Energy Reports 9 (2023) 631–644

Contents lists available at ScienceDirect

Energy Reports

journal homepage: www.elsevier.com/locate/egyr

Research paper

Energy cost optimization of globally distributed Internet Data Centers by copula-based multidimensional correlation modeling

Mohammad Ali Lasemi [a], Shahin Alizadeh [b], Mohsen Assili [b], Zhenyu Yang [a,*], Payam Teimourzadeh Baboli [c], Ahmad Arabkoohsar [a], Amin Raeiszadeh [c], Michael Brand [c], Sebastian Lehnhoff [c]

[a] AAU ENERGY, Aalborg University, Esbjerg, Denmark
[b] Faculty of Electrical Engineering, Shahrood University of Technology, Shahrood, Iran
[c] Resilient Monitoring and Control R&D group, Energy Division, OFFIS – Institute for Information Technology, Oldenburg, Germany

ARTICLE INFO

Article history:
Received 31 May 2022
Received in revised form 11 November 2022
Accepted 7 December 2022
Available online xxxx

Keywords:
Electricity market
Energy management
Energy storage system
Correlation modeling
Internet data center
Quality of service
Workload distribution

ABSTRACT

The high operating costs of Internet Data Centers (IDC) are a major challenge for their owners worldwide. Therefore, more attention has recently been paid to the energy and cost management of IDCs. This paper investigates the optimal operational strategy for minimizing the electricity costs of a group of globally distributed IDCs in different locations under various day-ahead electricity markets, and each is equipped with a high-performance energy storage system. For this goal, optimal workload dispatching and optimal energy management of the storage units of all IDCs are simultaneously perused by the proposed problem. The system is modeled regarding power balancing constraints, battery costs, and quality of service (QoS). For more practical results, a penalty function is also considered when QoS constraints are not perfectly met, and the impact of the batteries' depth of discharge on the cost of energy storage is also modeled. Moreover, the cross-correlations between the traffic of IDCs are also considered by the multidimensional copula function. The proposed energy cost optimization is linearized for increasing the accuracy of convergence. The results show that not only the power consumption pattern of the IDCs is significantly improved, but also the cost of power consumption is reduced by 34%. The results also prove the positive effect of battery discharge on workload dispatch and represent a compromise between battery costs and electricity cost savings.

1. Introduction

1.1. Motivation

The aggregated data generated every second by millions of Internet users should be processed by different servers. These servers are kept in places called Internet Data Centers (IDC) distributed geographically in the different locations of the world (Chen et al., 2020). Due to the growing demand for cloud computing and Internet services, there is significant growth in the construction of new IDCs or the development of available IDCs. As a result, the electric power consumption associated with IDCs has increased (Adrah et al., 2020). For instance, in 2014, 1.8% of the total electricity consumption in the United States was associated with the data centers, with a growth rate of 4% per year (Shehabi et al., 2016). The total electricity cost of an IDC comprises 30%–50% of its operating costs (Zhou et al., 2019), motivating experts to develop methods and new solutions to reduce the energy consumption of the IDCs. One of the critical factors in the operation of IDCs is the uncertain behavior of end-users. In addition to this uncertain behavior, the workloads of the various IDCs are spatially and temporally correlated. Therefore, more accurate models of these temporal–spatial correlations are one of the paradigms in the operation of IDCs.

1.2. Literature review

The research efforts in this context could be divided into three main categories. The first category includes investigations to decrease the IDCs' power consumption, which depends on the internal systems. In this category of works, the focus is on novel facilities and optimization procedures related to the cooling systems of data centers (i.e., computer room air conditioning units and fans). For this, some metrics and models of an IDC are necessary considering both computational and physical characteristics, as well as, their interactions (Koronen et al., 2020). Liu et al. (2020) presented a comprehensive review of the effect

[*] Corresponding author.
E-mail address: yang@energy.aau.dk (Z. Yang).

https://doi.org/10.1016/j.egyr.2022.12.033

Nomenclature

Sets

i	Index of the front-end web portal server
j	Index of the area where IDCs are located in
M	Set of IDCs
S	Set of front-end web portal servers
t	Index of time

Parameters

α	Penalty rate (\$/request)
$\overline{L_j}$	Maximum capacity of the battery in IDC j (kWh)
$\overline{N_j^{ser}}$	Total number of servers in IDC j
$\overline{P_j^{ch}}$	Maximum charging rate of battery in IDC j (kW)
$\overline{P_j^{dis}}$	Maximum discharging rate of battery in IDC j (kW)
ρ	Correlation coefficient
θ	Charging efficiency rate of battery
$\underline{L_j}$	Emergency level of battery in IDC j (kWh)
C_j^b	Battery cost per kWh (\$/kWh)
D^{SLA}	Delay bound of SLA (s)
$F(x_n)$	nth marginal distribution functions
N^c	Number of charging and discharging cycles
$T_{i,j,t}^d$	Transmission delay from front-end web portal server i to IDC j at time t (S)
$w_{i,t}$	Requests rate received by ith front-end web portal server at time t (request/s)
λ^{bat}	Battery price (\$)
$\mu_{j,t}$	Service rate of the IDC's server at area j at time t (request/s)
$\tilde{P}_{j,t}^s$	Power consumption of server s in IDC j at time t (kW)
$\tilde{P}^{dep}$	Average depth of discharge
Δt^Q	Period of requests transmission from front-end web portal server to data centers (S)
Δt^{bat}	Period of charging or discharging mode of the battery (h)
$\lambda_{j,t}^e$	Electricity price in IDC j at time t (\$/MWh)
DoD	Depth of discharge

Variables

$\beta_{j,t}$	Auxiliary binary variable
$E_{i,j,t}^d$	Total delay (S)
$L_{j,t}$	Power level of the battery in IDC j at time t (kWh)
$n_{j,t}$	Number of active servers in IDC j at time t
$P_{j,t}^g$	Transferred power from the electricity grid to IDC j at time t (kW)
$P_{j,t}^{ch}$	Charging power of battery in IDC j at time t (kW)
$P_{j,t}^{dis}$	Discharging power of battery in IDC j at time t (kW)
$P_{j,t}^{DC}$	Power consumption of IDC j at time t (kW)
$Q_{j,t}^d$	Queening delay (S)
$\xi_{i,j,t}$	Requested rate from front-end web portal server i to IDC j at time t (request/s)

of different thermal energy storage technologies on the energy management of IDCs. A comprehensive review related to the assessment criteria of the thermal performance of the internal interactions of IDCs has been carried out by Gong et al. (2020). In Li and Li (2020), an energy recovery system has been suggested to cool an IDC using a water-side economizer. This system has been presented by a model-based methodology to optimize the ambient wet-bulb and the cooling tower approach temperature in different cooling modes. He et al. (2021) presented an integrated heat pipe cooling system considering the heat transfer and the energy consumption model. The proposed model has defined the relationship between the operating parameters and energy efficiency and has been solved by Genetic Algorithm (GA). Water-cooled multi-chiller cooling systems have been considered for the cooling of IDCs, and the effects of different configurations of these systems have been investigated on the reliability and availability of IDCs in Cheung and Wang (2019). Authors in Temiz and Dincer (2022) propose a heat recovery system to reduce the cooling cost and increase the energy efficiency of IDCs, by applying the IDCs' waste heat in low-temperature district heating networks. Marshall and Duquette (2022) have presented a heat recovery system including a heat pump/heat pipe integrated unit based on the three-fluid heat exchange.

The second category of works in this framework includes the optimal workload distribution for a set of data centers to reduce the energy cost and conserve the quality of service (QoS). Generally, IDCs are in different geographic locations and operate under different markets (Hintemann and Hinterholzer, 2019). Users' requests are first sent to the front-end web portal servers and thereby are dispatched between the data centers. Front-end web portal servers choose the data center with the lowest electricity price to transfer workload. According to the Service Level Agreement (SLA), each IDC should provide the required QoS to the end-users. If an IDC cannot satisfy the QoS, the Internet provider must pay the penalty (Kwon, 2020). The penalty rate is determined through the SLA, and the Internet Service Providers (ISP) should guarantee the QoS accepted by the end-users; otherwise, they would be penalized based on SLA regulation. Considering QoS constraints, the power consumption cost optimization problem has been presented to distributed data centers based on dynamic voltage and frequency scaling in Ref. Gu et al. (2014).

The model presented by Gu et al. (2014) provided a balance between active servers and the operating frequency of each server to receive workload. When the request demand rate is more than the service rate of IDCs, the request outage is defined. Authors in Jin et al. (2020) introduced a novel model for reducing the electricity cost and meeting the outage probability constraint via dynamically adjusting server capacity and performing demand shifting in different time scales. To manage the IDC's workload, in Cheng et al. (2021) a new operation scheduling is presented considering the virtual data center allocation concept based on the scalable constraint. The energy costs of IDCs in the long term have been minimized by considering the optimal workload dispatching, the uncertainty in the price of electricity, and the renewable energy generation in Ref. Peng et al. (2021). Sun et al. (2020) have given a novel workload transfer strategy for IDCs considering the local electricity markets. Applying shifting workload capability among distributed IDCs, the energy management of IDCs has been investigated, using demand response programs (Zhang et al., 2021b). Zhang and Zavala (2021) carried out temporally shiftable electricity demand in large-scale IDCs to reduce the power consumption considering price sensitivity and cooling efficiency. Authors in Zhang et al. (2021a) used a proactive demand response considering the impact of IDCs' load redistribution on the power network for appropriate pricing and power load balancing.

