note that the remote cloud computing center receives the discounts from the cloud control center and publishes the discounts to the parking lot EVs since only the

TABLE I: Variables and their description.

Variables	Description
U_t	Electricity price at time t
$x_{i,t}$	Spatial location of EV i at time t
$d_{i,t}$	Charging demand of EV i at time t
S_t	Electricity supply of system operator at time t
Z_k	Power capacity of charging station k
P_{ch}^j	Power of charging level j
l_j	Power loss rate at charging level j
$c_{L,t}$	Number of local-cloud servers at time t
$c_{R,t}$	Number of remote-cloud servers at time t
E_r	Energy requested by an EV
$n_{k,t}$	Number of charg. outlets at charging station k at time t
$g_{k,t}$	Discount for delaying the charging at station k at time t
$Y_{PIM,t}$	Profit from price-incentive method (PIM) at time t
$\lambda_{k,t}$	EV arrival rate at charging station k at time t
$\lambda_{k,t}^d$	Updated arrival rate at charging station k at time t after offering discount
$\mu_{k,t}$	Service rate at charging station k at time t
$t_w^{L,t}$	Avg. waiting time of an EV in local cloud at time t
$t_w^{R,t}$	Avg. waiting time of an EV in remote cloud at time t
α_k	Time sensitivity at charging station k
$\rho_{k,t}$	Charging outlet utilization at charging station k at time t

Algorithm 1: Executed by an EV.

```
Input  : An EV $i$'s status: ID, location $x_{i,t}$, charging
         demand $d_{i,t}$, network status $NS$ =Not
         Connected, transm. attempt $w = 0$,
         max. transm. attempt $W$
Output: Discount $g_{k,t}^*$ , Contention delay $D_{con,t}^i$
while $w \leq W$ and $NS$==Not Connected do
    Send a preamble for connection request;
    $w \leftarrow w$+1;
    if No collision then
        $NS$=Connected;
        Calculate $D_{con,t}^i$ based on the time it takes
          EV $i$ to connect;
        Send charging request;
        Receive discount $g_{k,t}^*$;
        if $g_{k,t}^*$ is acceptable then
            Send ACK to accept the delay;
        else
            Send ACK to reject the delay;
end
if $NS$==Not Connected then
    Communication connection failed;
```

remote cloud computing center stores the status information of the parked EVs. The details of the hierarchical model are described in the following subsections.

B. Optimal Server Planning in the Cloud

To minimize the server rental cost in the cloud, we analyze the numbers $c_{L,t}^*$ and $c_{R,t}^*$ of local and remote cloud servers for given QoS thresholds. Specifically, the average waiting times of EVs in the queues of the local cloud $t_w^{L,t}$ and the remote cloud $t_w^{R,t}$ need to be below the thresholds ϵ_L and ϵ_R.

Algorithm 2: Executed by remote cloud.

```
Input  : Status (ID, location $x_{i,t}$, charging demand
         $d_{i,t}$) of each parking lot EV
Output: Discount $g_{k,t}^*$ to EVs
for $t$ = 1 to $T$ do
    if A remote cloud server is available then
        Receive and store EVs' information and
          charging request;
        Analyze each EV's charging request based on
          First-In-First-Out (FIFO) order;
        Synchronize EVs' information with local cloud;
    else
        Store charging request in buffer;
    Obtain $c_{R,t}^*$ and $g_{k,t}^*$ from cloud control center;
    Publish $g_{k,t}^*$ to EVs in parking lots;
end
```

Algorithm 3: Executed by cloud control center in local cloud.

```
Input  : Each EV's status (ID, $x_{i,t}$, $d_{i,t}$), $S_t$, $U_t$, $Z_k$,
         $l_j$, $P_{ch}^j$
Output: $c_{L,t}^*$, $c_{R,t}^*$, $Y_{PIM,t}^*$, $n_{k,t}^*$, $g_{k,t}^*$ $\forall k \in K$, $\forall t \in T$
for $t$ = 1 to $T$ do
    Collect status of EVs, $S_t$, $U_t$, and $Z_k$;
    Analyze $D_{e2e}^{L,t}$ and $D_{e2e}^{R,t}$ using (16);
    Execute server planning:
     The target is to minimize the server rental fee (3)
     with constraints $\epsilon_L$ (4) and $\epsilon_R$ (5), the cloud
     control center returns server numbers $c_{L,t}^*$ and
     $c_{R,t}^*$ to the local cloud and remote cloud;
    Execute capacity planning (10) with constraints
     (11)–(15). Return $n_{k,t}^*$ $\forall k \in K$, see Alg. 4;
    if $d < S_t$ and $b_{k,t} < \epsilon$ then
        Return $S1$ to SO and $g_{k,t}^* = 0$;
    else
        Execute profit optimization module: First,
     calculate $\lambda_{k,t}^*$, $\forall k \in K$ in (19). Second, execute
     PIM (23) with constraints (24)–(26) for load
     shifting. Third, calculate $C_{loss}^{k,t}$ in (7) and $C_{delay}^{k,t}$
     in (21). Get $g_{k,t}^*$, and $Y_{PIM,t}^*$;
    Publish $g_{k,t}^*$ and $Y_{PIM,t}^*$ to SO;
    Publish $g_{k,t}^*$ to EVs;
end
```

Following [41], [46], [47], we model both the local and remote cloud computing centers as $M/M/c$ queuing systems with arrival rates $\lambda_{L,t}$, $\lambda_{R,t}$ and service rates μ_L, μ_R that are independent and exponentially distributed. By the Poisson merging property, the aggregate arrival processes with rates $\lambda_{L,t} = \sum_{k=1}^{f} \lambda_{k,t}$ and $\lambda_{R,t} = \sum_{k=f+1}^{K} \lambda_{k,t}$ are Poisson processes, consistent with the $M/M/c$ model.

The cloud queueing delay t_w^t has an important influence on the end-to-end delay. Therefore, the target of our server planning is to optimize the cloud server numbers $c^*_{L,t}$ and $c^*_{R,t}$, given the thresholds ϵ_L and ϵ_R. Generally, for arrival rate λ and service rate μ, and thus traffic intensity $\rho = \lambda/(c\mu)$, the probability C that customers have to wait in an $M/M/c$ queue before getting served and the resulting mean waiting time t_w follow from queueing theory [48] as

$$C\left(c, \frac{\lambda}{\mu}\right) = \frac{1}{1 + (1-\rho) \cdot \frac{c!}{(c\rho)^c} \cdot \sum_{r=0}^{c-1} \frac{(c\rho)^r}{r!}} \tag{1}$$

$$t_w(c) = \frac{C(c, \frac{\lambda}{\mu})}{c \cdot \mu - \lambda}. \tag{2}$$

The total server rental cost is composed of the cost of the local cloud and the remote cloud servers. Let p_L and p_R denote the rental fee of a server in the local cloud and the remote cloud in one time slot, respectively. Due to the real-time function of the local cloud, the local cloud has higher rental fee $p_L > p_R$ and service rate $\mu_L > \mu_R$ than the remote cloud. The optimization problem at time t is formulated as:

$$\begin{aligned} \min \quad & p_L \cdot c_{L,t} + p_R \cdot c_{R,t} && (3)\\ \text{s.t.} \quad & t_w^{L,t}(c_{L,t}) \le \epsilon_L && (4)\\ & t_w^{R,t}(c_{R,t}) \le \epsilon_R && (5)\\ & t_w^{L,t}, t_w^{R,t}, c_{L,t}, c_{R,t} \ge 0. && (6) \end{aligned}$$

The objective function (3) minimizes the total cost of cloud services from both the local cloud and remote cloud. The constraints (4) and (5) guarantee the QoS to customers. Constraint (6) enforces that the server numbers $c_{L,t}$ and $c_{R,t}$ are non-negative, whereby we assume that $c_{L,t}$ and $c_{R,t}$ are integers. Eqs. (1) and (2) indicate that the waiting time t_w decreases with the number of servers c. Thus, the optimal cost is attained for $c^*_{L,t} = \lceil \arg_{c_{L,t}}(t_w^{L,t}(c_{L,t}) = \epsilon_L) \rceil$ and $c^*_{R,t} = \lceil \arg_{c_{R,t}}(t_w^{R,t}(c_{R,t}) = \epsilon_R) \rceil$, whereby the ceiling functions set the server numbers to the smallest integers that ensure that the waiting times are less than or equal to ϵ_L and ϵ_R, respectively.

C. Capacity Planning in the Network of Charging Stations

Due to the different geographical locations of the charging stations, the charging stations may encounter uneven charging demands. Therefore, it is necessary to provide more charging outlets to stations with heavy loads. On the other hand, a limited power supply may not satisfy all customers' charging demands in a timely manner during peak hours. Customers who are not served promptly are blocked. For gaining high reputation and providing a high QoS, it is critical to guarantee a low blocking probability. Accordingly, we consider the total weighted blocking probability as a performance metric.

