Joint Energy-QoE Efficient Content Delivery Networks Using Real-Time Energy Management

Pejman Goudarzi, Abolfazl Ghassemi, *Senior Member, IEEE*, Mohammad R. Mirsarraf, and Rajkumar Buyya, *Fellow, IEEE*

***Abstract*—In this paper, a joint optimization method has been proposed for energy cost and the user's perceived quality-of-experience (QoE) within a content delivery network (CDN). The proposed algorithm has been developed with an objective to minimize the total operational cost using real-time electricity pricing as well as the integration of green energy resources from the smart grid. To solve the joint optimization problem, a linear programming and a differential evolution algorithm have been used to provide a tradeoff between operational cost saving and computational complexity. We also formulate the joint problem as a noncooperative game in which the CDN providers act as players. A group utility function has been defined for the players and it is shown that a Nash equilibrium exists for the joint quality and energy cost saving when each CDN provider has single cluster or multiple clusters. Numerical results are presented to evaluate and validate the aforementioned solutions. These results illustrate significant operational/energy cost reductions while optimizing user's perceived QoE for a CDN or multiple CDN providers with respect to conventional methods.**

***Index Terms*—Differential evolution (DE), game theory, green content delivery network (CDN), quality-of-experience (QoE), request routing, smart grid.**

I. Introduction

THE rise of broadband multimedia services has a large share of the current Internet traffic. This leads to increased consumer demands for rich multimedia content from popular content providers, such as Netflix, Amazon, Hulu, or Youtube, and have eventually motivated the service providers to develop content delivery/distribution networks (CDNs).

A CDN typically uses a large number of distributed servers for the efficient delivery of web content and streaming media to global consumers. These servers are deployed in multiple clusters across the Internet in which each cluster includes a number of colocated edge servers in *a data centre*. Sizes of these clusters are varied from tens of servers to thousands of servers. As a result, data centres within a CDN can have a substantial amount of power consumption which can significantly increase energy costs. As an example, data centres in the USA consume 100 billion kWh or 7.4 billion dollars annually [1]. Many energy efficient techniques within data centres can reduce the energy consumption of a CDN, e.g., more than 55% in [2], while meeting quality-of-service (QoS) requirements. Other factors related to the energy loss in CDNs are inefficient load distribution between CDN clusters, inefficient content placement, and inefficient use of cooling system [3].

Moreover, the number of demand applications, particularly video streaming traffic [e.g., video on demand (VoD), online gaming, and IP television (IPTV)], has a substantial impact on the energy consumption of clusters, and can dynamically change according to the number of user demands. Furthermore, existing request redirection techniques in a CDN typically do not consider the energy consumption of the system and may result in inefficient total energy usage.

The above-mentioned factors increase the motivation for the development of energy-saving techniques as well as finding new solutions for alternative power supplies. Moreover, the new generation of power grid (the *smart grid*) enforces important features such as the integration of renewable generator and active consumer participation to near-instantaneously balance demand and power supply [4]. The first feature promotes the use of green and sustainable energy sources such as solar panels and wind. The latter feature enables the consumers to efficiently manage their energy consumption as well as track electricity price variations dynamically.

Based on the mentioned facts, we must use a smart grid-enabled real-time energy management facility for CDN. To describe more clearly, in the current paper, the total power consumption cost and quality-of-experience (QoE) [5] degradation cost within all server clusters must be jointly minimized for each time slot.

CDN providers typically employ geo-location or proximity-based information for request redirection [6]. However, this does not necessarily lead to energy cost saving and end user's QoE satisfaction.

We propose a joint energy cost and user perceived QoE optimization method for CDN. To do this, two different scenarios, single web/VoD traffic and mixed traffic types (VoD and web) have been considered. It is proved that there might be multiple optimal request redirection strategies for the first scenario while a single and optimal request routing strategy can exist for the second scenario. We have used a linear programming (LP), an evolutionary algorithm such as differential evolution (DE), and a noncooperative game theoretic approach to solve the proposed joint optimization problem.

While LP provides a low-complexity solution for joint optimization problem, DE technique reaches to global optimal

Manuscript received September 12, 2018; revised January 26, 2019, February 26, 2019, and June 17, 2019; accepted June 18, 2019. Date of publication July 9, 2019; date of current version March 2, 2020. *(Corresponding author: Pejman Goudarzi.)*

P. Goudarzi, A. Ghassemi, and M. R. Mirsarraf are with the Department of Information Technology, ICT Research Institute (ITRC), Tehran 1439955471, Iran (e-mail: pgoudarzi@itrc.ac.ir; a.ghassem@itrc.ac.ir; m.mirsaraf@itrc.ac.ir).

R. Buyya is with the Cloud Computing and Distributed Systems Lab, The University of Melbourne, Victoria, VIC 3010, Australia (e-mail: rbuyya@unimelb.edu.au).

Digital Object Identifier 10.1109/JSYST.2019.2924186

solution with large energy cost saving. In this case, we will show that both solutions to the joint optimization problem can provide an operational cost saving versus complexity tradeoff for various total number of solar panels. A noncooperative game is also considered for various CDN providers. They can create multiple groups to exchange their VoD/web connections according to the availability of the green energy resources. We show that a Nash equilibrium (NE) exists for the joint optimization problem for multiple scenarios.

Moreover, the performance of the proposed system has been investigated under real-time pricing regime, in which, the hourly operational cost for the proposed joint energy cost and QoE optimization has been obtained. The experimental results demonstrate that the proposed method can achieve a large amount of cost savings in comparison with the conventional systems. Also, it is shown that the proposed mechanism can allocate heterogeneous number of renewable sources, such as solar panels, to the server clusters based on assigned traffic.

The rest of the paper is organized as follows. Section II presents a review of related work. Section III includes system models and some assumptions that are used throughout the paper. Section IV is about the proposed energy-efficient content distribution techniques. Numerical results are discussed in Section V, while Section VI concludes the paper.

II. Related Work

Previous methods for reducing the power consumption of a CDN can largely be classified as content-placement strategies, energy-efficient routing schemes, energy harvesting methods, as well as cluster and/or server shutdown techniques. As the first category of energy-saving techniques for CDNs, content-placement techniques intelligently insert content into various caches to reduce the total energy consumption. Several content-placement strategies are presented in [7] for different types of CDNs over telecommunications networks to reduce energy consumption. In doing so, the authors in [7] exploit the variations between two power consumption measurements, storage and transmission, in order to achieve energy efficiency. The authors in [8] use a two-tier model for cache placement in wireless content delivery network while [9] proposes a QoE-driven cache management for HTTP adaptive bit rate streaming over wireless networks. The main objective in [8] is to reduce the complexity and latency associated with the placement of the content. Finally, [10] used monetary cost to efficiently place content in media clouds.

The second category typically uses caching techniques for developing energy-efficient routing within a CDN to reduce the total cost of energy consumption. As an example, the authors in [11] study the impact of using in-network caches and CDN cooperation on the energy consumption of request routing. They formulate this problem as energy efficient content distribution. The objective is to find a feasible routing in order to minimize the total energy consumption of the network subject to satisfying all the demands and link capacity. This paper also uses an integer LP and a heuristic algorithm to solve the raised optimization problem.

In [12], Yang *et al.* have described a learning system for implementing self-optimization-based dynamic server resource provisioning of data centers under deregulated electricity markets. They have proposed a postdecision state learning-based dynamic server resource-provisioning algorithm which has fast convergence by estimating and exploiting the workload arrival distribution. Yu *et al.* in [13], investigate the problem of developing a geographical load balancing scheme for distributed Internet data centers when they are price makers in the deregulated electricity markets. They have proposed a price-sensitivity aware geographical load-balancing scheme to address their so-called price-based challenges.

Further, the energy-efficient routing scheme proposed in [14] focuses on the problem of adaptive in-network caching for energy efficient content distribution. In this work, the authors introduce an information-centric optimization framework for energy-efficient caching in which an off-line solution based on an integer programming is obtained to maximize energy efficiency gains. These gains can be achieved using two major factors—global user requests and overall network resources. This can result in an on-line solution that allows network nodes to make caching decisions while taking into account the present estimate of global energy benefits. The authors in [15] considered the problem of minimizing the long-term energy cost for an Internet data center by joint workload and battery scheduling with heterogeneous service delay guarantees.

The third class of techniques, known as energy harvesting, use various power harvesting mechanisms to improve the energy efficiency of mobile CDNs. In doing so, the authors in [16] and [17] provide energy efficiency for content delivery via energy harvesting within small mobile cells. The proposed technique exploits the number of request arrivals for content delivery as well as energy harvesting from external power sources which are both unpredictable, therefore, proactive caching and pushing are jointly employed to address these uncertainties.

Another class of CDN energy reduction techniques has been focused on shutting down servers either partially or entirely within a cluster. This technique is considered in [18], where the entire collocated servers in a CDN cluster are turned OFF to save the power consumption. This is beneficial as the proposed approach creates energy savings for both servers as well as their cooling systems.

Other techniques that differ from the above-mentioned classes for energy reduction within a CDN, have been proposed in [19]–[23]. In [19], an integration of two network layers, metro and access, are considered for VoD service delivery. This method turns metro servers and network interfaces on and off to facilitate a balance between the energy consumption for content transport via the network and the energy consumption for processing and storage in the metro servers. In [20], energy efficiency considerations are addressed in the context of BitTorrent. The authors provide mechanisms for facilitating energy efficiency and energy proportionality, and employ these mechanisms to minimize energy consumption and consequently reduce the operational cost.