Cost reduction by using energy buffering forms the following category of research activities in this regard. Electricity prices are usually variable under different electricity markets (Lasemi et al., 2022). In day-ahead electricity markets, the power consumption of IDCs can be planned to reduce by using energy buffering. Considering a set of batteries with a high capacity for each data center, the batteries are charged while electricity price is low and discharged to supply the servers while electricity price becomes high. In Lyu et al. (2021), an online algorithm has been presented to reduce the cost of IDCs by energy management of storage systems through stochastic programming. This work disregards the impact of the depth of discharge in the life cycle of the battery. Sajid et al. (2019) proposed an integrated power management of IDCs and electric vehicles (EVs) for frequency regulation. Using renewable energy resources to supply the required energy of IDCs has also been taken into consideration via a concept of so-called green data centers (Hu et al., 2021). Wang et al. (2020) proposed an analytical model for improving energy management in large IDCs to facilitate wind power integration. Rahmani et al. (2020) investigated how to minimize energy cost and carbon emission in IDCs connected to a microgrid considering stochastic parameters such and microgrid design optimization.

1.3. Contributions

As discussed in the literature review section, most of the research works regarding the optimal operation of IDC have focused on only one of the above-discussed main categories, *i.e.*, optimization of the internal system, optimization of the workload distribution, and implementation of energy buffering. The proposed models of almost all these studies suffer from the complexity of the nonlinear optimization problem framework. Moreover, the papers, which focus on energy buffering management, have just consider the effect of battery price and do not investigate the effects of depth of discharge on charge and discharge patterns. On the other hand, the cross-correlations between the traffic of IDCs and its effect on workload distribution are other issues which are neglected in the literature. In this paper, the energy management of distributed IDCs, located in different geographic regions under different day-ahead electricity markets is investigated to maximize the profit of ISPs, covering all these gaps. Here, both energy buffering and the workload distribution are considered in the proposed model under the linear optimization problem framework. A penalty cost function is introduced to eliminate unacceptable transmission delay in QoS constraints due to the inappropriate QoS. The depth of discharge is also considered in battery cost, and its effects are studied on the time-patterns of charging and discharging. Moreover, the paradigm of temporal–spatial correlations between IDCs is discussed in more detail. Based on historical data from IDCs, the probability density functions (PDFs) are extracted from each IDC. Based on these PDFs, the joint PDF of all IDCs is then calculated using the multi-dimensional (-variable) copula approach. In fact, the computation of this multi-variable joint PDF is complicated, and a mathematical approach is required to decompose the proposed approach into a solvable model. The pair-wise copula function is used here as the decomposition approach.. The proposed scheme is formulated via nonlinear optimization problem (NLP) models by considering different constraints such as the QoS, the power balance, the workload distribution, and the battery power management. Then, the model is linearized and is given as a mixed-integer linear optimization problem (MILP). Finally, it is solved by the General Algebraic Modeling System (GAMS) software. In summary, the contributions of this article are listed as follow:

(1) An novel optimization problem is given to attain both optimal workload distribution and energy management of distributed IDCs equipped with energy storage.
(2) Optimal operation strategy for minimizing electricity cost of a group of distributed IDCs is done considering Copula-based Multidimensional Correlation Modeling.
(3) A mixed-integer linear optimization problem (MILP) is introduced to prevent the complexity of the nonlinear optimization problem framework.
(4) Modeling the energy buffering with considering the depth of discharge effect and battery cost in order to improve the electricity consumption pattern.

1.4. Paper organization

In the rest of the paper, the proposed problem would be presented and discussed. To this end, the next section comprehensively explains the mathematical formulation of the proposed problem. In Section 2, at the first, multidimensional copula function, carried out for the modeling of the cross-correlations between the traffic of IDCs, is described, and then, the proposed problem's objective function and its constraints are given by discussing regarding the linearization of the model. Section 3 introduces the case study, which has been applied to evaluate the proposed model and the simulation results interpretations would be provided on this section. Finally, the discussions and conclusions drawn from the study would be given in the last section.

2. Mathematical modeling

This paper addresses the energy management for the ISP's system that comprises some IDCs in different geographic locations under various electricity markets. Several front-end web portal servers receive requests and dispatch them between the data centers in this system. In the proposed system pictured in Fig. 1, an energy storage system is considered for each data center so that the part of the battery capacity is used for an emergency time as well as the remaining will be used for energy buffering to reduce electricity costs. Moreover, for optimal energy buffering the correlation between the IDCs are modeled using the Copula function. As shown in Fig. 1, S front-end web portal servers are considered to receive the client requests, and M data centers provide Internet services. In Fig. 1, $w_{i,t}$ denotes the total requests that are received by *ith* front-end web portal server at time t, and $\xi_{i,j,t}$ is defined as the request rate transferred from the *ith* front-end web portal server to the *jth* IDC at time t. The proposed methodology for energy cost optimization of IDCs consists of two main stages, *i.e.*, the workload modeling of IDCs and the optimization part. In the first stage, IDC's workload uncertainties are modeled using multivariable correlation models, and then this enhanced workload model is used in the optimization stage. The mathematical model of the proposed system is explained below.

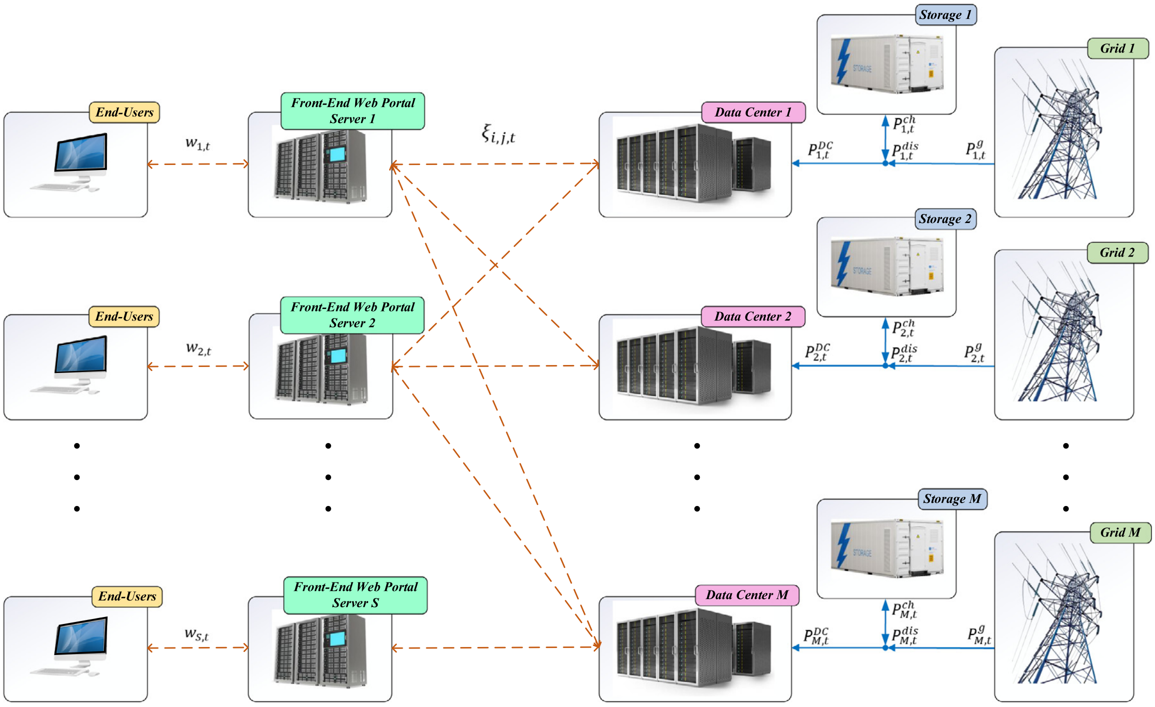

Fig. 1. Architecture of the ISP's system with considering energy storage system for each IDC.