Every charging station is modeled as an $M/M/c/c$ queue. The service rate μ_k is a function of the charging level j and the charging efficiency in charging station k. Power losses occur in a charging station because of the characteristics of the power electronics. Specifically, with P^j_{ch} and P^j_{out} denoting the power going in and out of a charging outlet at charging level j, the power losses rate is [49], [50]:

$$l^j = 1 - \frac{P^j_{out}}{P^j_{ch}}. \tag{7}$$

The charging efficiency of a charging station with charging level j is $\eta_j = 1 - l^j$. For a given EV battery capacity E, expected value of the requested state of charge (SoC) SoC_r, and expected value of the initial SoC SoC_i, the service rate in CS_k is

$$\mu_k = \frac{1}{\frac{d}{P^j_{ch}\eta^j}}, \tag{8}$$

where the average charging demand of an EV is denoted as $d = (SoC_r - SoC_i)E$. Then, with $n_{k,t}$ charging outlets in CS_k, the blocking probability of CS_k follows from loss system theory [51] as

$$b_{k,t} = \frac{\left(\frac{\lambda_{k,t}}{\mu_k}\right)^{n_{k,t}} / n_{k,t}!}{\sum\limits_{i=0}^{n_{k,t}} \left(\frac{\lambda_{k,t}}{\mu_k}\right)^i / i!}. \tag{9}$$

Due to the internal resistance of an EV battery, the charging time increases exponentially when the SoC is close to full, resulting in a general distribution of the service rate μ_k, see detailed battery characteristics in [52], [53]. For simplicity, we do not consider these detailed battery characteristics. Instead, we assume that μ_k follows an exponential distribution and therefore apply Eq. (8).

Customers in CS_P are willing to delay the charging if a discount is offered, while customers in CS_H would delay charging for a larger discount, since they are more time sensitive. Thus, the time sensitivity parameters α_H and α_P are considered in the capacity planning, whereby $\alpha_H > \alpha_P$. Hence, in CS_k, when $k \le f$, $\alpha_k = \alpha_H$; else, $\alpha_k = \alpha_P$.

The maximal power capacity Z_k of charging station k, i.e., of CS_k, with assigned charging power level P^k_{ch}, is considered as the power constraint. The variables in the optimization problem are the numbers of charging outlets $n_k, \forall k \in K$. Considering the performance of a network of charging stations, we introduce the weighted value $w_{k,t} = \lambda^d_{k,t} \Big/ \sum\limits_{k=1}^{K} \lambda^d_{k,t}$. and formulate the problem at time t as:

$$\begin{aligned} \min \quad & \sum_{k=1}^{K} \alpha_k \cdot (w_{k,t} \cdot b_{k,t}) && && (10)\\ \text{s.t.} \quad & \lambda^d_{k,t} = \lambda_{k,t} + \lambda_{k,t-1} \cdot \theta(g_{k,t-1}) && \forall k \in K && (11)\\ & b_{k,t} = \frac{\left(\frac{\lambda^d_{k,t}}{\mu_{k,t}}\right)^{n_{k,t}} / n_{k,t}!}{\sum\limits_{i=0}^{n_{k,t}} \left(\frac{\lambda^d_{k,t}}{\mu_{k,t}}\right)^i / i!} && \forall k \in K && (12)\\ & P^k_{ch} \cdot n_{k,t} \le Z_k && \forall k \in K && (13)\\ & \sum_{k=1}^{K} n_{k,t} P^k_{ch} \le S_t && && (14)\\ & n_{k,t} > 0 && \forall k \in K. && (15) \end{aligned}$$

Constraint (11) describes the updated arrival rate $\lambda^d_{k,t}$, after

Algorithm 4: Charging outlet allocation algorithm executed by the cloud control center in local cloud.

```
   Input : $\left[\frac{\lambda^d_{1,t}}{\mu_{1,t}}, \frac{\lambda^d_{2,t}}{\mu_{2,t}}, \ldots, \frac{\lambda^d_{K,t}}{\mu_{K,t}}\right]$, $\alpha_k$, $S_t$
           $P^k_{ch}$, $Z_k$, $n_{k,t}$ $\forall k \in K$, $\forall j \in J$
   Output: $n = [n_{1,t}, n_{2,t}, \ldots, n_{K,t}]$,
           $Result_t = \sum_{k=1}^{K} w_{k,t} b_{k,t}$
   Initialization: $n_{1,t} = \cdots = n_{K,t} = 1$,
           $temp = S - \sum_{k=1}^{K} n_k P^k_{ch}$,
           $w_{k,t} = \frac{\lambda^d_{k,t}}{\sum_{k=1}^{K} \lambda^d_{k,t}}$, $\rho_{k,t} = \frac{\lambda^d_{k,t}}{n_{k,t} \cdot \mu_{k,t}}$
1  $V_t = [\rho_{1,t} \cdot \alpha_1, \ldots, \rho_{K,t} \cdot \alpha_K]$;
2  while $temp \geq 0$ and $V_t \neq \{0\}$ do
3      $L = \arg_k \max(V_t)$;
4      if $(n_{L,t} + 1) P^j_{ch} \leq Z_L$ and $temp \geq P^k_{ch}$ then
5          $n_{L,t} \leftarrow n_{L,t} + 1$;
6          $temp \leftarrow temp - P^j_{ch}$;
7      else
8          $\lambda_{k,t} = 0$;
9      Update $V_t$ and $\rho_{k,t}$;
10 end
11 for $k$ = 1 to $K$ do
12     $b_{k,t} = \frac{(\rho_{k,t} \cdot n_{k,t})^{n_{k,t}}}{n_{k,t}!} \Big/ \sum_{i=0}^{n_{k,t}} \frac{(\rho_{k,t} \cdot n_{k,t})^i}{i!}$;
13 end
14 $Result_t = \sum_{k=1}^{k} w_{k,t} b_{k,t}$;
```

providing discount $g_{k,t}$. The updated arrival rate $\lambda^d_{k,t}$ is composed of both new arrivals (with rate $\lambda_{k,t}$) and delayed EVs from the preceding time slot, whereby $\theta(g_{k,t})$ is an EV's probability of delayed charging, which is described in Section II-E. Constraint (13) guarantees the amount of electrical power (kW) distributed to a charging station is no more than the maximal charging power rate of a charging station.

Based on our previous work [54], we proceed to develop a greedy algorithm to find near-optimal solutions for the charging outlets allocation. However, note that different from [54], we incorporate the maximal power capacity Z_k, multiple (J) charging levels, and time sensitivity parameter α_k in the charging station selection algorithm. With the help of these three parameters, our novel model flexibly offers different service level agreements (SLAs) based on the customer needs. The details are presented in Algorithm 4.

Algorithm 4 presents the procedure of allocating charging outlets. First, a charging station with a heavy load is selected in the further step of allocating one charging outlet. The time sensitivity α_k is incorporated with the server intensity $\frac{\lambda^u_{1,t}}{n_{1,t} \cdot \mu_{1,t}}$ in this operation, see Line 1. Second, after selecting a charging station, its charging level j and maximum power in this charging station Z_k are considered. At each iteration, if the remaining power supply can support an EV with its charging level and the power does not exceed Z_k, then one charging outlet is allocated to this charging station, see Line 4. Towards this end, we calculate the weighted blocking probabilities based on the charging outlet allocation.

D. Delay Analysis

1) End-to-end delay: In time slot t, the expected end-to-end delay D^t_{e2e} is the sum of the EVs' expected contention delay $D_{con,t}$ to connect to the 4G LTE base stations, the various expected packet transmission delays D_{trans}, the various expected propagation delays D_{prop}, and the expected queuing time in the cloud, and expected cloud servers' processing time. Specifically, D^t_{e2e} is computed as:

$$\begin{aligned} D^t_{e2e} = D_{con,t} + D^e_{trans,req} + 2D^e_{prop} + D^c_{trans,req} \\ +2D^c_{prop} + \frac{1}{\mu} + t^t_w + D^e_{trans,res} + D^c_{trans,res}, \end{aligned} \tag{16}$$

whereby the individual delay components are defined as follows. The contention delay $D_{con,t}$ is a function of the arrival rate of EVs and the number of available 4G LTE preambles and is analyzed in Section II-D2. With packet size P_{req} and 4G LTE network transmission rate R_{BS}, the transmission delay from EVs to BSs is $D^e_{trans,req} = P_{req}/R_{BS}$. With BS-to-cloud server link transmission rate R_c, the transmission delay from BSs to the cloud servers is $D^c_{trans,req} = P_{req}/R_c$. D^e_{prop} and D^c_{prop} are the expected propagation delays between EVs and BSs, and between BSs and the cloud servers, respectively. In particular, $D^c_{prop} = D^L_{prop}$ between BSs and the local cloud, while $D^c_{prop} = D^R_{prop}$ between BSs and the remote cloud.