Furthermore, a video traffic routing is optimized in terms of contention-delay in [21] for software-defined interdata center networks by managing video traffic among interdata centres. To reduce end-to-end delays between clients, [23] also considers connecting multiple clients through multiple relay servers and analyzes the server selection problem from a dense pool of content delivery network edge locations and data centres. It shows that the delay optimization problem is an extension of the well known Euclidean k-median problem. Moreover, stochastic optimization techniques were employed in [24] to minimize the total cost of ownership (TCO) of CDN and video CDN overlays. The main focus of this category is on the TCO monetary cost modeling and optimization without any concentration on energy conservation.

TABLE I
PERFORMANCE COMPARISON OF THE PROPOSED DE METHOD WITH RELATED STATE-OF-THE-ART

Method	Joint QoE-Energy Maximization	Bio-inspired	Smart-grid enabled	Cloud-based	Joint Cooling/Server Energy reduction
Proposed DE	√	√	√	×	×
Hu et al. [25]	×	×	×	√	×
Yang et al. [12]	×	√	×	×	×
Yu et al. [22]	×	×	×	√	×
Araujo et al. [11]	×	×	×	×	×
Ge et al. [27]	×	×	×	×	√

Within a cloud CDN, [25] proposed joint optimization for content placement and request dispatching strategies for minimizing the bandwidth, storage, and replication costs of video streaming services under QoS constraints. The minimization of the TCO associated with cloud mobile media services under capacity/QoS constraints was also studied in [26]. In [27], Ge *et al.* propose jointly optimizing the energy consumption of both server infrastructures and cooling systems in a holistic manner.

In contrast to previous techniques [7]–[27], the proposed methods in this paper jointly optimize the total users' perceived QoE and energy consumption costs within CDN in the context of smart grid. In Table I, the performance of the proposed method is compared with some important relevant ones.

III. SYSTEM MODELS AND ASSUMPTIONS

A. Assumptions and Considerations

In the current paper, we have assumed a real-time energy management system for energy/QoE-efficient request redirection of CDN users. These users are assumed to be nonreal-time (elastic) or real-time ones. As discussed in [28], these real-time/nonreal-time users have different characteristics. In the current study, it is assumed that like Akamai, Cisco ECDS, or Amazon's Cloudfront, the CDN is designed for both web acceleration and media delivery purposes. But, as we mention in Section III-C, the real-time connections have more diverse QoE degradation effect than their nonreal-time counterparts. Moreover, the real-time connections in this paper are assumed to be of adaptive bit rate (ABR) type. In the current paper, we have used real-time traffics of VoD types, not time-sensitive and live IPTV ones.

In the proposed optimal request redirection algorithm, it is assumed that the user requests are received and aggregated during time slot t and delivered to the CDN management system. Then, at the end of each time slot, after executing the joint energy cost/QoE optimal request redirection algorithm, each request is mapped to its appropriate/optimal server cluster. After mapping each request to an appropriate server cluster, the real-time user sticks to media server and its VoD traffic can be streamed in real-time.

It is assumed in each CDN server cluster that renewable energy sources are solar panels and surplus energy can be stored in properly designed stackable batteries. It is assumed that energy cannot be sold to grid. It is also assumed that actual hourly real-time prices for 24 h are adopted and the electricity price is updated every 2 h. In the game-theoretic CDN provider interconnection section, we have assumed that CDN providers are ISP-operated and may share infrastructure for user connection request exchanging.

B. CDN Architecture

A CDN has some major components—request redirection system, caching system, management system, server clusters in each point of presence (PoP),[1] and acquisition/distribution servers. A sample CDN topology is depicted in Fig. 1 where we only consider the major elements for simplicity. A CDN architecture has usually three hierarchical layers. The first layer, the acquisition layer, receives the content from content producers and distributes it over surrogate servers. The second layer, distribution layer, performs the request redirection. CDN typically consists of server clusters distributed over diverse geographical locations. Each PoP consists of multiple surrogate servers with a limited real-time request serving capability (concurrent sessions) and throughput [29].

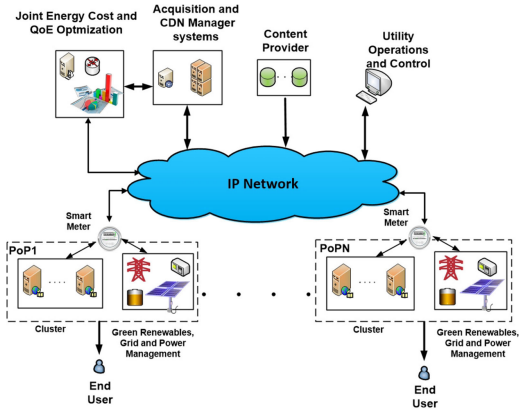

Fig. 1. Green CDN architecture.

As the last layer, the access layer provides connectivity between the users and the CDN network elements.

A CDN can deliver multiple services for the users such as multimedia services (e.g., IPTV or VoD) and elastic services (e.g., web browsing). Request redirection can be done by various mechanisms such as DNS-based, http redirection-based, geo-location-based, and server load-based mechanisms [6], [30]. Both caching and management systems are major players to do the request routing process. For example, in server load-based methods, surrogate servers inform the management system about their delivery throughput and their processing level through a network control protocol (e.g., simple network management protocol). The management system provides request redirection component with these info, and then, new incoming requests can be routed to the edge servers with minimal processing burden. This balances the edge server loads globally.

C. QoE Model

QoE presents user perception, experience, and expectations to application and network performance, and is a function of

[1]Hereafter, we use terms PoP and cluster interchangeably throughout the paper.

QoS [31]. In this paper, we consider a generic QoE model from [32], called IQX hypothesis, given as

$$\mathcal{Q} \triangleq \eta \exp\{-q\} + s \tag{1}$$

where η and s are positive constants and also, we denote the QoE by $\mathcal{Q}$ and QoS disturbance by q. Assume that $\zeta_k^v(t)$ and $\zeta_k^w(t)$ represent the number of requests/connections associated with VoD and Web applications within a cluster k at time slot t, respectively. The parameter q perceived by an end user feeding from that cluster can be considered as a weighted sum of the number of VoD and web instantaneous requests as follows:

$$q_k(t) \triangleq \nu_k \zeta_k^v(t) + \omega_k \zeta_k^w(t). \tag{2}$$

Positive constants ν_k and ω_k (normally $\omega_k << \nu_k$) represent the impact of VoD and web active requests on the users perceived QoE.[2] The selected numerical values for these parameters are selected such that the QoS disturbance values be in their standard range according to the IQX model presented in [32]. Hence, from (2) the quality perceived by users in cluster k at time slot t in (1), $\mathcal{Q}_k(t)$, can be rewritten as

$$\mathcal{Q}_k(t) = \eta \exp\{-\nu_k \zeta_k^v(t) - \omega_k \zeta_k^w(t)\} + s, \qquad k \in \mathcal{N}. \tag{3}$$

ν_k and ω_k are constants for video and web applications at cluster k. $\mathcal{N}$ is the set $\{1, 2, \ldots, N\}$ and defined as the number of clusters (PoPs). The value of $\mathcal{Q}$ can be maximized if there are no active VoD/web requests, therefore, we have $\mathcal{Q}^{\max} = \eta + s$. Without loss of generality, (1) becomes a valid model if $\mathcal{Q}^{\max}$ is accessible and the number of VoD/web requests is less than a predetermined threshold.

D. Power Consumption Model

The electric power consumed by a server within a cluster k is composed of two parts—the static part c_k' and the dynamic part c_k. c_k' is the base amount of power consumption when a server is active, idle, and ready to process an application request. c_k is a power consumption which depends on the computation load due to processing all application requests. Within a cluster, c_k' increases when adding an active server, but its c_k^d may decrease as more active servers can share the request demands; this is given by [33]

$$c_k' = e_k^{\text{idle}} + (U - 1)\, e_k^{\text{avg}} \tag{4}$$

where e_k^{idle} is the average idle power of a server at cluster k. U is the power usage effectiveness [27], [33], and e_k^{avg} is the average peak power of a server when the server processes the application request (i.e., user's demand). The dynamic power consumption can also be expressed as

$$c_k = \left(e_k^{\text{avg}} - e_k^{\text{idle}}\right) \sum_{m \in \mathcal{D}^m} \left(\frac{\zeta^m}{D^m}\right). \tag{5}$$

ζ^m is the number of randomly generated incoming requests of applications m while D^m denotes the total/serviceable amount of user's demands requesting for application m. $\mathcal{D}^m$ is the set $\{1, 2, \ldots, D^m\}$. Since static power consumption does not depend on the user's demand and our focus in this paper is on the dynamic part, we only consider the dynamic power consumption for the rest of this paper.

[2]The fact associated in using weight parameters ν_k and ω_k in (2) is that as the number of serving real-time/nonreal-time connections increases for a given cluster, the QoE for its associated end users deteriorates gradually. $\omega_k << \nu_k$ means the larger sharing of the real-time connections (which normally are typically broadband) in QoE reduction of other end users in comparison with the nonreal-time ones.

E. Real-Time Energy Management System

Within a cluster, we employ two major smart grid applications—smart metering and renewable resources, as illustrated in Fig. 1. Renewable energy generators such as solar panels are considered, while energy storage devices include batteries. The power management module performs several key functions. It manages solar power and battery storage to modulate their output power levels according to the power consumption of the cluster servers.

This module has also the capability to connect and disconnect from the electrical grid, according to measuring shortage and surplus in electrical power. In the case of disconnection, it locally balances supply and demand while also employing the power grid and renewable energy resources to balance supply and demand, when it is connected to the grid. The power management system also facilitates real-time management according to time-based pricing through bidirectional communications between a smart meter and the utility control center, as shown in Fig. 1.