2.1. Multidimensional copula function

In this section, the copula function is used to model the multivariate correlations between IDCs. The workloads of IDCs are naturally correlated, and this concept has been considered in many studies, such as Ref. Chen et al. (2020). The authors in Chen et al. (2020) proposed an algorithm for employing demand–response programs in IDC. In this regard, the load modeling of IDCs is modeled based on a bottom-up approach. Although spatial and temporal correlation is considered in the workload balance constraints of Chen et al. (2020), the proposed model is so straightforward. In fact, the spatial and temporal correlations are modeled by two simple equations. These equations represent that the sum of geo-distributed services must be served (spatial correlation), and the sum of services must be satisfied over a period of time (representing temporal correlation). In this manuscript, however, the paradigm of temporal–spatial correlations between IDCs is discussed in more detail. Since there are many IDCs with corresponding traffic correlations worldwide, the dimension of this global problem for modeling multivariate correlations is enormous. The copula function $C(.)$ for representing the joint distribution function $F(x_1, \ldots, x_n)$ could be calculated as:

$$F(x_1, \ldots, x_n) = C\big(F(x_1), \ldots, F(x_n), \rho\big) \tag{1}$$

Thus Every multivariate distribution function F can be split into its marginal distributions $F_1, \ldots, F_d$ and a copula C. The copula C describes the dependence structure of the random vector $X = x_1, \ldots, x_n$. It could be decoupled into the n-variate copula density function $c_{1\ldots n}(.)$ and marginal density functions $f_1(x_1), \ldots, f_n(x_n)$ using the chain rule.

$$\begin{aligned} f(x_1, \ldots, x_n) =&c_{1\ldots n}\big(F_1(x_1), \ldots, F_n(x_n)\big)\times \\ &f_1(x_1) \times \cdots \times f_n(x_n) \end{aligned} \tag{2}$$

where $c, f_1, \ldots, f_d$ are the probability density functions corresponding to $F_1, \ldots, F_d$, respectively. So the copula density function $c(x_1, \ldots, x_n)$ could be calculates as:

$$c(u_1, \ldots, u_n) = \frac{\partial C(u_1, u_2, \ldots, u_n)}{\partial u_1 \times \partial u_2 \times \cdots \times \partial u_n} \tag{3}$$

where $u_1 = F_1(x_1)$ and similarly $u_n = F_n(x_n)$. As can be seen, it is possible to compute the multidimensional copula function, but it is very complicated to solve for more than two variables. Hopefully, there is a mathematical way to model the correlations between the multidimensional variables through the two-dimensional copula functions based on Pair Copula functions. In fact, the Pair Copula function decouples the problem into solvable copula functions and provides a practical solution to a mathematical model of multidimensional correlations between IDCs. For details of Pair Copula method, Ref. Baboli et al. (2021) could be read. But in short, the main idea is decoupling the multivariate joint distribution function $f(x_1, \ldots, x_n)$, using conditional distribution as:

$$\begin{aligned} f(x_1, \ldots, x_n) =&f_n(x_n) \times f(x_{n-1}|x_n) \\ &\times f(x_{n-2}|x_{n-1}, x_n) \times \cdots \\ &\times f(x_1|x_2, \ldots, x_n) \end{aligned} \tag{4}$$

and later decoupling the conditional distributions to is equal to two-dimensional copulas. Thus all conditional functions in (4) could be decoupled to Pair Copula functions. For example:

$$\begin{aligned} f(x_{n-1}|x_n) =&c_{(n-1)n}\big(F_{n-1}(x_{n-1}), F_n(x_n)\big) \\ &\times f_{n-1}(x_{n-1}). \end{aligned} \tag{5}$$

There are many possible ways to construct Pair Copula for multivariate distributions, *e.g.*, 240 various constructions for 5-variate density (Baboli et al., 2021). However, in this paper, three IDCs are considered, and as a result, the 3-variable pair copula function has been written as an example. The general expression in the three-dimensional case is:

$$\begin{aligned} f(x_1,x_2, x_3) =& f_1(x_1) \times f_2(x_2) \times f_3(x_3) \\ &\times c_{12}\big(F_1(x_1), F_2(x_2)\big) \times c_{23}\big(F_2(x_2), F_3(x_3)\big) \\ &\times c_{13|2}\big(F(x_1|x_2), F(x_3|x_2)\big). \end{aligned} \tag{6}$$

where f_1, f_2, and f_3 in the first line are individual marginal density functions of each input variable, c_{12}, and c_{23} in the second line are unconditional Pair Copula functions and $c_{13|2}$ in the last line is the conditional Pair Copula function.

2.2. Objective functions

In the proposed model, the power consumption cost of IDCs includes three parts. The first part is the cost of the power absorbed by the whole system from the electricity grid. This part can be defined as follows:

$$F_1 = \sum_{t=1}^{24} \sum_{j=1}^{M} \lambda^e_{j,t} \cdot P^g_{j,t} \tag{7}$$

The second part is the cost of the battery, which can be defined as follows:

$$F_2 = \sum_{t=1}^{24} \sum_{j=1}^{M} C^b_j . P^{ch}_{j,t} \tag{8}$$

where, C^b_j defines the cost of battery and $P^{ch}_{j,t}$ indicates the amount of power charged into the battery in *jth* IDC at time *t*. Lithium-ion batteries are commonly used in UPS systems. The battery life cycle (N_{cycle}) depends on three factors: the number of charges and discharges cycles, depth of discharge, and the working temperature. Since N_{cycle} effects on the operation, maintenance (OM&R), and replacement costs of the battery, we can reach a better solution for the charging and discharging pattern with the aim of reducing OM&R costs by considering the depth of discharge effect on the system operation. Actually, the depth of discharge and the number of charging and discharging for a battery are inversely related. Hence, the cost of the battery could be converted into the cost of discharged power. The cost of power discharge according to Yao et al. (2013) is presented by:

$$C^b_j = \frac{\lambda^{bat}}{DoD.N^c.\tilde{P}^{dep}} \tag{9}$$

where, λ^{bat} denotes battery price, *DoD* indicates the maximum battery capacity, N^c defines the number of charging and discharging cycles in the battery life cycle and $\tilde{P}^{dep}$ shows the average depth of discharge. As mentioned, the ISPs have to pay the penalty, if QoS is not met. Therefore, the penalty cost should be considered as follows:

$$F_3 = \sum_{t=1}^{24} \sum_{j=1}^{M} \sum_{i=1}^{S} \Delta t^Q . \alpha . \xi_{i,j,t} \tag{10}$$

in which, Δt^Q is the timeframe that the requests are sent to the IDCs without regarding QoS constraint. $\xi_{i,j,t}$ denotes the request rates transferred from *ith* front-end web portal server to the *jth* data center at time *t*. α is the penalty rate stated in the service level agreement and is defined as dollars per request.

2.3. Constraints

2.3.1. Power consumption of servers

For simplicity, it is assumed that all servers in IDCs are similar and work with the same frequency. Therefore, the power consumption of all servers is equal, i.e., $\widetilde{P}^s_{j,t} = const. \forall t, \forall j.$

2.3.2. Active servers' constraint

The number of active servers in each data center is between zero and the total servers of the data center:

$$0 < n_{j,t} < \overline{N^{ser}_j}, \qquad \forall j,\ \forall t \tag{11}$$

2.3.3. Workload balance model

Each client working with the internet sends its requests to the front-end web portal servers, which transmit them to the IDCs. Therefore, the front-end web portal server receives the workload and dispatches them between IDCs. It can be formulated by (12).

$$\sum_{j=1}^{M} \xi_{i,j,t} = W_{i,t}\ , \qquad \forall i, \forall t \tag{12}$$

2.3.4. QoS constraint

The end-to-end delay requirement is one of the most important factors for clients to evaluate the QoS. It is divided into two parts: the transmission delay and the Queuing delay. The transmission delay is the time that the workload is transmitted from the front-end web portal server to IDC. Also, the Queuing delay is the amount of time the workload must wait to process by the server. (13) illustrates the total delay.

$$E^d_{i,j,t} = Q^d_{j,t} + T^d_{i,j,t}\ , \qquad \forall i,\ \forall t,\ \forall j \tag{13}$$

where $Q^d_{j,t}$ denotes queuing delay, $T^d_{i,j,t}$ indicates transmission delay and $E^d_{i,j,t}$ is the total delay. The delay must be lower than a certain amount that this amount is determined in the Service Level Agreement (SLA).

$$E^d_{i,j,t} < D^{SLA}\ , \qquad \forall i,\ \forall t,\ \forall j \tag{14}$$

where, D^{SLA} defines delay bound. The distance between each front-end web portal server to IDCs is different. The transmission delay is related to the distance between the web portal server to the IDC and workload rate. In the far distance, the transmission delay is becoming more. If the request rate becomes greater, the transmission delay time is increased. Also, the queuing delay model is considered for each data center with the M/M/n queuing mode. In this model, the queuing delay is represented using the following equation:

$$Q^d_{j,t} = \frac{1}{n_{j,t}\ .\mu_{j,t} - \sum_{i=1}^{S} \xi_{i,j,t}}\ , \qquad \forall t,\ \forall j \tag{15}$$

where $\mu_{j,t}$ represents the service rate for each server in the *jth* data center at time *t*.

2.3.5. Charge and discharge constraint

The power level of the battery is obtained following equation:

$$L_{j,t+1} = L_{j,t} + \Delta t^{bat} . (P^{ch}_{j,t} - P^{dis}_{j,t})\ , \qquad \forall t,\ \forall j \tag{16}$$

where $L_{j,t}$ denotes battery power level in the *j*th data center at time *t*. Δt^{bat} is a period that the battery is planned to charge and discharge. The battery power level must be between the emergency level and the battery's maximum capacity. This constraint is following as:

$$\underline{L_j} < L_{j,t} < \overline{L_j}\ , \qquad \forall t,\ \forall j \tag{17}$$

2.3.6. Power balance constraint

In this subsection, the balance constraint of the proposed system can be defined as follows:

$$P^{DC}_{j,t} = n_{j,t}\ .\ \widetilde{P}^s_{j,t}\ , \qquad \forall t,\ \forall j \tag{18}$$

$$P^{DC}_{j,t} = P^g_{j,t} - \frac{P^{ch}_{j,t}}{\theta} + P^{dis}_{j,t}\ , \qquad \forall t,\ \forall j \tag{19}$$

where, $\widetilde{P}^s_{j,t}$ denotes the power consumption of each server. $P^{DC}_{j,t}$ is the power consumption of *jth* data center at time *t*. $P^g_{j,t}$ defines the power absorbed from the electricity grid. In addition, it should be assumed that this power is always positive.