Similarly, with the response packet size P_{res}, $D^e_{trans,res} = P_{res}/R_{BS}$ and $D^c_{trans,res} = P_{res}/R_c$. The average waiting time t^t_w before getting served in the $M/M/c$ system is from Eq. (2). Note that the cloud server waiting times t^t_w and processing times $1/\mu$ for the local and remote cloud are different, depending on their service rates μ and server numbers $c_{L,t}$ and $c_{R,t}$. Note that the processing time of a cloud server is the time to process the data and return the result, while the waiting time is the time that a customer has to wait before getting served. The difference between the local and remote clouds is due to different service rates and different numbers of servers.

2) Contention delay: When EVs do not have a grant for uplink data transmission and no Physical Uplink Control Channel (PUCCH) has been configured for scheduling requests, a random access procedure is initiated [55]. The 4G LTE random access is based on preamble transmissions [56]. Hence, we consider only the contention delay for the uplink network.

We denote ΔT_s for the time slot duration of the contention process (and note that this contention time slot is different from the time slot denoted by t for the charging system. The charging slot time (minutes to hours) is much longer compared to the contention slot time (ms)). Furthermore, we denote W for the maximum number of transmission attempts, and O for the number of preambles. The total arrival rate of new requests is $\lambda_{s,t} = \sum_{k=1}^{K} \lambda_{k,t}$. We denote the expected number of EVs, including new and previously collided requests that transmit a preamble at time t by $N_{s,t}$. Then, the expected number of EVs per preamble is $x_{s,t} = N_{s,t}/O$ and the probability of successfully sending a preamble without collision in time slot t is $p_{s,t} = e^{-x_{s,t}}$ [56]. For $W \leq 8$, the balance equation

$$\frac{N_{s,t}}{\lambda_{s,t}} = \frac{1-(1-p_{s,t})^W}{p_{s,t}} \tag{17}$$

has a single unique solution, i.e., a unique value of $p_{s,t}$, for given $\lambda_{s,t}/O$ and W [56]. For the simplifying assumption that the backoff time is zero when an EV encounters a collision, the average delay from request generation to successful preamble transmission is [56]:

$$\begin{aligned} D_{con,t}(\lambda_{s,t}, O) = B \cdot \big(1 + (W-1)(1-p_{s,t})^W\big) \\ -B \cdot \big(W(1-p_{s,t})^{W-1}\big) + \frac{\Delta T_s}{2}, \end{aligned} \tag{18}$$

where $B = \Delta T_s \cdot \big(\frac{1}{p_{s,t}} - 1\big) \cdot \frac{1}{1-(1-p_{s,t})^W}$.

3) Cloud Server Connection Alternatives: We briefly note that our system model allows for flexible cloud server options. So far, we have considered a local cloud with short propagation delay D_{prop}^{L} (compared to the longer propagation delay D_{prop}^{R} of the remote cloud) and high compute service rate μ_L (compared to the lower compute service rate μ_R of the remote cloud). Depending on the available cloud computing infrastructures and services, our model accommodates flexible cloud computing arrangements. For instance, our system can operate with only a local cloud (to ensure low delay service), which would be modelled by setting $D_{prop}^{R} = D_{prop}^{L}$ and $\mu_R = \mu_L$, at the expense of not exploiting the cheaper remote cloud computing. Alternatively, our system can operate by renting c_L fast servers (providing a high compute service rate μ_L) and c_R slow servers (providing a low compute service rate μ_R) in a local cloud with short propagation delay D_{prop}^{L}, to exploit the cost savings of cheaper low service rate cloud computing.

E. Profit Optimization

The goal of the SO is to maximize the system profit and to guarantee service to a large percentage of the EVs. Capacity planning minimized the total weighted blocking probability. However, capacity planning cannot guarantee fairness among charging stations because the stations with high traffic intensities have priority to get the majority of the charging outlets to alleviate the high traffic intensities. However, this allocation may leave insufficient numbers of outlets for other stations. As a result, other stations may also experience the problem of high blocking rates. For better QoS and to be fair, our aim is to ensure that the blocking probability $b_{k,t}$ at every charging station k in (9) is at most ϵ. The corresponding optimal arrival rate is given by

$$\lambda_{k,t}^{*} \leq \underset{\lambda}{\operatorname{argmax}}(b_{k,t} \leq \epsilon). \tag{19}$$

The SO strives to serve the largest number of EVs given a prescribed ϵ threshold; thus, $\lambda_{k,t}^{*}$ is the maximal value of the arrival rate at charging station k and can be achieved when $b_{k,t} = \epsilon$.

When the supply satisfies the EVs' demands within the blocking rate threshold, i.e., when $b_{k,t} \leq \epsilon$ at all charging stations, then the cloud control center notifies the SO that there is sufficient supply. On the other hand, in peak hours, $\lambda_{k,t}$ may be higher than $\lambda_{k,t}^{*}$. Thus, we adapt the general strategy of offering discounts to users [16] to our system and propose the PIM to shift the load from peak hours to off-peak hours in Section II-E1. The CEM which purchases a sufficient amount of energy in introduced in Section II-E2, and the HPC, a hybrid method of the PIM and CEM, is presented in Section II-E3. All three methods are compared in Section III.

1) Price-Incentive Method (PIM): The PIM incentivizes EVs to shift their charging demands from peak hours to off-peak hours by offering a discount. When the supply is insufficient to guarantee the QoS level of blocking probability less than ϵ, the system provides a discount to encourage customers to shift their demands by one time slot, i.e., we apply load shifting concepts [16], [57]. However, we modify the existing load shifting models from one charging station to a network of charging stations, and include the penalty costs for not serving customers, the rental fee from cloud services, and our price-sensitivity function for delayed charging in our profit model.

Assume the price sensitivity function is known to the system. Applying the concept of a linear price sensitivity function [58], suppose the admission (charging) fee of an EV at charging station k of charging level j for a one-time charge is A_j and the offered discount is $g_{k,t}$. The discount ranges of EVs in CS_H and CS_P of charging level j are $[0,\ g_{\max}^{H,j}]$ and $[0,\ g_{\max}^{P,j}]$, respectively, mapping to the probability $\theta(g_{k,t}) \in [0,1]$ of a customer delaying the charging. The higher the offered discount $g_{k,t}$, the higher the probability $\theta(g_{k,t})$ that a customer delays charging to the next time slot. For a given charging level j of charging stations, $g_{\max}^{P,j} < g_{\max}^{H,j} \leq A_j$ because customers in CS_H are more time sensitive.

From the perspective of a charging station CS_k at time t, we consider the admission fee A_j of charging level j of newly arrived EVs without delaying charging and delayed EVs from the preceding time slot $t-1$ to the current time slot t, the penalty for not serving an EV p_{pen}, the cost of power losses, the penalty of an EV's $D_{e2e}^{k,t}$, and the maintenance cost of a charging station p_m. The number of EVs delayed from $t-1$ to t is $\acute{\lambda}_{k,t-1} = \lambda_{k,t-1} \cdot \theta(g_{k,t-1})$. The average EV charging demand is denoted by $d = (SoC_r - SoC_i)E$.

In time slot t, the cost of the energy loss in a charging station k with charging level j is computed as

$$C_{loss}^{k,t}(\lambda_{k,t}^{d}) = U_t \cdot l^j \cdot P_{ch}^{j} \cdot \frac{1}{\mu_{k,t}} \cdot \lambda_{k,t}^{d} \cdot (1 - b_{k,t}(\lambda_{k,t}^{d})), \tag{20}$$

where $\lambda_{k,t}^{d}$ is the number of EVs in the system at time t, including the newly arrivals rejecting to delay and the arrivals accepting to delay from time $t-1$. Hence, $\lambda_{k,t}^{d} = \lambda_{k,t} \cdot (1 - \theta(g_{k,t})) + \acute{\lambda}_{k,t-1}$. Eq. (20) describes the cost of the amount of energy loss in a unit of time in CS_k, which is related to the electricity price U_t, the power loss rate l^j, charging power level P_{ch}^{j}, the duration of the charging time $1/\mu_{k,t}$, and the number of served EVs in a time slot $\lambda_{k,t}^{d} \cdot (1 - b_{k,t}(\lambda_{k,t}^{d}))$.