IV. Energy Efficient Content Distribution

Assume that the composite random $\lambda(t)$ at time slot t is defined as: $\lambda_k(t) = [r_k(t), p_k(t), \zeta_k(t)]$; $r_k(t) \in R_k(t)$, $p_k(t) \in P_k(t)$, $\zeta_k(t) \in Z_k(t)$. $R_k(t)$, $P_k(t)$, and $Z_k(t)$ are the set of amount of power generated from renewable source, the set of spot power prices from electrical grid, and the set of users' demand at time slot t within cluster k, respectively. The random variables $r_k(t)$, $p_k(t)$, and $\zeta_k(t)$ take values from these sets. If we assume that the user requests include two major VoD and web applications within a CDN, then the random variable $Z_k(t)$ can be defined as

$$Z_k(t) = \{\zeta_k^v(t), \zeta_k^w(t)\}.$$

Assume that D^w and D^v are the total/serviceable number of web and VoD applications; so, we have, $D = D^w + D^v$. Thus, we can rewrite (5) as

$$c_k(t) = \left(e_k^{\text{avg}} - e_k^{\text{idle}}\right) \left(\frac{\zeta_k^v(t)}{D^v} + \frac{\zeta_k^w(t)}{D^w}\right). \tag{6}$$

A. Joint Energy Cost and QoE Optimization

To obtain the joint optimization of energy cost and QoE, let us first formulate the optimization problem for minimizing energy cost. We can formulate the minimization of energy cost as an optimization with uncertainty $\lambda(t)$. In other words, our object is to minimize the expected cost over all clusters, i.e., $\mathbb{E}[\cdot]$ over random variables $\lambda(t)$. To minimize the cost at time slot t, the cluster power consumption cost $f(\cdot)$ is defined for each k and t as follows:

$$f\left(c_k(t), \lambda_k(t-1), \lambda_k(t)\right) = c_k(t) p_k(t). \tag{7}$$

The decision variable $c_k(t)$ is the amount of power bought from an electrical grid. So we can formulate the cost minimization problem as

$$\underset{c(t)}{\operatorname{argmin}} \sum_{k\in\mathcal{N}} \mathbb{E}[f\left(c_k(t), \lambda_k(t-1), \lambda_k(t)\right)] \tag{8}$$

where T is the total number of time slots and

$$\sum_{k\in\mathcal{N}} \zeta_k(t) \le D \quad \forall t \in \mathcal{T},\ \forall \zeta_k(t) \in Z(t) \tag{9}$$

and $\mathcal{T} = \{1, 2, \ldots, T\}$.

Equation (9) ensures that the available servers are sufficient to handle all user's demand. Equations (8) and (9) provide optimal solutions for the decision variable $c(t)$. In order to solve the uncertainty model defined in (8) and (9) in a global manner, we can use the *the deterministic* equivalent LP model [34] given as

$$\underset{c(t)}{\operatorname{argmin}} \sum_{k\in\mathcal{N}} p_k(t)c_k(t) \tag{10}$$

subject to

$$\sum_{k\in\mathcal{N}} (\zeta_k^v + \zeta_k^w) \le D^v + D^w \tag{11}$$

$$0 \le \zeta_k^v \quad 0 \le \zeta_k^w \quad 0 \le c_k(t) \quad \forall t \in \mathcal{T},\ \forall k \in \mathcal{N}. \tag{12}$$

Using LP as an optimization technique, we can obtain the optimal results in which the cost function is a linear function subject to linear constraints. LP can practically implement with low computational complexity.

If we incorporate (1) and (2) into (10)–(12), we can define joint optimization of energy cost and QoE[3] at time slot t as

$$\Phi(\bar{\zeta}) = \arg\min_{\zeta(t)} \sum_{k=1}^{N} \Big(-\alpha_1 \mathcal{Q}_k(t) + \alpha_2 p_k(t)(e^v \zeta_k^v(t) + e^w \zeta_k^w(t)) \Big) \tag{13}$$

subject to

$$\sum_{k=1}^{N} \zeta_k^v(t) = D^v(t) \qquad \sum_{k=1}^{N} \zeta_k^w(t) = D^w(t)$$

$$0 \le \zeta_k^v \le D_k^v \quad 0 \le \zeta_k^w \le D_k^w \qquad \forall\, k, t \tag{14}$$

where e^v and e^w are some positive constants and $D^v(t)$ and $D^w(t)$ are the VoD and web current user demands in CDN at time slot t, respectively. D_k^v and D_k^w are the maximum serviceable VoD and web connections for cluster k, respectively.

We use LP to solve the optimization problem in (13).

Remark 1: Let $\bar{\zeta}^*$ denote the solution of the optimization problem (13), then there might be noninteger optimal values for ζ^v and ζ^w. In this case, we can interpret optimal values as partially serving each VoD/web application by multiple clusters/servers. A practical case for this scenario is to distribute the content of a website or video file in multiple/geographically distinct server clusters.

[3]In other words, (13) and (14) can achieve optimal request routing for each time slot.

Remark 2: The objective of (13) and (14) is to redirect user requests according to the global server load balancing strategy in [6]. We currently assume that geographical user location is not taken into account in redirection decisions. In fact, the optimization problem in (13) and (14) balances the user requests between all server clusters based on an objective function which is composed of total CDN energy consumption and sum of the users' QoE.

Remark 3: The proposed optimal request redirection algorithm (13) has inherent scalability and can be easily adopted for a CDN architecture with multiple levels of hierarchy. In the case of a hierarchical server cluster, the optimal $\zeta_k^{v^*}$ and $\zeta_k^{w^*}$ can be distributed between cluster servers using a load-balancing agent.

Theorem 1: If we consider a combination of web/VoD traffic which is served by a CDN, then multiple optimal request routing solutions may exist from (13). If we consider distinct web or VoD application which is served by CDN, a unique and optimal request redirection solution can be found from (13).

Proof: We consider two distinct scenarios: 1) mixed traffic types (VoD and web) and 2) single web/VoD traffic.

Scenario 1: Let $\tilde{H}_{2N\times 2N}$ denote Hessian matrix associated with the objective function $\Phi(\cdot)$ in (13) given by

$$\tilde{H} = \begin{bmatrix} \frac{\partial^2\Phi}{\partial_{\zeta_1^{v^2}}} & \cdots & \frac{\partial^2\Phi}{\partial_{\zeta_1^v}\partial_{\zeta_N^v}} & \frac{\partial^2\Phi}{\partial_{\zeta_1^v}\partial_{\zeta_1^w}} & \cdots & \frac{\partial^2\Phi}{\partial_{\zeta_1^v}\partial_{\zeta_N^w}} \\ \vdots & & \vdots & \vdots & & \vdots \\ \frac{\partial^2\Phi}{\partial_{\zeta_N^v}\partial_{\zeta_1^v}} & \cdots & \frac{\partial^2\Phi}{\partial_{\zeta_N^{v^2}}} & \frac{\partial^2\Phi}{\partial_{\zeta_N^v}\partial_{\zeta_1^w}} & \cdots & \frac{\partial^2\Phi}{\partial_{\zeta_N^v}\partial_{\zeta_N^w}} \\ \frac{\partial^2\Phi}{\partial_{\zeta_1^w}\partial_{\zeta_1^v}} & \cdots & \frac{\partial^2\Phi}{\partial_{\zeta_1^w}\partial_{\zeta_N^v}} & \frac{\partial^2\Phi}{\partial_{\zeta_1^{w^2}}} & \cdots & \frac{\partial^2\Phi}{\partial_{\zeta_1^w}\partial_{\zeta_N^w}} \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2\Phi}{\partial_{\zeta_N^w}\partial_{\zeta_1^v}} & \cdots & \frac{\partial^2\Phi}{\partial_{\zeta_N^w}\partial_{\zeta_N^v}} & \frac{\partial^2\Phi}{\partial_{\zeta_N^w}\partial_{\zeta_1^w}} & \cdots & \frac{\partial^2\Phi}{\partial_{\zeta_N^{w^2}}} \end{bmatrix}.$$

From (3), we can simplify the matrix $\tilde{H}$ as follows:

$$\tilde{H} = \begin{bmatrix} \tilde{H}_1 & \tilde{H}_2 \\ \tilde{H}_3 & \tilde{H}_4 \end{bmatrix}$$

where

$$\tilde{H}_1 = \operatorname{diag}\left[-\alpha_1 \nu_k^2 \eta \exp\{-\nu_k \zeta_k^v(t) - \omega_k \zeta_k^w(t)\}\right]_{N\times N}$$

$$\tilde{H}_2 = \tilde{H}_3 = \operatorname{diag}\left[-\alpha_1 \nu_k \omega_k \eta \exp\{-\nu_k \zeta_k^v(t) - \omega_k \zeta_k^w(t)\}\right]_{N\times N}$$

and

$$\tilde{H}_4 = \operatorname{diag}\left[-\alpha_1 \omega_k^2 \eta \exp\{-\gamma_k \zeta_k^v(t) - \omega_k \zeta_k^w(t)\}\right]_{N\times N}.$$

From identity matrix property, we have

$$\det(\tilde{H}) = \det(\tilde{H}_1)\det(\tilde{H}_4 - \tilde{H}_3\tilde{H}_1^{-1}\tilde{H}_2). \tag{15}$$

It is very straightforward to show, according to Scenario 1, that the second term in (15) is zero, thus we have $\det(\tilde{H}) = 0$, and the Hessian $\tilde{H}$ is not necessarily convex/concave. As a result, the optimization problem in (13) may have multiple optimal points.