$$0 < P^g_{j,t}\ , \qquad \forall t,\ \forall j \tag{20}$$

2.3.7. Battery constraints

Each battery has a specific capacity for charging and discharging. Also, a battery cannot be charged and discharged simultaneously. (21)–(23) show the battery constraints:

$$0 \le P^{ch}_{j,t} \le \overline{P^{ch}_{j}}\ , \qquad \forall t,\ \forall j \tag{21}$$

$$0 \le P^{dis}_{j,t} \le \overline{P^{dis}_{j}}\ , \qquad \forall t,\ \forall j \tag{22}$$

$$P^{ch}_{j,t}\ .\ P^{dis}_{j,t}{=}0\ , \qquad \forall t,\ \forall j \tag{23}$$

2.4. Linearization

In the optimization problems, attaining the global best solution for NLP is complicated and strict due to the non-convex feasible set. As can be seen, the proposed problem is introduced as NLP due to (15) and (23). QoS constraint can be converted and rewritten as (24) by substituting (13) and (15) into (14).

$$\frac{1}{D^{SLA}{-}T^{d}_{i,j,t}} \le n_{j,t}\ .\ \mu_{j,t}{-}\sum_{i=1}^{S}\xi_{i,j,t}, \qquad \forall i,\ \forall t,\ \forall j \tag{24}$$

$$T^{d}_{i,j,t}{<}D^{SLA}\ , \qquad \forall i,\ \forall t,\ \forall j \tag{25}$$

Moreover, for linearizing (23), an auxiliary binary variable $\beta_{j,t}$ can be defined, and then (26)–(28) can be considered against the battery constraints described in (21)–(23).

$$\beta_{j,t}{\in}\{0,\ 1\}\ , \qquad \forall t,\ \forall j \tag{26}$$

$$P^{ch}_{j,t} \le \overline{P^{ch}_{j}}.\beta_{j,t}\ , \qquad \forall t,\ \forall j \tag{27}$$

$$P^{dis}_{j,t} \le (1{-}\beta_{j,t}).\overline{P^{dis}_{j}}\ , \qquad \forall t,\ \forall j \tag{28}$$

2.5. Proposed problem

The proposed operation problem of Distributed IDC is presented by (29) as a MILP, where, to find the optimal operation point of the problem, the Weighted Sum Method (WSM) is utilized to convert the problem into single-objective optimization form, then it is solved by GAMS software. Respectively, X_1 and X_2 are considered as the vector of the decision variables, and the vector represents the dependent variables of the proposed model.

$$\underset{X_1}{min} \qquad C{=}w_1\frac{F_1}{F^*_{1\ min}}{+}w_2\frac{F_2}{F^*_{2\ min}}{+}w_3\frac{F_3}{F^*_{3\ min}} \tag{29}$$

$$s.t. \qquad Eqs.\ (1)\text{–}(6),\ (10)\text{–}(14),\ (18)\text{–}(21) \tag{30}$$

$$X_1 = [\beta_{j,t}, P^{ch}_{j,t}, P^{dis}_{j,t}, n_{j,t}, P^{g}_{j,t}, \xi_{i,j,t}] \tag{31}$$

$$X_2 = [P^{DC}_{j,t}, L_{j,t}, F_1,\ F_2, F_3] \tag{32}$$

3. Simulation and results

3.1. Case study

In this paper, the ISP's system includes three Google IDC located in different locations in the United States of America, under distinct electricity markets: Mountain View, Houston, and Atlanta (Shao et al., 2013). Also, four web portal end servers are assumed for distributing the workload to IDCs. The electricity prices are determined through a day-ahead process. The electricity price on a particular day in the three locations is shown in Fig. 2 (Shao et al., 2013). As can be seen in Fig. 2, the electricity price in Mountain View and Houston is variable, while the electricity price in Atlanta is constant during the day. For example, in Houston, TX, the electricity price at 15 is almost triple rather than the electricity price at 3. Hence, the battery can be considered to the reliability of the system, it plays the role of energy buffer to reduce the cost of the power consumption of the IDCs in these locations. Also, the transmission delays from the front-end web portal servers to the IDCs have been shown in Figs. 3 to 5 for each location (Shao et al., 2013).

Table 1
The parameter of ISP's servers in IDCs located in a different area.

J	Nmj	P (kW)	μ
Mountain View, CA	50 000	120	2
Atlanta, GA	30 000	120	2
Houston, TX	40 000	120	2

Table 2
The parameters of the batteries installed in Mountain View and Houston.

J	P1 ($)	nt	CM (kWh)	Cr (kWh)	Dr (kWh)
1	240,000	2000	24,000	3000	3000
3	160,000	2000	16,000	2000	2000

Table 3
The amount of the penalty rate (requests/s) for the proposed scenarios in different ranges of time delay (ms).

Scenario #	Time delay (TD) range (ms)			
	<80	80–110	110–150	> 150
S1	0	1.25×10^{-7}	2.5×10^{-7}	3.75×10^{-7}
S2	0	2.5×10^{-7}	5×10^{-7}	7.5×10^{-7}
S3	0	5×10^{-7}	1×10^{-6}	1.5×10^{-6}

The workload of every front-end web portal server is presented as the amount of request per second or request rate in Figs. 6 to 9 by dash line. Based on the proposed methodology in subsection II.A, the multivariate correlation modeling has been performed and the pair copula structures for these four front-end web portal servers are constructed. Based on the forecast error model and the calculated correlations, synthetic data are then generated to model the probable scenarios. The distributions of these generated scenarios are also shown in Figs. 6 to 9 by a boxplot. In these figures, the maximum, minimum, and average values, as well as a standard deviation below and above the average values are shown. In this paper, the average value of the synthetic data is considered instead of the individual forecasts of the front-end web portal servers, which provides a more realistic representation of the servers' workload in the day-ahead energy market.

The parameters of servers included the number of servers in IDCs, the power consumption of all servers, and the service rates have also been shown in Table 1. Due to the time-variability of electricity, two high-capacity batteries are considered for IDCs located in Houston, TX, and Mountain View, CA for energy buffering. A portion of the capacity of each battery is considered for the reliability of IDC and the remaining capacity is for energy buffering. The parameters of batteries have been listed in Table 2. These parameters include the battery's price, the time of charge and discharge of the battery in the life cycle, the maximum capacity, and the charging and discharging rate. It also supposes that the battery charging efficiency is equal to 0.96.

The penalty rate for IDC is applied when QoS is not satisfied. Three scenarios with three different levels (low, medium, and high level) for penalty rates are considered. Table 3 demonstrates the penalty rate (requests/s) for these three scenarios in different ranges of time delay.

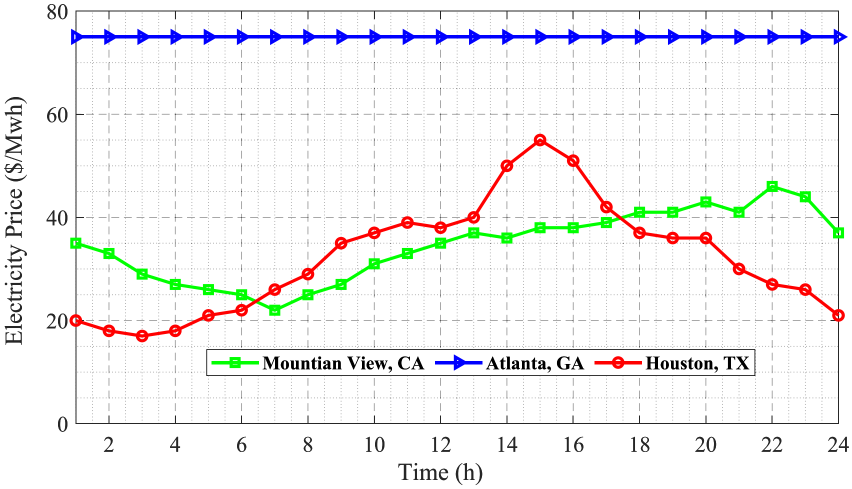

Fig. 2. Electricity price for the different locations where ISP's data centers are located in the different areas (Shao et al., 2013).

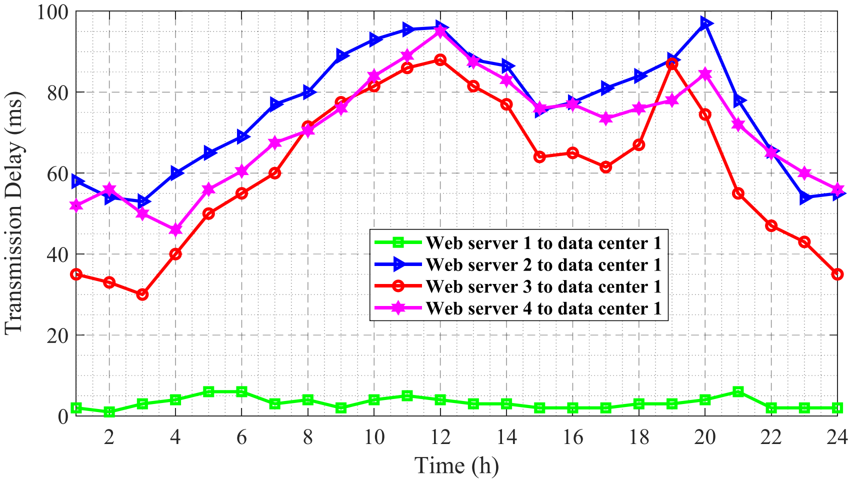

Fig. 3. Transmission delay from front-end web portal servers to data centers located in Mountain View.