$D_{e2e}^{k,t}$ is differentiated by the connection of either local cloud or remote cloud. Therefore, the cost of delay in a charging station k is given by

$$C_{delay}^{k,t}(\lambda_{k,t}) = \begin{cases} \zeta_H \cdot \lambda_{k,t} \cdot e^{(D_{e2e}^{L,t} - D_H)} & k \leq f \\ \zeta_P \cdot \lambda_{k,t} \cdot e^{(D_{e2e}^{R,t} - D_P)} & \text{otherwise,} \end{cases} \tag{21}$$

where ζ_H and ζ_P are the penalty factors, and D_H and D_P are the delay thresholds for CS_H and CS_P, respectively. Referring to [59], the exponential function is used to quantify the impact of an EV's D_{e2e}. Therefore, at time t, the profit $R_{k,j,t}^{PIM}$ after offering discount at CS_k with corresponding charging level j is given by

$$\begin{aligned} R_{k,j,t}^{PIM} &= (1 - b_{k,t}) \cdot \lambda_{k,t} \cdot (1 - \theta(g_{k,t})) \cdot (A_j - U_t \cdot d) \\ &\quad + (1 - b_{k,t}) \cdot \acute{\lambda}_{k,t-1} \cdot (A_j - g_{k,t-1} - U_t \cdot d) \\ &\quad - b_{k,t} \cdot \lambda_{k,t}^{d} \cdot p_{pen} - C_{loss}^{k,t}(\lambda_{k,t}^{d}) - C_{delay}^{k,t}(\lambda_{k,t}) - p_m. \end{aligned} \tag{22}$$

Considering the rental cost Eq. (3), the total profit in a network of charging stations with offered discounts is formulated as

$$Y_{PIM,t} = \max \sum_{k \in K} R_{k,j,t}^{PIM} - p_L \cdot c_{L,t} - p_R \cdot c_{R,t} \tag{23}$$

$$\text{s.t. } \lambda_{k,t}(1 - \theta(g_{k,t})) + \lambda_{k,t-1}\theta(g_{k,t-1}) \leq \lambda_{k,t}^{*} \quad \forall k \in K \tag{24}$$

$$0 \leq g_{k,t}, g_{k,t-1} \leq g_{\max} \qquad \forall k \in K \tag{25}$$

$$0 \leq \theta(g_{k,t}) \leq 1 \qquad \forall k \in K. \tag{26}$$

The objective function (23) characterizes the total profit in a network of charging stations. The constraint (24) guarantees that the sum of customers who reject delaying charging and customers who accept delaying charging from the preceding time slot is no larger than the optimal arrival rate. In constraint (25), $g_{\max}$ denotes the applicable maximum discount $g_{\max}^{H,j}$ or $g_{\max}^{P,j}$ for the considered location of CS_k (highway or parking lot) and charging level j. Our model in this article is limited to delaying the charging by one time slot. Modeling delays of multiple time slots, possibly in conjunction with different discounts for different delay times, is an interesting direction for future research.

2) Capacity Expansion Method (CEM): With the CEM, the SO purchases extra electricity to satisfy the peak demand, with the penalty of expanding the system capacity. The CEM algorithms executed by the cloud control center, an EV, and the remote cloud are similar to the PIM algorithms in Algorithm 3. Thus, the CEM algorithms are not shown. The CEM considers the QoS performance ($b_{k,t} \leq \epsilon$). Hence, in CS_k, the optimal number of charging outlets $n_{k,t}^{*}$ should satisfy $n_{k,t}^{*} \geq \arg_n \min(b_{k,t} \leq \epsilon)$. From Eq. (9), $b_{k,t}$ is decreasing with $n_{k,t}$, so $n_{k,t}^{*} = \lceil \arg_n(b_{k,t} = \epsilon) \rceil$. Therefore, the number of extra charging outlets that need to be deployed at CS_k is

$$\Delta n_{k,t} = \begin{cases} n_{k,t}^{*} - n_{k,t} & \text{if } n_{k,t} < n_{k,t}^{*} \\ 0 & \text{otherwise.} \end{cases} \tag{27}$$

Applying the general penalty concept [60], when the demand is larger than the pre-decided supply, the SO has to pay $\acute{U}_t = U_t \cdot (1 + \sigma)$ for extra power, whereby σ is the real-time price penalty factor.

The arrival rate at t is $\lambda_{k,t}$ since no discount is offered. Hence, the profit $Y_{CEM,t}$ when purchasing electricity is:

$$\begin{aligned} Y_{CEM,t} = \sum_{k=1}^{K} \big(&\lambda_{k,t} \cdot (1 - b_{k,t}(\lambda_{k,t}, n_{k,t}^{*})) \cdot (A_j - U_t \cdot d) \\ &- \lambda_{k,t} \cdot (1 - b_{k,t}(\lambda_{k,t}, n_{k,t}^{*})) \cdot \frac{\Delta n_{k,t}}{n_{k,t}^{*}} \cdot U_t \cdot \sigma \cdot d \\ &- b_{k,t}(\lambda_{k,t}, n_{k,t}^{*}) \cdot \lambda_{k,t} \cdot p_{pen} - C_{loss}^{k,t}(\lambda_{k,t}) \\ &- C_{delay}^{k,t}(\lambda_{k,t}) - p_m - p_L \cdot c_{L,t} - p_R \cdot c_{R,t}, \end{aligned} \tag{28}$$

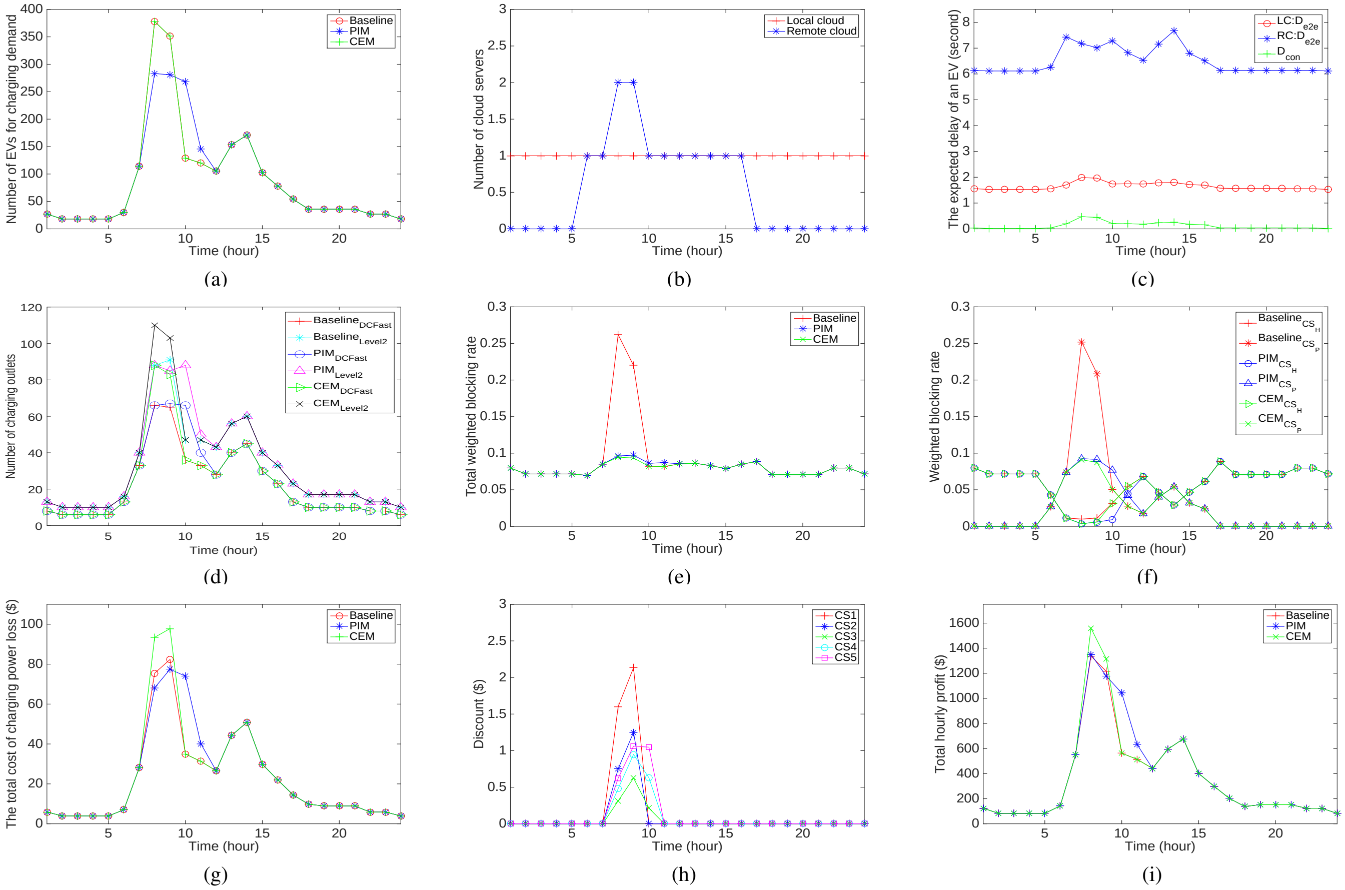
where the first term is the profit gained from the served EVs. The second term is the penalty for the extra energy; specifically, the formula $\lambda_{k,t} \cdot (1 - b_{k,t}(\lambda_{k,t}, n_{k,t}^{*})) \cdot \frac{\Delta n_{k,t}}{n_{k,t}^{*}}$, is the number of customers using the extra power to charge their EVs. The third term includes the penalty of blocking EVs and the cost of energy loss. The fourth term is composed of the penalty of communication delay, the maintenance cost, and the rental fees for the cloud servers. The CEM ensures high QoS for customers. However, by purchasing more electricity, the CEM burdens the grid during peak hours.