Scenario 2: For the case of single VoD traffic, we have $\tilde{H} = \tilde{H}_1$, therefore, we can rewrite (15) as

$$\det(\tilde{H}) = (-\alpha_1\eta)^N \left(\prod_{k=1}^{N} \nu_k^2 \exp\{-\nu_k \zeta_k^v - \omega_k \zeta_k^w\} \right)$$

where Hessian matrix $\tilde{H}$ is symmetric and concave. The first derivative with respect to all ζ_k is positive (strictly increasing) and the optimal solution satisfies the following equations:

$$\alpha_1 \eta \nu_k \exp\{-\nu_k \zeta_k^{v^*}(t)\} + \alpha_2 e^v p_k(t) - L^* = 0 \tag{16}$$

and

$$\sum_{k=1}^{N} \zeta_k^{v^*}(t) = D^v(t) \tag{17}$$

where L is the Lagrange multiplier. If we assume that $\nu_k = \nu$ for all k, we will reach the trivial solution $\zeta_k^{v^*}(t) = D^v(t)/N$, $\forall k$.

To prove that optimal value $\bar{\zeta}^*$ is a unique value, we assume that there are two differently optimal values $\bar{\zeta}_1^*$ and $\bar{\zeta}_2^*$ corresponding to different L_1^* and L_2^*. Using the optimality conditions in (16) and letting $L^* - \alpha_2 p_k(t) e^v > 0$ we can obtain

$$\sum_{k=1}^{N} \ln \left(\frac{L_1^* - \alpha_2 p_k(t) e^v}{L_2^* - \alpha_2 p_k(t) e^v} \right)^{\frac{1}{\nu_k}} = 0. \tag{18}$$

As $\alpha_2 p_k(t) e^v$ term is constant, (18) is valid only for $L_1^* = L_2^*$ which implies that the optimal solution $\bar{\zeta}^*$ must be unique for single VoD traffic. Similarly, we can prove the optimality conditions for single web traffic. ■

Next, we discuss the DE algorithm to solve the joint optimization problem in which each application/user is served either from a distinct edge server or multiple edge servers.

B. DE Approach to Energy-Efficient Content Distribution

The DE algorithm can be adopted to solve the nonlinear/nonconvex optimization problem. DE iteratively optimizes a problem by enhancing the number of candidate solutions with regard to a predefined quality metric [35]. We define parameters and variables for DE algorithm as shown in Table II. In this case, the optimization in (13) executes Algorithm 1 including the four stages of initialization, mutation, recombination, and selection to obtain the optimal value $\bar{\zeta}^*$. Furthermore, since the problem is constrained, we require to modify the last stage according to [36]. At the first stage, the values of D^v or D^w are updated according to the present condition on arrival of each new VoD/web request, then, we run the DE algorithm. For the next stage, we need to assign the service to each eligible server. This stage also includes the computation of the optimal number of assignable VoD/web connection to each server on every new arrival. Then, the new request is routed to the server based on the maximum available service capacity[4] which can be used for request redirection.

[4]This capacity is the difference between the optimal serving capacity determined by the DE algorithm and the number of present VoD/web sessions.

TABLE II
NOMENCLATURE

Parameters	Description
$\mathcal{A}$	The length of parameter vector ψ
$\mathcal{B}$	Total population number, typically $\mathcal{B} \geq 4$
G	Generation number
ψ	Parameter vector in DE algorithm
α_1, α_2	Positive cost function weighting parameters
$\mathcal{L}$	Constant from $[0, 2]$ for weighting vectors
Υ	Donor vector for mutation
Ξ	Trial vector for recombination
ρ	Number of CDN providers/groups
$\Phi(\cdot)$	CDN's Joint energy/quality cost function
$f(\cdot)$	Cluster power consumption cost
$GUF(\cdot)$	Group utility function
Variables	**Description**
$c_k(t)$	Dynamic power consumption for cluster k at slot t
$p_k(t)$	Power consumption price/cost for cluster k at time t
$\ell(t)$	Normalized per unit connection monetary gain

Algorithm 1: Executed for Each Application/User.

Initialization:
1: Randomly select the initial parameter values uniformly on interval $[0, \zeta^{\max}]$

Mutation:
1: Randomly select three arbitrary and distinct vectors $\psi_{i+r,G}$, $\psi_{i+r\prime,G}$, and $\psi_{i+r\prime\prime,G}$
2: Compute weighted donor vector as $\Upsilon_{i,G+1} = \psi_{i+r,G} + \mathcal{L}\left(\psi_{i+r\prime,G} - \psi_{i+r\prime\prime,G}\right)$

Recombination:
1: Obtain $\Xi_{i,G+1}$ from elements of target vector $\psi_{i,G}$ and donor vector $\Upsilon_{i,G+1}$

Selection:
1: Compare $\psi_{i,G}$ with $\Xi_{i,G+1}$
2: Consider target vector $\psi_{i,G}$ for the next generation based on [36]
3: Repeat tree stages mutation, recombination, and selection till meeting stopping criteria

C. Game-Theoretic Method to Energy-Efficient Content Distribution

A noncooperative game model focuses on the case that all players make decisions independently. Each decision maker tries to optimize its pay-off unilaterally. After converging to NE, the optimal objective of all players can be satisfied [37].

Consider a scenario where multiple ISP-operated CDN providers (CDNPs) interact with each other to enhance their footprint for delivering the content, as illustrated in Fig. 2. These ISP-operated CDN providers may share infrastructure with each other. In other words, they may exchange user connection requests between one another for load balancing and other efficiency-related purposes. In order to gain a profit, a CDN provider is required to interact with others using noncooperative game theory mechanism, since there does not exist any centralized CDN control authority.

Let us assume that the CDNs can form multiple groups. Each group can practically be associated with a CDN service provider in a country where it can exchange some resources, e.g., a percentage of web/VoD connections, from other groups. We assume that the total number of groups is ρ. In this case, if a typical group serves other groups, it can gain a monetary benefit from that

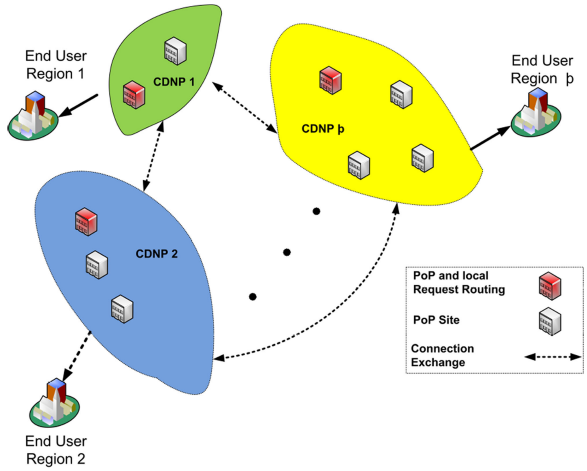

Fig. 2. Block diagram to illustrate game-theoretic approach for energy-efficient content distribution.

group; instead, its QoE for existing users deteriorates while its power consumption cost increases accordingly. Hence, we can form a *noncooperative game framework* for problem formulation. To do so, we first need to define a group utility function (GUF) as follows [38]:

$$\begin{aligned}\mathrm{GUF}(\Delta_y, \Delta_{-y}) \triangleq \sum_{\substack{x=1\\x\neq y}}^{\rho} \left(\sigma_y^v(t)\Delta\zeta_{x,y}^v(t) + \sigma_y^w(t)\Delta\zeta_{x,y}^w(t)\right) \\ - \sigma_y^v(t)\Delta\zeta_y^v(t) - \sigma_y^w(t)\Delta\zeta_y^w(t) - \delta_y\beta_y^+(t) \\ \forall y = 1, 2, \ldots, \rho \qquad (19)\end{aligned}$$

where $\sigma_y^v(t) = \ell^v(t) - \hat{p}_y^v(t) > 0$ and $\sigma_y^w(t) = \ell^w(t) - \hat{p}_y^w(t) > 0$, and they are the difference between normalized (per unit connection) monetary gain $\ell(t)$ and power consumption cost $\hat{p}_y(t)$ associated with VoD and web connection, respectively, borrowing from any other group at time slot t. δ_y is some positive constant. [5] $\Delta\zeta_{x,y}^v(t)$ and $\Delta\zeta_{x,y}^w(t)$ are the percentage of VoD and web connections served for group x by group y in time slot t, respectively. $\Delta\zeta_y^v(t)$ and $\Delta\zeta_y^w(t)$ are the total percentage of VoD and web connections associated with group y and served by other groups, i.e., $\Delta\zeta_y^v(t) = \sum_{\substack{x=1\\x\neq y}} \Delta\zeta_{y,x}^v(t)$ and $\Delta\zeta_y^w(t) = \sum_{\substack{x=1\\x\neq y}} \Delta\zeta_{y,x}^w(t)$ where

$$\begin{aligned}&\beta_y^+(t) \\ &\triangleq \alpha_1\left(\eta\exp\left\{-\nu_y\left(\zeta_y^v(t) + \sum_{\substack{x=1\\x\neq y}}\Delta\zeta_{x,y}^v(t) - \Delta\zeta_y^v(t)\right)\right.\right. \\ &\qquad \left.\left. - \omega_y\left(\zeta_y^v(t) + \sum_{\substack{x=1\\x\neq y}}\Delta\zeta_{x,y}^w(t) - \Delta\zeta_y^w(t)\right)\right\}\right) \\ &\qquad - \alpha_2 p^y(t)\left(e^v\zeta_y^v(t) + e^w\zeta_y^w(t)\right). \qquad (20)\end{aligned}$$

[5] For simplicity of notations, we assume that each group comprised of a single PoP/cluster for the beginning.