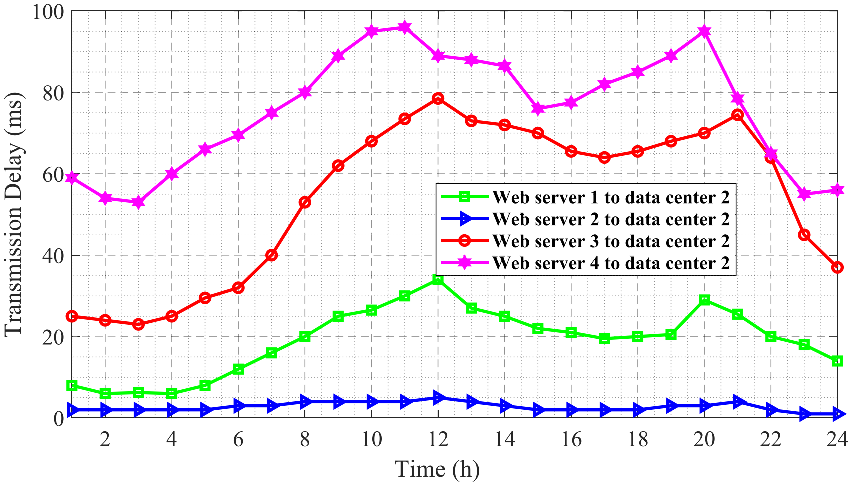

Fig. 4. Transmission delay from front-end web portal servers to data centers located in Atlanta, GA.

3.2. Results and discussions

In this section, we analyze the results obtained from the simulation of the proposed problem. The number of active servers is shown for the scenarios in Figs. 10–12. It is obvious that the number of active servers in location j at time t should be proportional to its costs. Fig. 6 demonstrates the obtained result of the scenario I in which the penalty rate has been considered with a low-level rate. As can be seen in Fig. 6, the servers located in Atlanta are

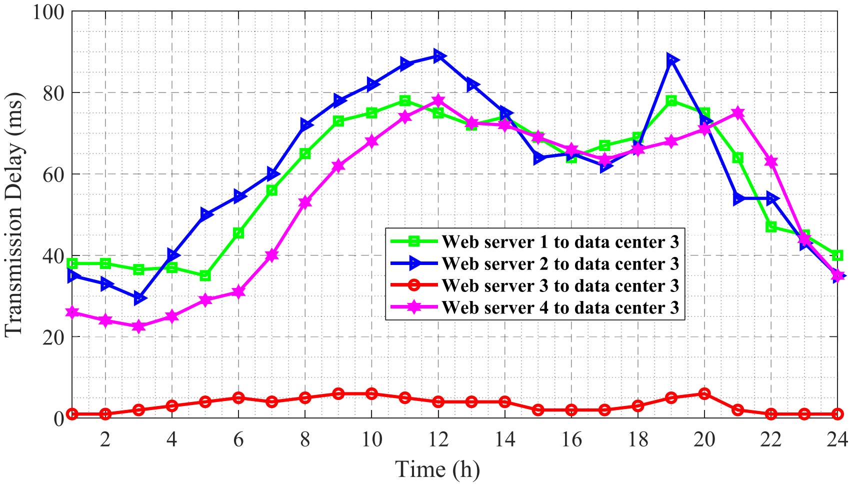

Fig. 5. Transmission delay from front-end web portal servers to data centers located in Houston, TX.

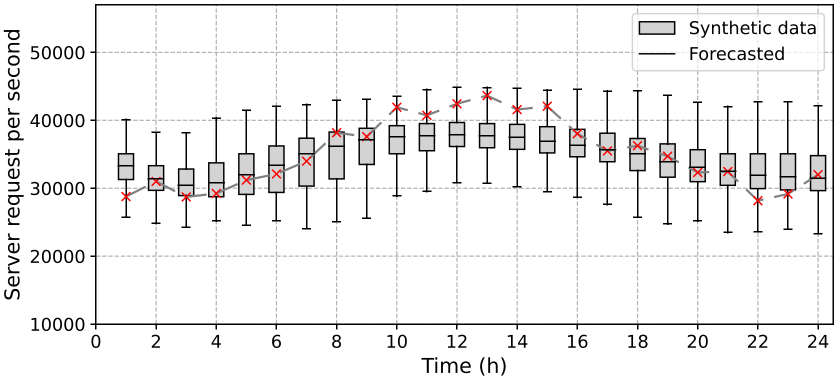

Fig. 6. Request rate from users to front-end web portal of 1st server.

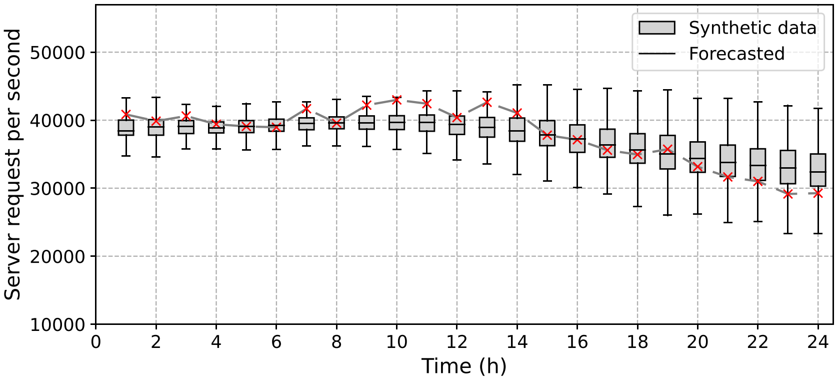

Fig. 7. Request rate from users to front-end web portal of 2nd server.

not utilized for the Internet services in the operation duration due to high electricity prices in the local electricity market of this area. In the first scenario, regarding the obtained results, decision making for utilization of the ISP's server is done only base on electricity price and the TD index does not affect this decision-making due to being low the penalty rate. It is clear from Fig. 2 that the electricity price has the lowest rate in Mountain View in comparison to other areas from 7 A.M to 5 P.M. Therefore, in this duration, all ISP's servers of this city have been selected to process some part of client's request. The remained part of the client's request has been distributed to the servers located Houston because the electricity price of this city is lower than the electricity price of Atlanta.

In contrast to the first scenario, the TD index affects the selection of the active servers to carry out the service of the internet provider company for other scenarios. Fig. 11 shows the result obtained from scenario II, in which the penalty rate with a medium level rate is considered. Comparing this figure with Fig. 6, it is extracted that the TD index changes the decision-making for the duration from 10 A.M. to 1 P.M. In this way, the priority for carrying out the processing of the client request is transferred to the servers located in Houston, TX, despite the higher electricity price than the price of Mountain View. Also, for Scenario III (Fig. 12), the servers' participation in Atlanta is seen. As an interesting point, in the mentioned duration, there is not only a priority problem but also the main problem of how

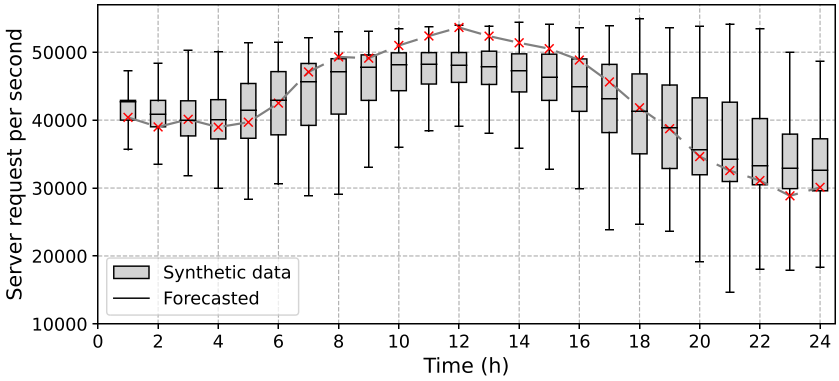

Fig. 8. Request rate from users to front-end web portal of 3rd server.

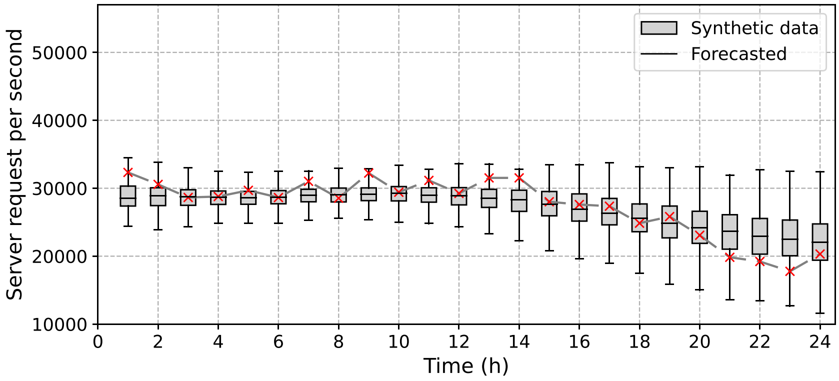

Fig. 9. Request rate from users to front-end web portal of 4th server.