3) Hybrid PIM and CEM (HPC): The HPC is a hybrid method of the PIM and CEM. In the PIM, the discount $g_{k,t}$ determines the percentage $\theta(g_{k,t})$ of customers that are willing to delay charging; whereby, a higher percentage $\theta(g_{k,t})$ of customers will delay their charging if a higher discount $g_{k,t}$ is offered. However, customers who are under time pressure (i.e., need to arrive at their destination on time), will reject to delay charging even if a higher discount is offered. For these customers who cannot delay their charging, the SO may adopt the CEM by purchasing more power to satisfy these customers while performing the PIM for the remaining customers. Specifically, the HPC mainly differs from the PIM through the percentage β, $0 < \beta < 1$, of customers, who reject to delay their charging due to time pressures to arrive at their destinations; the CEM will be performed for this percentage β of customers. Correspondingly, the percentage of customers who will be involved in the PIM decreases to $(1 - \beta)$.

The number of EVs that delay charging from $t - 1$ to t is $\acute{\lambda}_{k,t-1}^{HPC} = \lambda_{k,t-1} \cdot (1 - \beta) \cdot \theta(g_{k,t})$. We define $\lambda_{k,t}^{HPC}$ as the number of EVs arriving at time t, including the new arrivals that reject to delay charging and the arrivals that have accepted to delay from time $t - 1$ to time t, i.e., $\lambda_{k,t}^{HPC} = \lambda_{k,t} \cdot (1 - (1 - \beta) \cdot \theta(g_{k,t})) + \acute{\lambda}_{k,t-1}^{HPC}$.

Since only a proportion $1 - \beta$ of customers participate in the PIM, charging stations need to serve more customers compared to the case of the pure PIM, which corresponds to $\beta = 0$. To maintain the same QoS (e.g., same blocking probability), more charging outlets will need to be deployed. The optimal number $n_{k,t}^{(HPC)^*}$ of charging outlets still needs to satisfy $n_{k,t}^{(HPC)^*} \geq \arg_n \min(b_{k,t}^{HPC} \leq \epsilon)$. The blocking probability $b_{k,t}^{HPC}$ with the HPC is a function of the arrival rate $\lambda_{k,t}^{HPC}$ and number $n_{k,t}^{(HPC)*}$ of outlets. From Eq. (9), $b_{k,t}^{HPC}$, decreases with $n_{k,t}^{HPC}$, thus $n_{k,t}^{(HPC)^*} = \lceil \arg_n(b_{k,t}^{HPC} = \epsilon) \rceil$.

Considering the penalty of purchasing extra power, the



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 3: Performance comparison: (**a**) Charging demand comparison; (**b**) Cloud servers in local and remote cloud; (**c**) End-to-end delay and contention delay; (**d**) charging outlet allocation; (**e**) Weighted blocking rate comparison; (**f**) Weighted blocking rate of different locations; (**g**) Total cost of charging power loss ($); (**h**) Discount in PIM ($); (**i**) Total profit comparison ($).

penalty of blocking customers, the energy loss, communication delay, and maintenance cost of charging station, the profit of station k at charging level j at time t is

$$\begin{aligned} R^{HPC}_{k,j,t} &= (1-b^{HPC}_{k,t})\lambda_{k,t}(1-(1-\beta)\theta(g_{k,t}))(A_j - U_t d) \\ &+(1-b^{HPC}_{k,t})\cdot \acute{\lambda}^{HPC}_{k,t-1}\cdot(A_j - g_{k,t-1} - U_t \cdot d) \\ &-\lambda_{k,t}\cdot(1-b^{HPC}_{k,t})\cdot \frac{n^{HPC*}_{k,t} - n_{k,t}}{n^{HPC*}_{k,t}}\cdot U_t\cdot \sigma \cdot d \\ -b^{HPC}_{k,t}&\cdot \lambda^{HPC}_{k,t}\cdot p_{pen} - C^{k,t}_{loss}(\lambda^{HPC}_{k,t}) - C^{k,t}_{delay}(\lambda_{k,t}) - p_m, \end{aligned} \tag{29}$$

whereby the first term is the profit gained from the EVs which arrive at the current time t and are served. The second term is the profit gained from the EVs from the previous time slot $t-1$ which are served at the current time, whereby the discount is applied. The third term is the penalty for extra energy. The formula $\lambda_{k,t}\cdot(1-b^{HPC}_{k,t})\cdot\frac{n^{HPC*}_{k,t}-n_{k,t}}{n^{HPC*}_{k,t}}$ is the number of customers using the extra power to charge EVs. The fourth term is composed of the penalty of blocking EVs, the cost of energy loss, the penalty of communication delay, and the maintenance cost of a station. The total profit at time t of a network of charging stations is the sum of each station's profit in Eq. (29). The HPC method can also be applied to the case when the traffic load is extremely heavy. If the traffic is so heavy that the QoS (e.g., low blocking probability) cannot be maintained for the maximum SO discount ($g^{H,j}_{\max}$ and $g^{P,j}_{\max}$), then the HPC can be employed to alleviate the situation.

III. Numerical Results

The performance of the proposed PIM and CEM with respect to a baseline approach without load shifting nor capacity expansion is evaluated in Section III-A. The performance of the HPC is evaluated in Section III-B.

A. Comparison of PIM, CEM, and a Baseline Approach

This subsection compares the performance of the PIM and CEM with a baseline approach (without load shifting or capacity expansion). We note that this evaluation is fundamentally different from the evaluation in [1] in that the evaluation in [1] compared for each time slot (i.e., each hour) the maximum of the PIM and CEM to a baseline approach. In contrast, we consistently evaluate and present both the PIM and CEM for each time slot in this section.

We consider $T = 24$ hours with an hour as the charging system time unit. The performance metrics include the end-to-end delay D^t_{e2e}, cloud server allocations $c_{L,t}$ and $c_{R,t}$,

TABLE II: Parameters, their description, and values

Parameter	Description	Values
p_L, p_R	Rental fee local, remote cloud server	0.15, 0.1 \$/h
P^1_{ch}, P^2_{ch}	Charging power levels	50, 19.2 kW
A_1, A_2	Admission (charging) fees of EVs	10\$, 8\$ [61]
ϵ_L, ϵ_R	Waiting time thresholds	0.5 s, 3 s
D_H, D_P	Delay thresholds for CS_H, CS_P	2 s, 10 s
ζ_H, ζ_P	Penalty factors	0.2\$, 0.1\$
$g^{H,1}_{\max}, g^{H,2}_{\max}$	Max. discounts for highway EVs	4\$, 3\$ [58]
$g^{P,1}_{\max}, g^{P,2}_{\max}$	Max. discounts for parking lot EVs	2\$, 1.5\$ [58]
D^L_{prop}, D^R_{prop}	Prop. delay of local, remote cloud	0.01, 50 ms [41]
SoC_i	Initial state of charge of an EV	$\mathcal{N}(30, 15)$ (%) [62]
η_1, η_2	Charging effic. of charging station	88.7, 99 % [63]
R_c	Link transmission rate	10 Gbps [41]
P_{req}, P_{res}	Packet size	256 Bytes [64]
D^e_{prop}	Prop. del. between EV and BS	0.33 μs [41]
J	Charging levels in charging station	2
ϵ	Blocking threshold	0.1
p_m	Maint. cost of charging station	0.05 \$/h
S	Real-time electricity supply	5 MW
σ	Price penalty factor	0.7 [60]
p_{pen}	Penalty for not serving an EV	2\$
W	Max. # transm. attempts	8 [56]
ΔT_s	Time slot durat. of cont. process	10 ms
R_{BS}	Average wireless transmission rate	2 Mbps
SoC_r	Requested SoC of an EV	85%
α_H, α_P	Time sensitivity parameter for highway exit and parking lot	1, 0.7
μ_L, μ_R	Service rate of servers at local and remote clouds	2400, 600
Z_{1-5}	Power cap. lim. of a charging sta.	4000 kW
O	Number of preambles	54 [56]

hourly charging demand $d_{i,t}$, charging outlets allocation $n_{k,t}$, discount $g^*_{k,t}$, weighted blocking rate $b^u_{k,t}$, and total profit $Y^*_{PIM,t}$. Two highway exit charging stations CS_H and three parking lot charging stations CS_P are considered and denoted as CS_1–CS_5, respectively. Specifically, we suppose that CS_1 as well as CS_4 and CS_5 are direct-current (DC) fast chargers with charging level $j = 1$, while CS_2 and CS_3 are level 2 three-phase chargers with charging level $j = 2$. The 2017 Nissan leaf with 30 kWh Li-ion battery is considered. The scheduled supply in the network of charging stations is constant $S = 5$ MW, and the EV arrivals to the five charging stations are [500, 400, 400, 500, 300] EVs per day. The hourly arrival time distribution of CS_H and CS_P follows [65], [66]. In 4G LTE networks, both downlink and uplink average wireless transmission rates are set to $R_{BS} = 2$ Mbps. Other parameters are summarized in Table II.