1) NE and Dominant Strategies: Let $\Delta_{-y} \triangleq \{\Delta_1, \Delta_2, \ldots, \Delta_{y-1}, \Delta_{y+1}, \ldots, \Delta_\rho\}$ denote the set of strategies adopted by all groups except y where $\Delta = \Delta_{-y} \bigcup \{\Delta_y\}$. The following strategies can be obtained.

1) The best strategy $\Lambda_y(\Delta)$ adopted by the group y is to select its connection share Δ_y such that

$$\Lambda_y(\Delta) = \arg\max_{\Delta_y} \mathrm{GUF}(\Delta_y, \Delta_{-y}) \qquad \forall y. \qquad (21)$$

2) The best strategy set of all the users, i.e., $\Delta^* = \{\Delta_1^*, \ldots, \Delta_N^*\}$ constructs the NE of the joint energy cost and QoE control game if and only if we have [39]

$$\Delta_y^* = \max(0, \Lambda_y(\Delta^*)) \qquad \forall y. \qquad (22)$$

Equation (22) states, for equilibrium strategies [39], that any user cannot unilaterally change its strategy (connection share) and improve its utility (pay-off) without compromising other users and decreasing the utility of at least one of them.

Remark 4: It must be mentioned that for simplicity of mathematical relations, we assume that all of the clusters in each group are homogeneous and have the same request handling capacity. Moreover, it is assumed that group manager (denoted by the PoP and local request router in Fig. 2), distributes the offered request load uniformly between group clusters.

Theorem 2: A NE exists for the joint energy cost and QoE control game when there is single cluster or multiple clusters in each group.

Proof: We first assume that each group is comprised of a single cluster. Let $G = \{I, \{\chi_y\}, \{\mathrm{GUF}_y(\cdot)\}\}$, $y \in I$ denote the game where $I = \{1, 2, \ldots, \rho\}$ is the index set for the groups, $\chi_y = \{0, \Delta_1^v, \Delta_2^v, \ldots, \Delta_F^v\} \bigcup \{0, \Delta_1^w, \Delta_2^w, \ldots, \Delta_F^w\}, 0 \leq \Delta_1^v \leq \Delta_2^v \leq \cdots \leq \Delta_F^v \leq D^v, 0 \leq \Delta_1^w \leq \Delta_2^w \leq \cdots \leq \Delta_F^w \leq D^w$ and is the strategy space and $\mathrm{GUF}_y(\cdot)$ is the utility function of group y. F is the index of strategies set. Each group determines the required connection share size such that $\Delta_y \in \chi_y$. Let the share vector $\Delta \in \chi$ denote the game outcome in terms of the connection share size required by all the groups where χ is the set of all connection share vectors. The strategy space of all the groups except group y is denoted by χ_{-y}.

According to [39], an NE exists for game G if the following two conditions are met: 1) χ_y is a nonempty, convex, and compact subset of some Euclidean space $\mathbb{R}^\rho$ and 2) $U_y(\Delta_y, \Delta_{-y})$ is continuous in Δ_y and quasi-concave in Δ_y. Each group has a strategy for the amount of required connection defined by a minimum value 0, and a maximum value Δ_F, and all the other values in between. We also assume that $\Delta_F \geq 0$, thus χ_k is clearly a closed and convex subset of $\mathbb{R}^\rho$, and the first condition is satisfied. It remains to show that the utility function $\mathrm{GUF}_y(\cdot)$ is quasi-concave in Δ_y for all y in the joint quality/energy control game.

In order to show the quasi-concavity property of the function $\mathrm{GUF}_y(\cdot)$, it is sufficient to show that its second derivative with respect to all values of $\Delta\zeta_{x,y}(t)$ is not positive for all values of m. From (19)–(20), it is very straightforward to verify

$$\frac{\partial^2 \mathrm{GUF}_y}{(\partial\Delta\zeta_{m,y}^v)^2} < 0 \qquad \frac{\partial^2 \mathrm{GUF}_y}{(\partial\Delta\zeta_{m,y}^w)^2} < 0 \qquad \forall m, y.$$

Hence, $\mathrm{GUF}_y(\cdot)$ is a continuous and strictly concave function of Δ_y for all y. A strictly concave function must also be a quasi-concave one, so that the second condition is also satisfied and an NE does exist in the joint energy cost and QoE control game.

If a group consists of multiple clusters, we can assume that connection share is divided uniformly between clusters by coalition manager. Therefore, the new group y utility can be written based on (19) as

$$\mathrm{GUF}_y(\Delta_y, \Delta_{-y}) \triangleq \sum_{\substack{x=1 \\ x\neq y}}^{\rho} \left(\sigma_y^v(t)\Delta\zeta_{x,y}^v(t) + \sigma_y^w(t)\Delta\zeta_{x,y}^w(t)\right) - \sigma_y^v(t)\Delta\zeta_y^v(t) - \sigma_y^w(t)\Delta\zeta_y^w(t) - \sum_{h=1}^{n_y} \delta_h \beta_h^-(t), \quad \forall y = 1, 2, \ldots, \rho \tag{23}$$

where n_y is the number of PoPs/clusters in group y. Also, we have for each h

$$\beta_h^-(t) \triangleq \alpha_1 \left(\eta \exp\left\{ -\nu_h\left(\zeta_h^v(t) + \frac{1}{n_y}\sum_{\substack{m=1 \\ m\neq h}} \Delta\zeta_{m,h}^v(t) - \Delta\zeta_h^v(t)\right) - \omega_h\left(\zeta_h^v(t) + \frac{1}{n_y}\sum_{\substack{m=1 \\ m\neq h}} \Delta\zeta_{m,h}^w(t) - \Delta\zeta_h^w(t)\right)\right\} + s\right) - \alpha_2 p_h(t)\left(e^v\zeta_h^v(t) + e^w\zeta_h^w(t)\right). \tag{24}$$

We can similarly use the above steps to prove the existence of NE for the multiple clusters case. ■

Remark 5: For reaching the NE point in a distributed manner, each player (CDN provider in this specific case) must deploy the DE algorithm presented in Algorithm I individually based on the information available from previous time slot $t-1$ to solve the nonlinear optimization in (21). The price of anarchy (PoA) for the game can be defined as follows [38]:

$$\mathrm{PoA}(\Gamma) = \frac{\sum_{y=1}^{\rho} \mathrm{GUF}(\Gamma_y, \Gamma_{-y})}{\sum_{y=1}^{\rho} \mathrm{GUF}(\Delta_y^*, \Delta_{-y}^*)} \tag{25}$$

where Δ^* is the optimal strategy set in a centralized solution and $\Gamma \subset \Delta$ is a nonoptimal subset of the strategy set.

V. Empirical Results and Discussion

In this section, numerical results are presented to demonstrate the performance of the proposed real-time energy management algorithm, joint energy cost and QoE optimization, and noncooperative game approach to energy-efficient content distribution. For the purpose of this study, we use the actual hourly real-time prices during 24 h as is shown in Fig. 3. Without loss of generality, we assume rectifiers with efficiency 95%. For simplicity, we also assume that there is no selling energy to the grid. The price of electricity is updated every two hours while the real-time energy management executes Algorithm 1 every one hour. Further, it is assumed that the video packet generation process follows a heavy-tail distribution such as Markov modulated Poisson process (MMPP) [41], [42]. MMPP is a doubly stochastic Poisson process whose average number of events in an interval, i.e., event rate varies according to a Markov process.

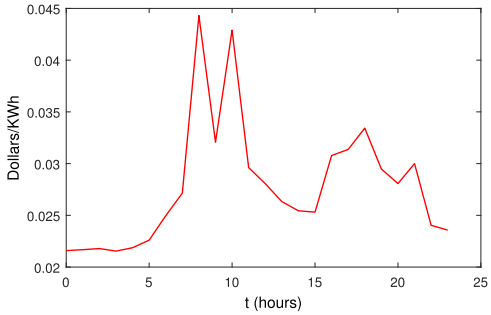

Fig. 3. Actual hourly real-time prices for 24 h daily measured in 2/1/2013 from the Illinois power company [40], [43].

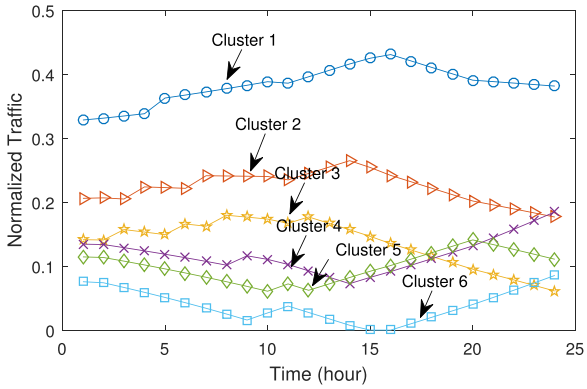

Fig. 4. CDN traffic distribution over various clusters.