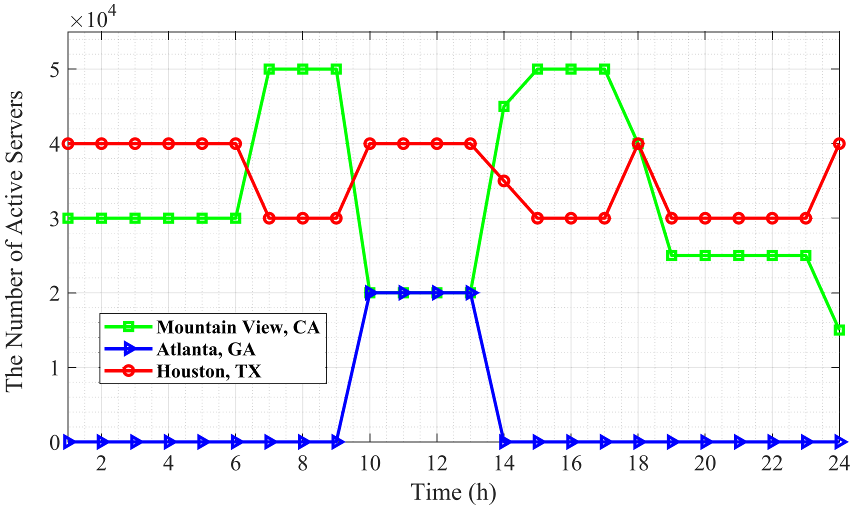

Fig. 10. The obtained result regarding the number of active servers for scenario I (with low-level penalty rates).

much the requests should be dispatched to the servers located in Mountain View and Atlanta. It can be seen that none of the servers of these two areas are utilized in their maximum capacity.

Figs. 13–15 show the obtained results of scenario III about the workload distributed from front-end web portal servers to the IDCs of Mountain View, Atlanta, and Houston, respectively. In all transfers, the transmission delay time is under 80 ms. Due to the high penalty rate, when the delay is closer to 80 ms, the rate of workload received by the IDC should be adjusted to assist the reduction of Queuing delay to keep the general delay under 80 ms. As shown in Fig. 13, the transmission delay from front-end web portal servers 2, 3, and 4 to the IDCs located in Mountain View is more than 80 ms, from 10 A.M. to 1 P.M. (Fig. 3), and this would make the workload transmission to be not economically affordable. Hence, the front-end web portal server 1 transfers the requests to the IDC of this area.

At the mentioned period, another part of the requests is transferred by the front-end web portal servers 3 and 4 to the IDC at Houston (Fig. 15), where the electricity cost is lower than in Atlanta. Finally, the workload of front-end web portal server 2 is transferred to the IDCs in Atlanta (Fig. 14) because the transmission delay from this server to the IDCs of other areas is more than 80 ms at this period. By considering the medium rate for the penalty rate, the distribution of workload will be changed,

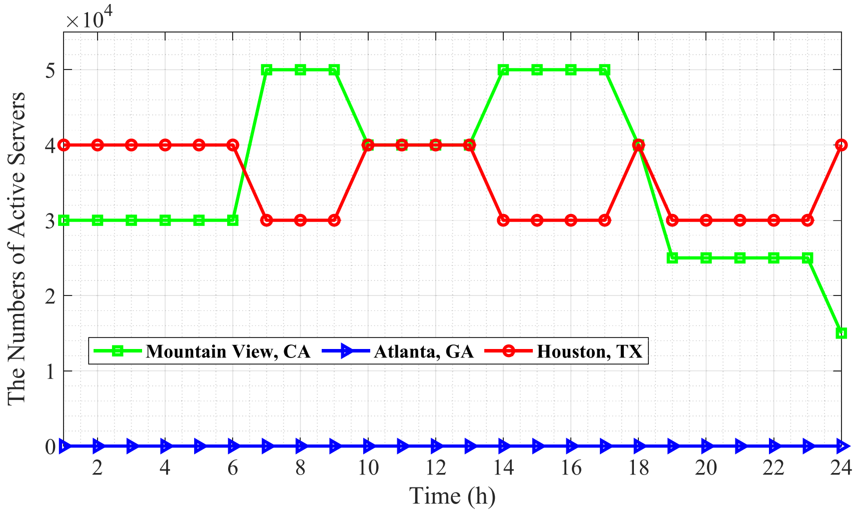

Fig. 11. The obtained result regarding the number of active servers for scenario II (with medium-level penalty rates).

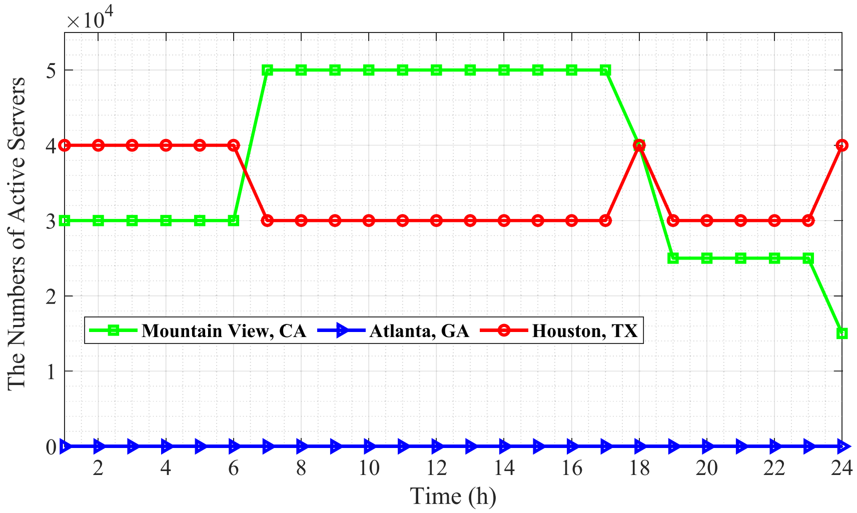

Fig. 12. The obtained result regarding the number of active servers for scenario III (with high-level penalty rates).

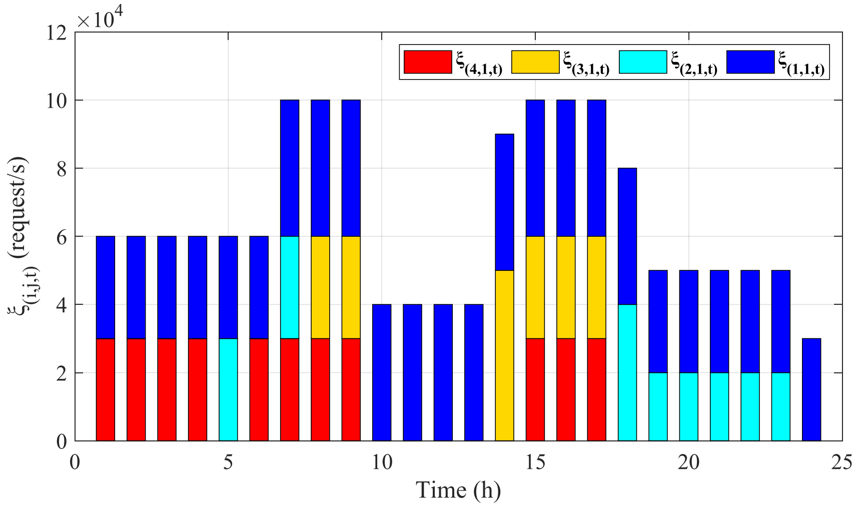

Fig. 13. The workload assigned from front-end web portal servers to IDCs located in Mountain View.

and the workload of front-end web portal server 2 is transferred to the IDCs at Mountain View. Because, in this case, the power consumption cost for transferring requests to Atlanta's IDCs is more than the total cost for transferring requests to Mountain

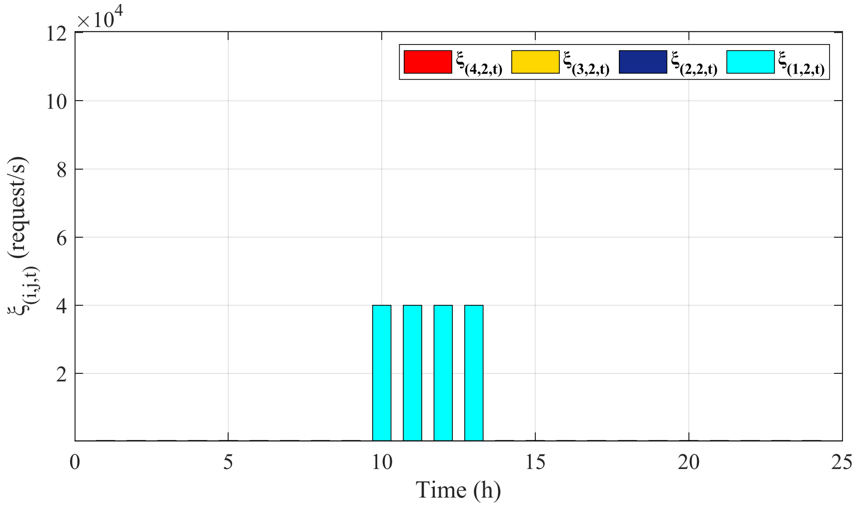

Fig. 14. The workload assigned from front-end web portal servers to IDCs located in Atlanta.