Fig. 3 presents the obtained results. During off-peak hours [1, 7] and [12, 24], the statistics are the same in all cases (Baseline, PIM, and CEM) because the supply is sufficient to satisfy the charging demands of the EVs and to ensure high QoS. However, the results for the three cases differ in hours [8, 11] when the system cannot satisfy all EV charging demands.

The number of EVs requesting to charge the battery per hour for three cases (Baseline, PIM, and CEM) in the network of charging stations is shown in Fig. 3a. The charging demands of every hour for baseline and the CEM are the same because no loads are shifted. In contrast, the PIM shifts some load from the peak hours 8 and 9 to off-peak hours, so as to satisfy the QoS requirement in Eq. (19).

Fig. 3b depicts the allocation of the numbers of local ($c_{L,t}$) cloud and remote ($c_{R,t}$) cloud servers, which correspond to their service rates, the average waiting time thresholds, and the numbers of EVs in Eqs. (1)–(6). More cloud servers are allocated to the remote cloud than the local cloud during peak hours [8, 9]. This is because of the comparably larger number of EVs in parking lots during peak hours and the smaller remote cloud service rate. In Fig. 3c, a large number of EVs request to charge during peak hours [8, 9], hence D_{con} in Eq. (18) is relatively high. It is noticeable that D^R_{e2e} (see Eq. (16)) at time $t = 14$ is higher than at peak hours [8, 9], because the dominant parameter t^R_w at time $t = 14$ is larger than at [8, 9]. In particular, the number of remote cloud servers c_R at time $t = 14$ is smaller than at [8, 9] (see Fig. 3b), resulting in the larger t^R_w.

Due to the CEM characteristics, the numbers of charging outlets $n_{k,t}$ both in DC fast charging stations and Level 2 three phase charging stations are the largest among the three cases during peak hours [8, 9], as shown in Fig. 3d. Fig. 3e compares the total weighted blocking rates for the three cases. During the peak hours [8, 9], the baseline charging load is heavy (over 350 EVs requesting demand), resulting in high blocking probabilities (over 20%). On the contrary, the PIM alleviates the heavy load during peak hours by load shifting. The total weighted blocking rate is within 0.1 (threshold in Eq. (19)) with the PIM and CEM. Fig. 3f depicts the detailed blocking rates of charging stations in different locations (highway exits and parking lots).

In Fig. 3g, the cost of the total charging power loss in a network of stations per hour $\sum\limits_k C^{k,t}_{loss}(\lambda_{k,t})$ in the CEM is the highest due to more EVs being served during peak hours [8, 9]. In hours 9 and 10, the PIM cost is relatively higher because the load from the two peak hours was shifted to hours [9, 10].

The discounts $g^*_{k,t}$ offered in the PIM are depicted in Fig. 3h. To satisfy the arrival rate constraint (24) and maximize the profit (23), the optimal discounts of CS_1 at hours 8 and 9 are [\$1.6, \$2.13]. As discussed in Section II-E, customers in CS_H are time sensitive and only delay charging if offered a large discount. Hence, we observe a greater discount for CS_H than for CS_P in Fig. 3h.

Compared to the total daily profit of the baseline approach, the PIM and CEM approaches increase the total profit by 6.9% and 3.9%, respectively. Fig. 3i shows the hourly profit, which mainly depends on the arrival rates, the average blocking rates, and the real-time electricity price. The CEM profits during hours 8 and 9 are larger than for the other two cases because more EVs can be served. At hours 9 and 10, the baseline profit

TABLE III: The HPC performance changes relative to the PIM as a function of percentage β of customers who require the CEM (i.e., do not participate in the PIM).

β	Total profit change (%)	Avg. discount change (%)	Avg. # charg. outl. change (%)
0.25	0.26	-25.5	$+10.9$
0.5	-0.85	-33.4	$+21.7$
0.75	-1.4	-52.5	$+28.2$

is slightly higher than the PIM profit, due to the relatively larger demand served with the baseline approach and the PIM discount. However, the total PIM profit for 24 hours is 6.9% higher than the baseline profit. The PIM guarantees a low blocking probability while gaining a higher profit. This is crucial for satisfying the charging demands of the EVs and for gaining a high reputation for the SO.

On one hand, the obtained results show that by purchasing more energy, the CEM can ensure high system performance and gain higher profit compared to the baseline. However, the CEM burdens the load of the grid, which is not good for the long-term grid development. On the other hand, the PIM is able to alleviate the heavy load and maintain high system performance (low blocking probability and high system profit) without increasing the grid load.

B. Hybrid Approach of PIM and CEM (HPC)

In this section, we evaluate the HPC performance for the settings from Table II. We vary the percentage β of customers who adopt the CEM (with $100 - \beta$ percent of the customers participating in the PIM). Table III reports the percentage changes achieved by the HPC relative to the performance of the pure PIM (for $\beta = 0$). In particular, Table III reports the percentage change of the total profit in a network of charging stations in a day (24 hours), the percentage change of the average discount offered to each customer that participates in the PIM during peak hours [8-10], and the percentage change of the average number of charging outlets.

When $\beta = 25\%$ of the new customers in the HPC require the CEM (i.e., cannot delay charging due to time pressures), the total profit increases by 0.26% compared to the PIM profit, i.e., the profit very slightly increases. This result indicates that the hybrid HPC strategy for a realistic scenario (where some customers are under time pressures and cannot delay charging) achieves a total profit that is comparable to the profit in a pure PIM scenario (where all customers are potentially willing to delay charging if offered a high discount). This is because the offered discount is 25.5% lower (thus less income is lost due to giving out discounts), while the penalty for purchasing extra power is still comparably low as indicated by the relatively modest around 11% increase of the charging outlets. The goal of the HPC is not to try to improve upon the profit of the pure PIM. Instead, the HPC accommodates a more realistic scenario (e.g., some percentage of customers cannot delay charging no matter what discount is offered) and balances the charging needs of customers with the need to perform demand control.

When the percentage β of customers that absolutely require immediate charging increases to 50% and 75%, then the total profit slightly decreases. This is because the penalty for purchasing additional power increases substantially (by a factor of roughly two as β increases from 0.25 to 0.5), while the offered discounts decrease relatively slower (by a factor of less than 1.5 as β increases from 0.25 to 0.5). Overall, the HPC method can balance the immediate charging needs of customers with the alleviation of the power grid load through demand control based on discounts for delayed charging.

IV. Conclusion

In this paper, we have proposed a hierarchical charging model for heterogeneous EVs. To accommodate the diverse service requirements of customers, we considered a two-layered cloud computing infrastructure consisting of local and remote clouds. To solve the issues of heavy load demands and uneven charging demands at charging stations, we proposed to combine cloud server planning with capacity planning in charging stations and profit maximization in the model design. Different EV service requirements, end-to-end delays, and different charging levels have been considered and analyzed. Given the QoS metrics, the proposed models guarantee that only a low percentage of customers is not getting served. The obtained results demonstrate the efficiency of our models. The system profits increase with both the PIM and CEM, and more customers can be served compared to the baseline case with uncontrolled customers. The model ensures EVs' blocking probability and the queuing time in the clouds are bounded by a small threshold. Further, to balance the immediate charging needs of customers with the alleviation of the power grid load through demand control, a hybrid method of the PIM and CEM has been proposed. A potential direction for future work is to involve the real energy consumption profiles and energy storage systems to improve the system model and performance.

References

[1] C. Kong, B. P. Rimal, B. P. Bhattarai, and M. Devetsikiotis, "Cloud-based charging management of electric vehicles in a network of charging stations," in *Proc. IEEE Int. Conf. on Commun.*, 2018, pp. 1–6.

[2] I. S. Bayram and A. Tajer, *Plug-in Electric Vehicle Grid Integration*. Artech House, Norwood, MA, 2017.

[3] Bloomberg, "Cumulative Global EV Sales Hit 4 Million," https://about.bnef.com/blog/cumulative-global-ev-sales-hit-4-million/, 2018, [Accessed Feb-2019].

[4] International Energy Agency, "Global EV Outlook 2017," https://www.iea.org/publications/freepublications/publication/GlobalEVOutlook2017.pdf, 2017, [Accessed Oct-09-2017].

[5] K. Clement-Nyns, E. Haesen, and J. Driesen, "The impact of charging plug-in hybrid electric vehicles on a residential distribution grid," *IEEE Trans. Power Syst.*, vol. 25, no. 1, pp. 371–380, 2010.

[6] Y. He, B. Venkatesh, and L. Guan, "Optimal scheduling for charging and discharging of electric vehicles," *IEEE Trans. Smart Grid*, vol. 3, no. 3, pp. 1095–1105, 2012.

[7] H. Hu, Y. Wen, L. Yin, L. Qiu, and D. Niyato, "Coordinating workload scheduling of geo-distributed data centers and electricity generation of smart grid," *IEEE Transactions on Services Computing, in print*, 2020.

[8] R. Kaewpuang, S. Chaisiri, D. Niyato, B. Lee, and P. Wang, "Cooperative virtual machine management in smart grid environment," *IEEE Transactions on Services Computing*, vol. 7, no. 4, pp. 545–560, Oct 2014.