We further assume that video sources use adaptive bit rate with two distinct high/low bit rate levels. Video encoders assume to equally likely switch between the two states. We choose $\eta = 3.8$, $s = 1.2$, $\nu = 0.04$, and $\omega = 0.004$ for each cluster based on the nonlinear regression procedure in [32]. The network is assumed to be stationary during the simulation period and it is assumed that the number of real-time/nonreal-time requests and real-time prices are fixed during each time slot. We have selected[6] $\alpha_1 = 1$, $\alpha_2 = 100$, $\alpha^v \approx 200\frac{\mathrm{mW}}{\mathrm{conection}}$, and $\alpha^w \approx 1\frac{\mathrm{mW}}{\mathrm{conection}}$. The number of VoD and web connections are assumed to vary approximately between mean value of 100 and 1000 for each cluster, respectively. We set $N = 6$ and consider initial traffic-load based on average event rate extracted from MMPP according to Fig. 4 where the normalized traffic versus different times is shown. We have used the different prices from various dates in [43] for all the six clusters. We have assumed practical capacity 5 KWh for the batteries according to [44] and HIP-200NHE1 Sanyo solar panel modules for this section. We also assume that the clusters are heterogeneous in terms of the number of solar panels, i.e., the clusters have different number of solar panels which are allocated based on their traffic. The joint proposed algorithm is evaluated in the following using three different cases: 1) LP-based optimization; 2) differential evolution method; and 3) noncooperative game approach.

Part 1) We present the number of allocated solar panels for the variation in traffic within different clusters which are illustrated

[6]In order to obtain the realistic power-related parameters, the energy consumption of the server is collected with ipmitool [48] directly from the power supply units (PSU) of the server.

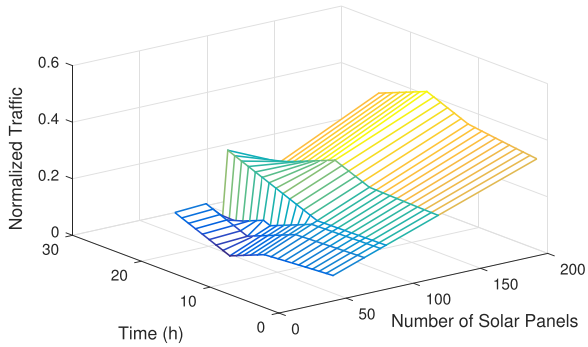

Fig. 5. Normalized traffic versus time and the number of solar panels within various clusters.

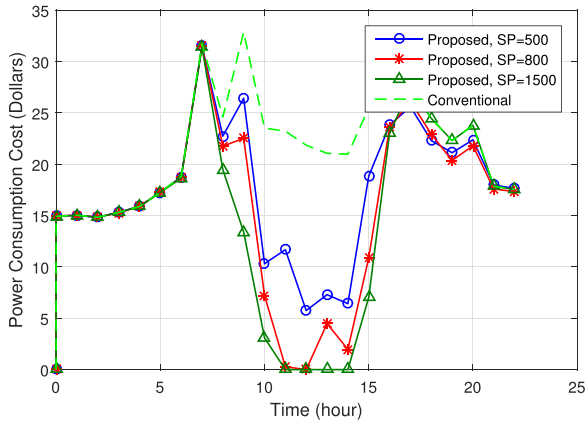

Fig. 6. Hourly operational cost comparison for different total number of solar panels.

in Fig. 5. In this case, the total number of solar panels for all clusters are $SP = 800$. As seen when the traffic within a cluster is increased, the number of solar panels for the corresponding cluster is also increased in order to use more green energy as well as to reduce the energy cost. Moreover, Fig. 6 presents the hourly operational cost for the proposed joint energy cost and QoE optimization in comparison with the system without real time energy management. In fact, we compare a regular CDN, namely a *conventional CDN*, which only uses the grid for supplying power, and *a proposed green CDN*, which balances demand and supply using real-time energy management, while it uses both the grid and the solar panels. We employ LP for solving the joint problem while considering different total number of solar panels in order to see the impact of this parameter on the energy cost reduction. The proposed joint optimization has a large cost saving over various time slots for the different total number of solar panels.

Part 2) In order to solve the joint optimization problem, we use DE algorithm as a nonlinear method and practical approach to obtain global optimization. In this case, Fig. 7 presents the hourly operational cost for the proposed joint energy cost and QoE optimization using DE and LP compared to the conventional system while LP and DE algorithms use real-time energy pricing. The total number of solar panels for all clusters is $SP = 1500$. As seen, DE algorithm performs much better than the LP in terms of operational cost reduction over various time slots. This performance improvement is summarized in Table III. We

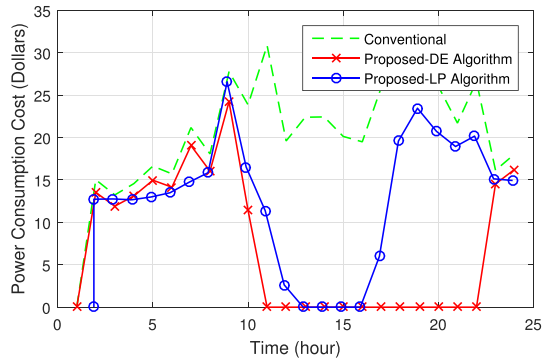

Fig. 7. Hourly operational cost for the proposed DE method and LP in comparison with the conventional system.

TABLE III
OPERATIONAL COST AND COMPUTATIONAL COMPLEXITY COMPARISON OF LP AND DE ALGORITHMS FOR SOLVING JOINT ENERGY COST AND QOE OPTIMIZATION

	Cost Reduction Ratio %		Execution Time Reduction Ratio %
SP	LP	DE	
500	22	57	59
800	30	60	64
1500	41	66	62

Mean Execution time (Second)

SP	LP	DE
500	1.19	2.93
800	1.48	4.17
1500	3.09	8.16

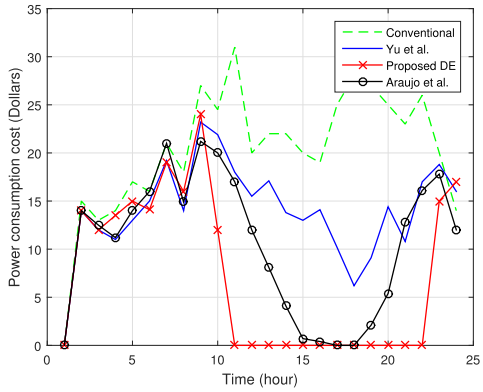

Fig. 8. Hourly operational cost comparison while the total number of solar panels is $SP = 1500$.

define the operational cost reduction ratio for the proposed joint optimization (using LP and DE) as $1 - \frac{\text{Cost}_{\text{Proposed}}}{\text{Cost}_{\text{Conventional}}}$. The DE algorithm can significantly save operational cost between 57% and 66% over the conventional system for different values of SP. This saving decreases to almost between 22% and 41% if we consider LP over the conventional system. We can also obtain that the DE algorithm is up to 34% cost saving compared with the linear LP method.

In Fig. 8, the proposed DE method is compared with that of conventional CDN without energy management, Araujo *et al.* [11] and Yu *et al.* [13]. The work in [11] study the impact of using in-network caches and CDN cooperation on the energy consumption of request routing, but it does not consider real-time energy management using renewable energy resources. The main

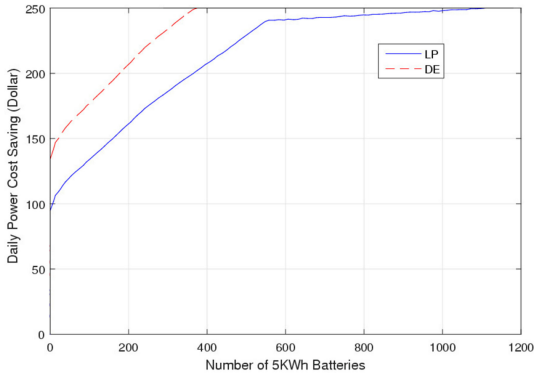

Fig. 9. Cost saving versus battery capacity for both LP and DE algorithms.

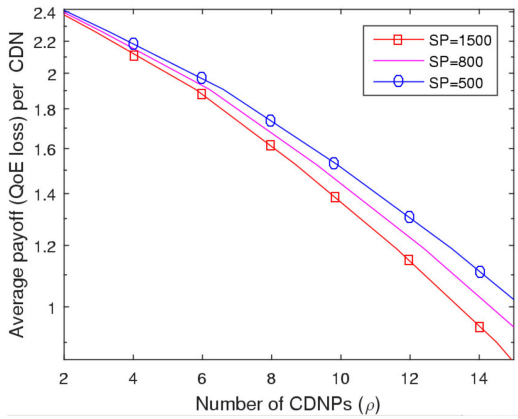

Fig. 10. Average QoE loss per CDN resulting from applying a noncooperative game for traffic exchange by CDN providers.

objective in [13] is geographical load balancing for distributed Internet data centers using a price-sensitivity aware geographical load balancing scheme, but, it is not tailored to fit well in a smart grid-enabled CDN environment which benefits from renewable and green energy. As it can be verified, the proposed DE method outperforms other related methods in energy cost reduction.

Further, we measure the computation time of the DE algorithm and LP approximation for different values of SP, as shown in Table III. We use the ratio of the LP optimization execution time for the proposed technique to that of the DE algorithm in order to compare the computational complexity, i.e., $1 - \frac{\text{Time}_{\text{LP}}}{\text{Time}_{\text{DE}}}$. The optimization time for each algorithm is averaged over 1000 running joint optimization algorithm. Simulation was done using MATLAB on a Quad-Core Intel processor with a 3-GHz clock speed. Table I shows the ratios versus the total number of solar panels SP. This shows that the computation time of the LP approach can be as small as 59% of that of the DE algorithm. Thus, our results can provide an operational cost saving versus complexity tradeoff for various total number of solar panels SP. However, it seems that the value of mean execution time for DE algorithm is not much longer than the value of mean execution time for LP algorithm. Thus, DE has superiority over LP in terms of operational cost saving with reasonable computational complexity.