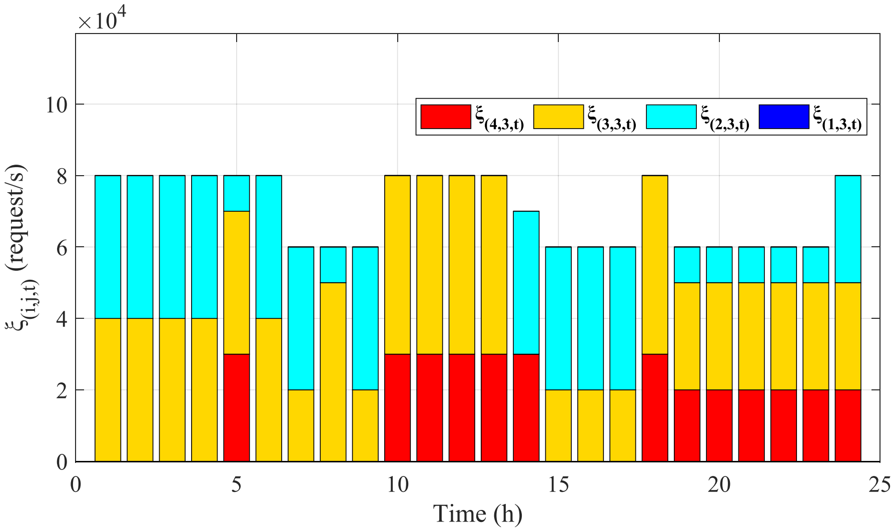

Fig. 15. The workload assigned from front-end web portal servers to IDCs located in Houston.

View's IDCs and to be penalized (because the QoS constraint is not satisfied).

Figs. 16 and 17 show the charging and discharging batteries in Mountain View and Houston, respectively. Moreover, Fig. 18 displays the energy level of these batteries. Respectively, the power level of batteries in Mountain View and Houston are always at least 3 MW and 2 MW for emergency time. Considering the electricity price (Fig. 2) in these two areas, the logical trend of charging and discharging batteries is seen from Figs. 16 and 17. Considering the obtained results demonstrated in Figs. 12 and 16, it can be seen that the parts of the workload requests are also transferred to the Mountain View's IDC from 7 P.M. to 11 P.M., despite being higher the electricity price in this area than the Houston price. This issue shows the advantage of energy buffering.

The hourly total electricity cost for three different dispatching methods, including the average dispatch method, the optimal dispatch method with and without considering energy storage, is depicted in Fig. 19. Because of the too high electricity price in Atlanta, GA, the total electricity cost in the average dispatch workload is high. The workload transmission to the IDCs of this area has economic justification just in the case of QoS is not satisfied, and the penalty rate is too high. On the other hand, the power consumption cost is reduced by optimal workload balancing and peak shaving when the energy buffering is applied. As a result, the power consumption scheme of the IDCs would be improved.

The cost of the overall power consumption of IDC in the average dispatch is 9716.2 $ and in the optimum dispatch method without considering the battery is equal to 6830.36 $ and with considering the battery is 6337.45 $. Respectively, the electricity cost in optimum dispatch with energy buffering has decreased %34.77 and %7.22 compared to the average dispatch and optimum dispatch.

As another critical factor, the battery price affects the energy buffering scheme. The battery charge and discharge scheme are significantly changed by changing the battery price. To demonstrate this issue, different battery prices are considered for the batteries installed in Mountain View and Houston IDCs, which are 72 000 $ and 48 000 $, respectively. The energy level of batteries in Mountain View and Houston IDCs are as follows in Fig. 20. As can be seen in this figure, due to the high price of the battery, the utilization of the entire battery capacity in energy buffer operation is not economical. For instance, in Mountain View's IDC, the battery is charged only for three hours; In other words, about %38 of the battery capacity is only used in the buffering operation.

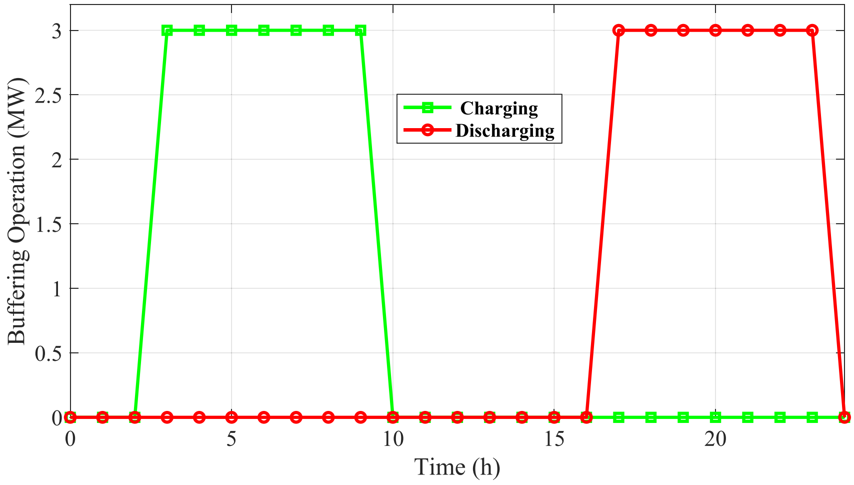

Fig. 16. The buffering operation for the batteries installed for the IDCs at Mountain View at the operation period.

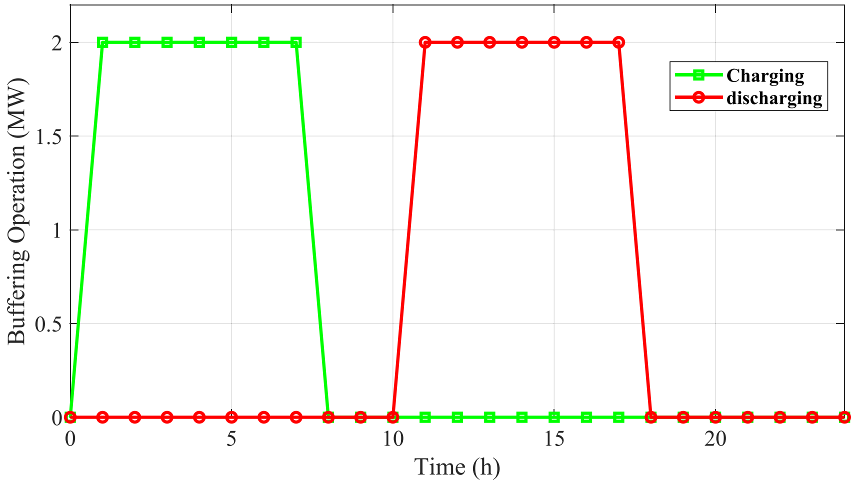

Fig. 17. The buffering operation for the batteries installed for the IDCs at Houston at the operation period.

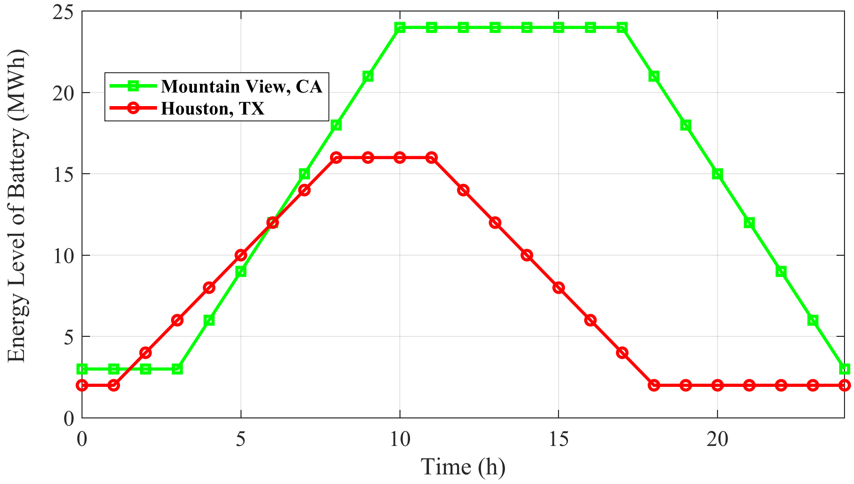

Fig. 18. The power level of batteries installed for IDCs located in Mountain View and Houston at the operation period.

In Table 4, the system actions are shown at two different hours. At 1 A.M., Houston's electricity cost is cheaper than to other regions. Therefore, transferring the workload to the IDCs in this area is prioritized. The workload of front-end web portal servers 2 and 3 are transferred to IDCs at Houston. Then, front-end web portal servers 1 and 4 transfer the workload to IDCs Mountain, which is cheaper than Atlanta. At this time, QoS is met. Also, the battery used in Houston's IDCs is charged during

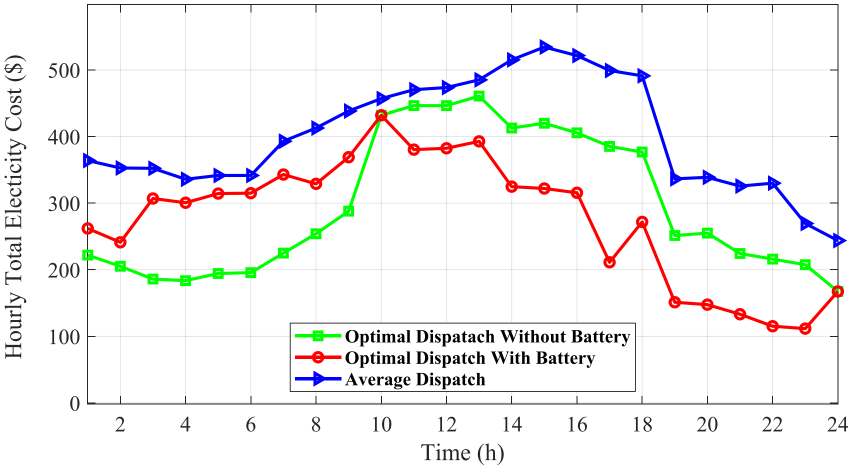

Fig. 19. Total hourly total electricity cost in three different dispatching methods.