[9] W. Wu, W. Wang, X. Fang, L. Junzhou, and A. V. Vasilakos, "Electricity price-aware consolidation algorithms for time-sensitive VM services in cloud systems," *IEEE Transactions on Services Computing, in print*, 2020.

[10] D. S. Markovic, D. Zivkovic, I. Branovic, R. Popovic, and D. Cvetkovic, "Smart power grid and cloud computing," *Renewable and Sustainable Energy Reviews*, vol. 24, pp. 566–577, 2013.

[11] B. P. Rimal and I. Lumb, "The rise of cloud computing in the era of emerging networked society," in *Cloud Computing: Principles, Systems and Applications*. Springer, Cham, Switzerland, 2017, pp. 3–25.

[12] Q. Wang, S. Guo, J. Liu, C. Pan, and L. Yang, "Profit maximization incentive mechanism for resource providers in mobile edge computing," *IEEE Transactions on Services Computing, in print*, pp. 1–1, 2020.

[13] Fortune, "Cloud computing will keep these electric vehicle chargers humming," http://fortune.com/2015/10/22/cloud-electric-vehicle-charger/, 2015, [Accessed Jan-10-2017].

[14] D. A. Chekired and L. Khoukhi, "Smart grid solution for charging and discharging services based on cloud computing scheduling," *IEEE Trans. on Industrial Inform.*, vol. 13, no. 6, pp. 3312–3321, Dec. 2017.

[15] S. Bae and A. Kwasinski, "Spatial and temporal model of electric vehicle charging demand," *IEEE Trans. Smart Grid*, vol. 3, no. 1, pp. 394–403, 2012.

[16] I. S. Bayram, M. Ismail, M. Abdallah, K. Qaraqe, and E. Serpedin, "A pricing-based load shifting framework for EV fast charging stations," in *Proc., IEEE SmartGridComm*, 2014, pp. 680–685.

[17] O. Hafez and K. Bhattacharya, "Integrating EV charging stations as smart loads for demand response provisions in distribution systems," *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 1096–1106, 2018.

[18] M. Ş. Kuran, A. C. Viana, L. Iannone, D. Kofman, G. Mermoud, and J. P. Vasseur, "A smart parking lot management system for scheduling the recharging of electric vehicles," *IEEE Trans. Smart Grid*, vol. 6, no. 6, pp. 2942–2953, 2015.

[19] P. Rezaei, J. Frolik, and P. D. Hines, "Packetized plug-in electric vehicle charge management," *IEEE Trans. Smart Grid*, vol. 5, no. 2, pp. 642–650, 2014.

[20] E. Yao, V. W. Wong, and R. Schober, "Robust frequency regulation capacity scheduling algorithm for electric vehicles," *IEEE Trans. Smart Grid*, vol. 8, no. 2, pp. 984–997, 2017.

[21] W. Tang and Y. J. Zhang, "Online electric vehicle charging control with multistage stochastic programming," in *Proc., IEEE CISS*, 2014, pp. 1–6.

[22] J. Zhao, J. Wang, Z. Xu, C. Wang, C. Wan, and C. Chen, "Distribution network electric vehicle hosting capacity maximization: A chargeable region optimization model," *IEEE Transactions on Power Systems*, vol. 32, no. 5, pp. 4119–4130, Sep. 2017.

[23] J. Yan, H. Zheng, and N. Lu, "Temperature-load sensitivity study for adjusting MISO day-ahead load forecast," in *Proc. Power and Energy Society General Meeting (PESGM)*, 2016, pp. 1–5.

[24] M. C. Kisacikoglu, F. Erden, and N. Erdogan, "Distributed control of PEV charging based on energy demand forecast," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 1, pp. 332–341, 2018.

[25] B. Hashemi, M. Shahabi, and P. Teimourzadeh-Baboli, "Stochastic-based optimal charging strategy for plug-in electric vehicles aggregator under incentive and regulatory policies of dso," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3234–3245, 2019.

[26] R. Mehta, D. Srinivasan, A. M. Khambadkone, J. Yang, and A. Trivedi, "Smart charging strategies for optimal integration of plug-in electric vehicles within existing distribution system infrastructure," *IEEE Transactions on Smart Grid*, vol. 9, no. 1, pp. 299–312, 2018.

[27] Y. Cao, T. Wang, O. Kaiwartya, G. Min, N. Ahmad, and A. H. Abdullah, "An ev charging management system concerning drivers trip duration and mobility uncertainty," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 4, pp. 596–607, 2018.

[28] B. Sun, Z. Huang, X. Tan, and D. H. K. Tsang, "Optimal scheduling for electric vehicle charging with discrete charging levels in distribution grid," *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 624–634, 2018.

[29] H. Yang, Y. Deng, J. Qiu, M. Li, M. Lai, and Z. Y. Dong, "Electric vehicle route selection and charging navigation strategy based on crowd sensing," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 5, pp. 2214–2226, 2017.

[30] X. Tang, S. Bi, and Y. A. Zhang, "Distributed routing and charging scheduling optimization for internet of electric vehicles," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 136–148, 2019.

[31] J. Liu, H. Guo, J. Xiong, N. Kato, J. Zhang, and Y. Zhang, "Smart and resilient EV charging in SDN-enhanced vehicular edge computing networks," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 1, pp. 217–228, 2020.

[32] H. Kim, Y.-J. Kim, K. Yang, and M. Thottan, "Cloud-based demand response for smart grid: Architecture and distributed algorithms," in *Proc. IEEE SmartGridComm*, 2011, pp. 398–403.

[33] S. Bera, T. Ojha, S. Misra, and M. S. Obaidat, "Cloud-based optimal energy forecasting for enabling green smart grid communication," in *Proc., IEEE GLOBECOM*, 2015, pp. 1–6.

[34] M. Ammous, S. Belakaria, S. Sorour, and A. Abdel-Rahim, "Optimal cloud-based routing with in-route charging of mobility-on-demand electric vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 7, pp. 2510–2522, 2019.

[35] Y. Cao, H. Song, O. Kaiwartya, B. Zhou, Y. Zhuang, Y. Cao, and X. Zhang, "Mobile edge computing for big-data-enabled electric vehicle charging," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 150–156, 2018.

[36] N. Kumar, S. Zeadally, and J. J. P. C. Rodrigues, "Vehicular delay-tolerant networks for smart grid data management using mobile edge computing," *IEEE Communications Magazine*, vol. 54, no. 10, 2016.

[37] S. Cowley, S. Singh, A. Krishna, and L. Kesterson-Townes, "Assembling your cloud orchestra: A field guide to multicloud management, Executive Report, IBM Institute for Business Value," Oct. 2018, last accessed June 11, 2020. [Online]. Available: https://www.ibm.com/thought-leadership/institute-business-value/report/multicloud#

[38] A. J. Ferrer, J. M. Marquès, and J. Jorba, "Towards the decentralised cloud: Survey on approaches and challenges for mobile, ad hoc, and edge computing," *ACM Computing Surveys (CSUR)*, vol. 51, no. 6, pp. 1–36, 2019.

[39] J. L. Lucas-Simarro, R. Moreno-Vozmediano, R. S. Montero, and I. M. Llorente, "Scheduling strategies for optimal service deployment across multiple clouds," *Future Generation Computer Systems*, vol. 29, no. 6, pp. 1431–1441, 2013.

[40] M. Mehrabi, D. You, V. Latzko, H. Salah, M. Reisslein, and F. H. Fitzek, "Device-enhanced MEC: Multi-access edge computing (MEC) aided by end device computation and caching: A survey," *IEEE Access*, vol. 7, pp. 166 079–166 108, 2019.

[41] B. P. Rimal, D. P. Van, and M. Maier, "Mobile-edge computing versus centralized cloud computing over a converged FiWi access network," *IEEE Trans. Network and Service Management*, vol. 14, no. 3, pp. 498–513, Sept 2017.

[42] B. P. Rimal, M. Maier, and M. Satyanarayanan, "Experimental testbed for edge computing in fiber-wireless broadband access networks," *IEEE Communications Magazine*, vol. 56, no. 8, pp. 160–167, 2018.

[43] ——, "Experimental testbed for edge computing in fiber-wireless broadband access networks," *IEEE Communications Magazine*, vol. 56, no. 8, pp. 160–167, August 2018.

[44] Z. Xiang, F. Gabriel, E. Urbano, G. T. Nguyen, M. Reisslein, and F. H. Fitzek, "Reducing latency in virtual machines: Enabling tactile internet for human-machine co-working," *IEEE J. on Sel. Areas in Commun.*, vol. 37, no. 5, pp. 1098–1116, 2019.

[45] M. J. Hodgson, "A flow-capturing location-allocation model," *Geographical Analysis*, vol. 22, no. 3, pp. 270–279, 1990.

[46] B. P. Rimal, D. P. Van, and M. Maier, "Cloudlet enhanced fiber-wireless access networks for mobile-edge computing," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3601–3618, 2017.