Fig. 9 presents the cost saving versus storage capacity while using both LP and DE algorithms for the joint optimization problem. The total number of batteries are shared based on mean normalized cluster traffic (see Fig. 4) in 24 h between different clusters. As seen, DE performs better than LP in terms of cost saving when both algorithms use the small number of storage batteries. Further, both algorithms have a limited cost saving for a small number of batteries. As we increase the storage capacity, the benefits of cost saving for both algorithms become constant. It is obvious that if we have the capability of selling green energy to the electrical grid, then certainly more benefits can be obtained and we can buy sufficient electricity from the grid for a long period.

Part 3) Using group formation for CDN providers for traffic/connections exchange explained in Section III-D, combined with a DE optimization solution, leads to promising results as illustrated in Fig. 10. The results show that the use of noncooperative games yields a performance advantage, in terms of QoE loss per CDN provider, which is increasing with the number of CDN providers ρ and reaching up to almost zero loss reduction at $\rho = 15$ relative to the scheme without traffic/connections exchange. If we increase the number of CDNPs, QoE level can also increase. In this case, the total CDN system can get more benefits from large elasticity and also, there exists more chance for an incoming request to find spare capacity in order to satisfy its demand in one of the CDN providers. The QoE was measured in peak signal to noise ratio (PSNR) in this experiment. The average initial PSNR was set to 40 dB.[7]

Part 4) In the sequel, we have considered CDNsim testbed which has been designed for large-scale CDN simulation. We have also used Oversim class in OMNET++ for application layer simulation. The selected network topology is consisted to be similar to Abilene with 11 clusters (PoPs) which are assumed to be located on different geographical locations [45]. Each cluster is assumed to have 20 Taurus servers equipped with two Intel Xeon E5-2630 CPUs with six cores each, 32GB memory, 598GB storage, and a 10-gigabit ethernet interface. Furthermore, we have added one request router/redirector server. We have used realistic 24-h SDSC traffic traces [46] for simulating the request traffic in clusters. The requests are assumed to be of streaming video type and similar to [27], and are distributed between clusters based on a PoP-specific scaling factor to the request volume in the trace. We have used video traces Verbose_ARDTalk (which use MPEG4-AVC/H.264 coding) for transmission over the CDN. The mean video streaming bit rate is selected to be 512 Kbps. For simulating realistic video popularity distribution, we have used Zipf distribution with parameter equal to 0.9 [45]. The number of video traces with different bit rates is 500 and these video realizations are distributed between edge servers based on Zipf distribution. The maximum allowable number of concurrent video user requests is limited to 25 000 for the entire CDN system (about 2300 for each cluster). As the mean request volumes differ in different time frames [46], we have selected a time frame of 1 h and denoted its mean request volume to the CDN system as client density in simulations. For real-time energy prices, we have used realistic price information based on the Ontario power company during 24 h [47]. It is assumed that the price profiles are similar among different CDN clusters but, as CDN clusters are assumed to be located in different time zones, the graph is circularly shifted in a time span of 3 h between different 11 PoPs. The total number of solar panels is set to 1500 and it is assumed that is distributed uniformly

[7]For mapping between the objective PSNR and subjective MOS, a sigmoidal curve has been used [31].

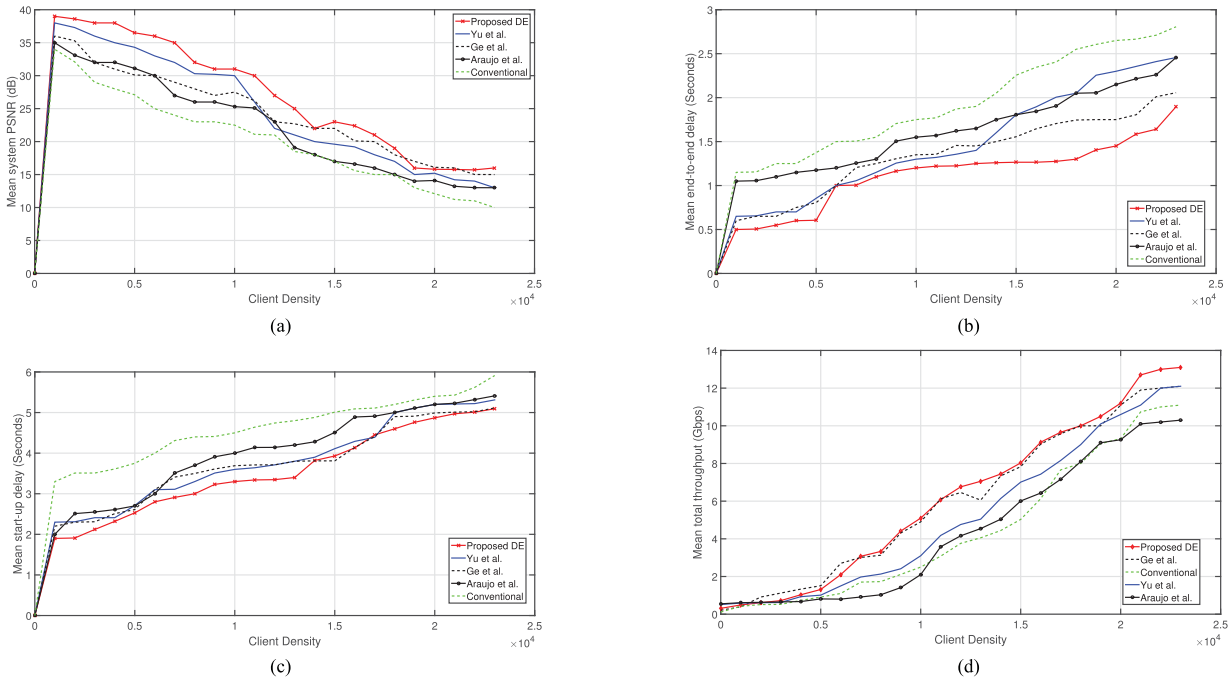

Fig. 11. QoE performance comparison versus client density while the total number of solar panels is SP = 1500. (a) Mean system PSNR. (b) Mean end-to-end delay of the CDN system. (c) Mean video start-up delay of the CDN system. (d) Mean total throughput of the CDN system.

between clusters. We have implemented the proposed evolutionary real-time energy management algorithm (DE method) in the request redirector functionality[8] and compared it with conventional CDN and the work of Ge *et al.* [27], Yu *et al.* [22], and Araujo *et al.* [11] in QoE performance metrics, which are considered to be mean system PSNR, mean end-to-end delay, mean total throughput, and mean video start-up delays of the CDN system. These average values are obtained by averaging between 10 000 independent runs. As can be verified from Fig. 11, the proposed DE method employing real-time energy management outperforms other methods in the selected QoE performance metrics. For example, based on Fig. 11(a), even in large client density, the proposed method can outperform by up to 5 dB with respect to a conventional CDN in mean system PSNR. On the other hand, as can be verified in Fig. 9(d), as the proposed method benefits from an intelligent QoE-based request redirection mechanism, almost all of the video users can receive video with mean bit rate of 512 Kbps and achieve maximum available quality.

VI. Conclusion

In this paper, we considered greening CDN networks within a smart grid context. This enabled us to provide real-time energy management for the CDN clusters. The joint energy cost and QoE problem was optimized within a CDN using LP and DE algorithms. This showed that using LP provides lower complexity than DE while significantly deteriorating energy cost savings. Both algorithms provide us with a complexity and cost saving tradeoff. A noncooperative game was proposed for CDN providers to exchange traffic/connections according to the availability of their green generations. Our results showed that noncooperative games could become a foundation for traffic exchange between CDN providers in future green CDN networks. Finally, we presented that the new proposed method provides significant energy cost reduction with reasonable complexity over conventional CDN without using real-time energy management and jointly optimized energy cost and QoE. The future work arising from our studies will be exploiting real-time energy management as well as jointly optimized energy cost and QoE in the context of cloud and mobile CDNs.

[8]As CDNsim has a DNS-based request redirection mechanism, we have used it for conventional CDN and modified the request redirection mechanism both at DNS redirector and client side for implementing the proposed DE mechanism.

References

[1] "Data center report to congress," U.S. Environmental Protection Agency, Washington, DC, USA, Tech. Rep. 2075-0401, 2007.
[2] V. Mathew, R. K. Sitaraman, and P. Shenoy, "Energy-aware load balancing in content delivery networks," in *Proc. IEEE Infocom*, Dec. 2012, pp. 954–962.
[3] C. Ge, Z. Sun, and N. Wang, "A survey of power-saving techniques on data centers and content delivery networks," *IEEE Commun. Surv. Tut.*, vol. 15, no. 3, pp. 1334–1354, Third Quarter 2013.
[4] T. A. Gulliver, J. M. Cioffi, and G. K. Karagiannidis, "Radio over fiber based networks for the smart grid," in *Proc. IEEE Globecom*, Dec. 2014, pp. 2605–2611.
[5] J. Song, F. Yang, Y. Zhou, S. Wan, and H. R. Wu, "QoE evaluation of multimedia services based on audiovisual quality and user interest," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 444–457, Mar. 2016.
[6] R. Buyya, M. Pathan, and A. Vakali, *Content Delivery Networks*. Berlin, Germany: Springer-Verlag, 2008.
[7] U. Mandal, P. Chowdhury, C. Lange, A. Gladisch, and B. Mukherjee, "Energy-efficient networking for content distribution over telecommunications network infrastructure," *Opt. Switching Netw.*, vol. 10, no. 4, pp. 393–405, Nov. 2013.
[8] J. Sung, M. Kim, K. Lim, and J. K. Rhee, "Efficient cache placement strategy in two-tier wireless content delivery network," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1163–1174, Jun. 2016.
[9] W. Zhang, Y. Wen, Z. Chen, and A. Khisti, "QoE-driven cache management for HTTP adaptive bit rate streaming over wireless networks," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1431–1445, Oct. 2013.

[10] Y. Jin, Y. Wen, and K. Guan, "Toward cost-efficient content placement in media cloud: modeling and analysis," *IEEE Trans. Multimedia*, vol. 18, no. 5, pp. 807–819, May 2016.
[11] J. Araujo, F. Giroire, J. Moulierac, Y. Liu, and R. Modrzejewski, "Energy efficient content distribution," *Comput. J.*, vol. 58, no. 12, pp. 192–207, 2015.
[12] J. Yang, S. Zhang, X. Wu, Y. Ran, and H. Xi, "Online learning-based server provisioning for electricity cost reduction in data center," *IEEE Trans. Control Syst. Technol.*, vol. 25, no. 3, pp. 1044–1051, May 2017.
[13] L. Yu, T. Jiang, and Y. Zou, "Price-sensitivity aware load balancing for geographically distributed internet data centers in smart grid environment," *IEEE Trans. Cloud Comput.*, vol. 6, no. 4, pp. 1125–1135, Oct.–Dec. 2018.
[14] J. Llorca *et al.*, "Dynamic in-network caching for energy efficient content delivery," in *Proc. IEEE Infocom*, 2013, pp. 245–249.
[15] L. Yu, T. Jiang, Y. Cao, and Q. Qi, "Joint workload and battery scheduling with heterogeneous service delay guarantees for data center energy cost minimization," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 7, pp. 1937–1947, Jul. 2015.
[16] S. Zhou, J. Gong, Z. Zhou, W. Chen, and Z. Niu, "Green delivery: Proactive content caching and push with energy-harvesting-based small cells," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 142–149, Apr. 2015.
[17] J. Gong, S. Zhou, Z. Zhou, and Z. Niu, "Proactive push with energy harvesting based small cells in heterogeneous networks," in *Proc. IEEE Infocom*, 2015, pp. 25–30.
[18] V. Mathew, K. Ramesh, P. Sitaraman, and J. Shenoy, "Energy-efficient content delivery networks using cluster shutdown," in *Proc. Int. Green Comput. Conf.*, 2013, pp. 1–10.
[19] M. Savi, G. Verticale, M. Tornatore, and A. Pattavina, "Energy-efficient VoD content delivery and replication in integrated metro/access networks," in *Proc. IEEE LATINCOM*, 2014, pp. 1–6.
[20] M. Forshaw and N. Thomas, "A novel approach to energy efficient content distribution with BitTorrent," in *Computer Performance Engineering* (Lecture Notes in Computer Science, vol. 7587), M. Tribastone and S. Gilmore, Eds., Berlin, Germany: Springer, 2012, pp. 188–196.
[21] Y. Liu, D. Niu, and B. Li, "Delay-optimized video traffic routing in software-defined interdatacenter networks," *IEEE Trans. Multimedia*, vol. 18, no. 5, pp. 865–878, May 2016.
[22] L. Yu, T. Giang, and Y. Zou, "Real-time energy management for cloud data centers in smart microgrids," *IEEE Access*, vol. 4, pp. 941–950, 2016.
[23] Y. Hu, D. Niu, and Z. Li, "A geometric approach to server selection for interactive video streaming," *IEEE Trans. Multimedia*, vol. 18, no. 5, pp. 840–851, May 2016.
[24] P. Goudarzi, "Stochastic total cost of ownership optimization for video streaming services," *Telematics Inform.*, vol. 31, no. 1, pp. 79–90, 2014.
[25] H. Hu, Y. Wen, T.-S. Chua, J. Huang, W. Zhu, and X. Li, "Joint content replication and request routing for social video distribution over cloud CDN: A community clustering method," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 7, pp. 1320–1333, Jul. 2016.
[26] Y. Wen, X. Zhu, J. J. P. C. Rodrigues, and C. W. Chen, "Cloud mobile media: Reflections and outlook," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 885–902, Jun. 2014.
[27] C. Ge, Z. Sun, N. Wang, K. Xu, and J. Wu, "Energy management in cross-domain content delivery networks: A theoretical perspective," *IEEE Trans. Netw. Service Manage.*, vol. 11 no. 3, pp. 264–277, Sep. 2014.
[28] M. A. Salahuddin, J. Sahoo, R. Glitho, H. Elbiaze, and W. Ajib, "A survey on content placement algorithms for cloud-based content delivery networks," *IEEE Access*, vol. 6, pp. 91–114, 2017.
[29] *Cisco Media Delivery Engine (MDE) 1100*, 2011. [Online]. Available: https://www.cisco.com/c/en/us/products/collateral/video/media-delivery-engine-1100/datasheet_c25-677670.html
[30] J. Broberg, R. Buyya, and Z. Tari, "MetaCDN: Harnessing 'storage clouds' for high performance content delivery," *J. Netw. Comput. Appl.*, vol. 32, no. 5, pp. 1012–1022, 2009.
[31] D. Liu *et al.*, "Mapping algorithm from PSNR (peak signal-to-noise ratio) to MOS (mean opinion score) in video system," *Chinese Patent Appl.* CN102630037A, 2012.
[32] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Netw.*, vol. 24, no. 2, pp. 36–41, Mar./Apr. 2010.
[33] M. Dayarathna, Y. Wen, and R. Fan, "Data center energy consumption modeling: A survey," *IEEE Commun. Surv. Tut.*, vol. 18, no. 1, pp 732–794, First Quarter 2016.
[34] J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming*. New York, NY, USA: Springer, 1997.
[35] K. Price, M. R. Storn, and A. J. Lampinen, *Differential Evolution: A Practical Approach to Global Optimization*. Berlin, Germany: Springer-Verlag, 2005.
[36] A. W. Mohamed and H. Z. Sabry, "Constrained optimization based on modified differential evolution algorithm," *Inf. Sci.*, vol. 194, pp. 171–208, 2012.
[37] P. D. Straffin, *Game Theory and Strategy*. Washington, DC, USA: Mathematical Association America, 1993.
[38] T. Basar and G. J. Olsder, *Dynamic Noncooperative Game Theory*. London, U.K.: Academic, 1982.
[39] C. U. Saraydar, N. B. Mandayam, and D. J. Goodman, "Efficient power control via pricing in wireless data networks," *IEEE Trans. Commun.*, vol. 50, no. 2, pp. 291–303, Feb. 2002.
[40] P. D. Diamantoulakis and G. K. Karagiannidis, "Smart hybrid power system for base transceiver stations with real-time energy management," in *Proc. IEEE Globecom*, Dec. 2013, pp. 2773–2778.
[41] S. Asmussen, *Applied Probability and Queues (Stochastic Modelling and Applied Probability)*. New York, NY, USA: Springer-Verlag, 2000.
[42] D. R. Cox, "Some statistical methods connected with series of events," *J. Roy. Statist. Soc., Ser. B*, vol. 17, no. 2, pp. 129–164, Jul. 1995.
[43] *Ameren: Real Time Prices*, Jan. 2013. [Online]. Available: https://www2.ameren.com/RetailEnergy/realtimeprices.aspx
[44] *4.8 kWh SMA Battery Backup System*. [Online]. Available: http://www.energymatters.com.au/48-kwh-sma-battery-backup-system-p-2637.html
[45] G. Rossini and D. Rossi, "Large scale simulation of CCN networks," *14mes Rencontres Francophones sur les Aspects Algorithmiques des Tlcommunications*, pp. 1–4, 2012.
[46] *Traces in the Internet Traffic Archive*. [Online]. Available: http://ita.ee.lbl.gov/html/traces.html
[47] *Power Data*. [Online] Available: http://www.ieso.ca/power-data
[48] *ipmitool(1)—Linux Man Page*. [Online]. Available: https://linux.die.net/man/1/ipmitool

Pejman Goudarzi received the B.Sc. degree in electronics from the Sharif University of Technology, Tehran, Iran, in 1995 and the M.Sc. and Ph.D. degrees in electrical engineering from the Isfahan University of Technology, Isfahan, Iran, in 1998 and 2004, respectively.

He is a Faculty Member, Senior Research Engineer, and Associate Professor with the Research Institute for ICT (ITRC), Tehran, Iran. His research interests include wireless video communication, distributed resource allocation algorithms, game theory, soft computing, data science, and quality control in data networks.

Abolfazl Ghassemi (SM'12) received the Ph.D. degree in electrical engineering from the University of Victoria, BC, Canada, in 2009.

His research interests include Internet of Things, big data analytics, green communications and computing, smart grids, radio over fiber based networks, and compressed sensing.

Mohammad R. Mirsarraf received the Ph.D. degree in telecommunications in 2001.

His main research interest is cloud computing, Internet of Things, real-time operating systems, fifth-generation wireless networks, and next-generation networks. He is also interested in using semiology and pragmatism theory on human–computer interaction design.

Rajkumar Buyya (F'15) is currently a Redmond Barry Distinguished Professor and the Director of the Cloud Computing and Distributed Systems (CLOUDS) Laboratory, The University of Melbourne, Victoria, Australia. He is also the Founding CEO of Manjrasoft Pty. Ltd. His research interests include cloud computing, fog computing, and parallel and distributed systems.