[47] J. Cao, K. Hwang, K. Li, and A. Y. Zomaya, "Optimal multiserver configuration for profit maximization in cloud computing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 6, pp. 1087–1096, 2013.

[48] L. Kleinrock, *Queueing Systems, Volume 2: Computer Applications*. Wiley, New York, 1976.

[49] R. W. Erickson and D. Maksimovic, *Fundamentals of Power Electronics*. Springer Science & Business Media, 2007.

[50] E. Apostolaki-Iosifidou, P. Codani, and W. Kempton, "Measurement of power loss during electric vehicle charging and discharging," *Energy*, vol. 127, pp. 730–742, 2017.

[51] A. O. Allen, *Probability, Statistics, and Queueing Theory*. Academic Press, 2014.

[52] P. Fan, B. Sainbayar, and S. Ren, "Operation analysis of fast charging stations with energy demand control of electric vehicles," *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 1819–1826, 2015.

[53] C. Kong, I. S. Bayram, and M. Devetsikiotis, "Revenue optimization frameworks for multi-class PEV charging stations," *IEEE Access*, vol. 3, pp. 2140–2150, 2015.

[54] C. Kong, R. Jovanovic, I. S. Bayram, and M. Devetsikiotis, "A hierarchical optimization model for a network of electric vehicle charging stations," *Energies*, vol. 10, no. 5, pp. 675.1–675.20, 2017.

[55] 3GPP, "LTE; evolved universal terrestrial radio access (E-UTRA); medium access control (MAC) protocol specification, TS 36.321, Ver. 9.2.0, Release 9, Apr. 2010."

[56] R. R. Tyagi, F. Aurzada, K.-D. Lee, S. G. Kim, and M. Reisslein, "Impact of retransmission limit on preamble contention in LTE-advanced network," *IEEE Systems Journal*, vol. 9, no. 3, pp. 752–765, 2015.

[57] N. Keon and G. Anandalingam, "A new pricing model for competitive telecommunications services using congestion discounts," *INFORMS J. Computing*, vol. 17, no. 2, pp. 248–262, 2005.

[58] D. Ban, G. Michailidis, and M. Devetsikiotis, "Demand response control for PHEV charging stations by dynamic price adjustments," in *Proc. IEEE ISGT*, 2012, pp. 1–8.

[59] M. G. Kallitsis, G. Michailidis, and M. Devetsikiotis, "A framework for optimizing measurement-based power distribution under communication network constraints," in *Proc. IEEE SmartGridComm*, 2010, pp. 185–190.

[60] C. Jin, J. Tang, and P. Ghosh, "Optimizing electric vehicle charging with energy storage in the electricity market," *IEEE Trans. Smart Grid*, vol. 4, no. 1, pp. 311–320, 2013.

[61] Blink Charging, "EV Charging Fees," http://www.blinknetwork.com/drivingelectric, 2008, [Accessed Nov-27-2017].

[62] Z. Luo, Z. Hu, Y. Song, Z. Xu, and H. Lu, "Optimal coordination of plug-in electric vehicles in power grids with cost-benefit analysis-Part II: A case study in China," *IEEE Trans. Power Syst.*, vol. 28, no. 4, pp. 3556–3565, 2013.

[63] "ENERGY STAR Market and Industry Scoping Report Electric Vehicle Supply Equipment (EVSE)," https://www.energystar.gov/sites/default/files/asset/document/Electric_Vehicle_Scoping_Report.pdf, 2008, [Accessed Dec-16-2017].

[64] Z. H. Mir and F. Filali, "Lte and ieee 802.11 p for vehicular networking: a performance evaluation," *EURASIP Journal on Wireless Communications and Networking*, vol. 2014, no. 1, p. 89, 2014.

[65] H. Cai, X. Jia, A. S. Chiu, X. Hu, and M. Xu, "Siting public electric vehicle charging stations in beijing using big-data informed travel patterns of the taxi fleet," *Transportation Research Part D: Transport and Environment*, vol. 33, pp. 39–46, 2014.

[66] Y. Guo, J. Xiong, S. Xu, and W. Su, "Two-stage economic operation of microgrid-like electric vehicle parking deck," *IEEE Trans. Smart Grid*, vol. 7, no. 3, pp. 1703–1712, 2016.

Cuiyu Kong received the B.S degree in Information Engineering from South China University of Technology, Guangzhou, China in 2011, M.S degree in Communication Engineering from National Central University, Taiwan in 2013, and Ph.D degree in the department of Electrical and Computer Engineering in North Carolina State University, USA in 2018. Currently she is a faculty in the department of Computer Science and Computer Information Science at Highline College.

Bhaskar Prasad Rimal (Senior Member, IEEE) received the Ph.D. degree in telecommunications engineering from the University of Quebec, Canada. He was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, USA. He was a Visiting Scholar with the Department of Computer Science, Carnegie Mellon University (CMU), Pittsburgh, USA, in 2016. From 2018 to 2019, he was a Visiting Assistant Professor and Academic Specialist at the Department of Computer Science at Tennessee Tech University, Cookeville, USA.

Martin Reisslein (Fellow, IEEE) received the Ph.D. degree in systems engineering from the University of Pennsylvania, Philadelphia, in 1998. He is currently a Professor with the School of Electrical, Computer, and Energy Engineering, Arizona State University (ASU), Tempe. He serves as an Associate Editor-in-Chief of the *IEEE Communications Surveys and Tutorials* and a Co-Editor-in-Chief of *Optical Switching and Networking*.

Martin Maier (Senior Member, IEEE) is a full professor with the Institut National de la Recherche Scientifique (INRS), Montreal, Canada. He was educated at the Technical University of Berlin, Germany, and received M.Sc. and Ph.D. degrees (summa cum laude) in 1998 and 2003, respectively. In the summer of 2003 he was a postdoc fellow at the Massachusetts Institute of Technology (MIT), Cambridge. He was a visiting professor at Stanford University, Stanford, from October 2006 through March 2007. Further, he was a co-recipient of the 2009 IEEE Communications Society Best Tutorial Paper Award. He was a Marie Curie IIF Fellow of the European Commission from March 2014 through February 2015. In March 2017, he received the Friedrich Wilhelm Bessel Research Award from the Alexander von Humboldt (AvH) Foundation in recognition of his accomplishments in research on FiWi enhanced networks. In May 2017, he was named as one of the three most promising scientists in the category "Contribution to a better society" of the Marie Skodowska-Curie Actions (MSCA) 2017 Prize Award of the European Commission. He is the founder and creative director of the Optical Zeitgeist Laboratory (www.zeitgeistlab.ca).

Islam Safak Bayram received his B.S. degree in electrical and electronics engineering from Dokuz Eylul University, Izmir, Turkey in 2007, his M.S. degree in telecommunications from the University of Pittsburgh in 2010, and his Ph.D. degree in electrical and computer engineering from North Carolina State University, in 2013. He received the Best Paper Awards at the Third IEEE International Conference on Smart Grid Communications and the First IEEE Workshop on Renewable Energy and Smart Grid in March 2015. From January to December 2014, he worked as a postdoctoral research scientist at Texas A& M University at Qatar. From 2015 to 2018, he was an assistant professor at College of Science and Engineering and staff scientist at Qatar Environment and Energy Research Institute both at Hamad Bin Khalifa University. Since 2018, he is a Research Assistant Professor in the Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, USA.

Michael Devetsikiotis (Fellow, IEEE) was born in Thessaloniki, Greece. He received the Diploma degree in Electrical Engineering from the Aristotle University of Thessaloniki, Greece, in 1988, and the M.S. and Ph.D. degrees in Electrical Engineering from North Carolina State University, Raleigh, in 1990 and 1993, respectively. In 1993 he joined the Department of Systems and Computer Engineering at Carleton University, Ottawa, Ontario, Canada, as a Post-Doctoral Fellow. He became a tenure track Assistant Professor in 1996, and an Associate Professor and Department Associate Chair in 1999.

Michael returned to the Department of Electrical and Computer Engineering at North Carolina State as an Associate Professor in 2000, and became a Professor in 2006. He served as the coordinator of the Masters of Science in Computer Networking until 2011, when he became the ECE Director of Graduate Programs, managing one of the largest graduate ECE programs in the country with over 800 students. On August 1, 2016, Michael joined the University of New Mexico as a Professor, and the Chair of the ECE Department in the School of Engineering.

His research work has resulted in 39 published refereed journal articles, 128 refereed conference papers, and 61 invited presentations, in the area of design and performance evaluation of telecommunication networks, complex socio-technical systems, and smart grid communications. Michael joined the IEEE as a student member in 1985, and he became a Fellow in 2012. He has served as Chairman of the IEEE Communications Society Technical Committee Communication Systems Integration and Modeling, and as a member of the IEEE ComSoc Education Board. Between 2008 and 2011 he was an IEEE ComSoc Distinguished Lecturer. He has also served as an Associate or Area Editor of several publications of the IEEE and the ACM; and as technical program committee Chair and in other roles for numerous conferences.