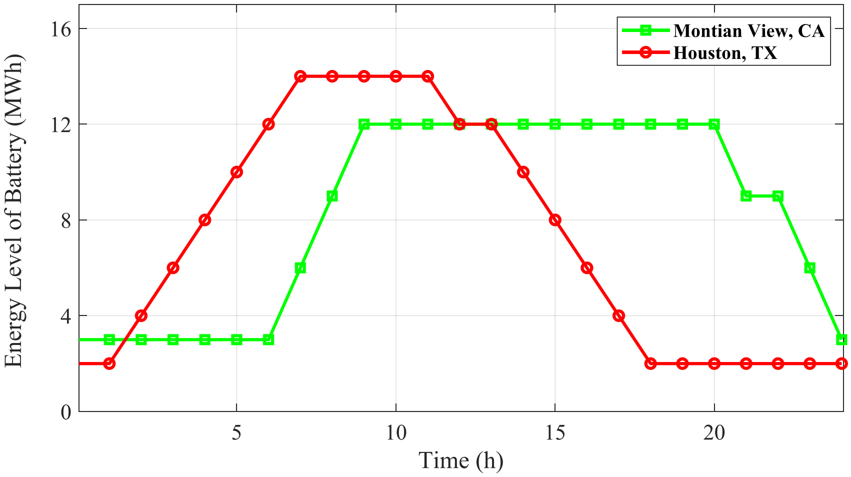

Fig. 20. The power level of the batteries after considering different battery prices.

this time because the cost of electricity is cheap. At 1 P.M., transferring the workload to Mountain's IDCs is prioritized; But only the front-end web portal server 1 transfers its requests to these IDCs because other requests transmission does not meet QoS constraints. Therefore, servers 3 and 4 transfer workload to Houston's IDCs because the electricity price in Houston is lower than in Atlanta. At this time, the transmission of requests in the IDCs located in Atlanta, despite its high electricity cost, is more economical than the transmission of requests in the IDCs located in Mountain View. The battery in Houston's IDCs gets discharged to supply part of the power consumption of IDCs and decreases the total energy cost.

Table 4
The analyze result for two hours.

J	t	M_i	$u_j(t)$	$q_j(t)$	C_{opt}	C_{ave}	C	ΔC
1	1	30 031	0	0	222	364	262	%28
2	1	0	–	–				
3	1	40 000	3e6	0				
1	13	20 007	0	0	461	485	393	%19
2	13	20 107	–	–				
3	13	40 000	0	3e6				

4. Conclusion

The high power consumption of IDCs, especially during peak hours, in addition to the economic losses, increases the risk of power outages and has become a critical concern for ISPs. In this paper, a novel scheme for energy cost optimization is presented to reduce the total cost of an ISP by optimal workload dispatch between IDCs and battery energy management. To make the work comprehensive, the literature gap in this context is identified and addressed. The main gaps in the field, in addition to the lack of works covering the three main possible measures for energy cost minimization (i.e., optimal internal system, optimal workload distribution, and energy buffering), include:

- The complexity of the nonlinear optimization problem framework in the previous studies.

- Neglecting the effects of depth of discharge on charge and discharge patterns of batteries.
- Neglecting the impact of cross-correlations between the IDCs' traffic.

To address these all, a copula-based multidimensional correlation analysis method has been used to model the multivariate correlations between IDCs, as well as, energy buffering has been considered in the proposed model, and the effect of depth of discharge has been given in battery price. Therefore, an optimal balance has been created between energy cost saving and battery cost. Moreover, a penalty term has been added to the objective function to prevent non-compliance with the QoS conditions, and the effect of the penalty rate has been shown and investigated on the workload distribution scheme. At first, the proposed problem has been presented as an NLP, and then it has been converted to a MILP framework with linearization techniques and solved by the BARON solver in GAMS software. Finally, the results have demonstrated that the proposed method has improved the electricity consumption pattern, furthermore, proved the importance of the depth of discharge to reach a more economical and operational solution. Moreover, the results have presented how energy buffering can affect distributing the workload and maximize the ISP's profit in a competitive energy market.

CRediT authorship contribution statement

Mohammad Ali Lasemi: Conceptualization, Methodology, Software, Writing – original draft. **Shahin Alizadeh:** Methodology, Software, Writing – original draft. **Mohsen Assili:** Supervision, Conceptualization, Writing – reviewing and editing. **Zhenyu Yang:** Supervision, Writing – reviewing and editing. **Payam Teimourzadeh Baboli:** Conceptualization, Writing – reviewing and editing. **Ahmad Arabkoohsar:** Writing – reviewing and editing. **Amin Raeiszadeh:** Methodology, Software. **Michael Brand:** Methodology, Software. **Sebastian Lehnhoff:** Writing – reviewing and editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research is partially supported via AAU Bubble Project - A Novel Zero/Negative Emission Energy System Integrating Power-to-X, Allam Cycle, Carbon Capture, Utilization and Storage (CCUS) and CO2-based Energy Storage (AAU Project No.762985).

Data availability

The authors do not have permission to share data.

References

Adrah, C.M., Palma, D., Kure, Ø., Heegaard, P.E., 2020. A network design algorithm for multicast communication architectures in smart transmission grids. Electr. Power Syst. Res. 187, 106484.

Baboli, P.T., Brand, M., Lehnhoff, S., 2021. Stochastic correlation modelling of renewable energy sources for provision of ancillary services using multidimensional copula functions. In: 2021 11th Smart Grid Conference (SGC). IEEE, pp. 1–6.

Chen, M., Gao, C., Shahidehpour, M., Li, Z., Chen, S., Li, D., 2020. Internet data center load modeling for demand response considering the coupling of multiple regulation methods. IEEE Trans. Smart Grid 12 (3), 2060–2076.

Cheng, H., Liu, B., Lin, W., Ma, Z., Li, K., Hsu, C.-H., 2021. A survey of energy-saving technologies in cloud data centers. J. Supercomput. 77 (11), 13385–13420.

Cheung, H., Wang, S., 2019. Reliability and availability assessment and enhancement of water-cooled multi-chiller cooling systems for data centers. Reliab. Eng. Syst. Saf. 191, 106573.

Gong, X., Zhang, Z., Gan, S., Niu, B., Yang, L., Xu, H., Gao, M., 2020. A review on evaluation metrics of thermal performance in data centers. Build. Environ. 177, 106907.

Gu, L., Zeng, D., Barnawi, A., Guo, S., Stojmenovic, I., 2014. Optimal task placement with QoS constraints in geo-distributed data centers using DVFS. IEEE Trans. Comput. 64 (7), 2049–2059.

He, Z., Xi, H., Ding, T., Wang, J., Li, Z., 2021. Energy efficiency optimization of an integrated heat pipe cooling system in data center based on genetic algorithm. Appl. Therm. Eng. 182, 115800.

Hintemann, R., Hinterholzer, S., 2019. Energy consumption of data centers worldwide. In: Business, Computer Science (ICT4S).

Hu, X., Li, P., Sun, Y., 2021. Minimizing energy cost for green data center by exploring heterogeneous energy resource. J. Mod. Power Syst. Clean Energy 9 (1), 148–159.

Jin, C., Bai, X., Yang, C., Mao, W., Xu, X., 2020. A review of power consumption models of servers in data centers. Appl. Energy 265, 114806.

Koronen, C., Åhman, M., Nilsson, L.J., 2020. Data centres in future European energy systems—energy efficiency, integration and policy. Energy Effic. 13 (1), 129–144.

Kwon, S., 2020. Ensuring renewable energy utilization with quality of service guarantee for energy-efficient data center operations. Appl. Energy 276, 115424.

Lasemi, M.A., Arabkoohsar, A., Hajizadeh, A., Mohammadi-ivatloo, B., 2022. A comprehensive review on optimization challenges of smart energy hubs under uncertainty factors. Renew. Sustain. Energy Rev. 160, 112320. http://dx.doi.org/10.1016/j.rser.2022.112320.

Li, J., Li, Z., 2020. Model-based optimization of free cooling switchover temperature and cooling tower approach temperature for data center cooling system with water-side economizer. Energy Build. 227, 110407.

Liu, L., Zhang, Q., Zhai, Z.J., Yue, C., Ma, X., 2020. State-of-the-art on thermal energy storage technologies in data center. Energy Build. 226, 110345.

Lyu, J., Zhang, S., Cheng, H., Yuan, K., Song, Y., Fang, S., 2021. Optimal sizing of energy station in the multienergy system integrated with data center. IEEE Trans. Ind. Appl. 57 (2), 1222–1234.

Marshall, Z., Duquette, J., 2022. A techno-economic evaluation of low global warming potential heat pump assisted organic Rankine cycle systems for data center waste heat recovery. Energy 242, 122528.

Peng, Y., Li, J., Hai, H., Jiang, X.-Q., Fawaz, A.-H., Park, S., 2021. Cost optimization of distributed data centers via computing workload distribution for next generation network systems. Phys. Commun. 46, 101340.

Rahmani, R., Moser, I., Cricenti, A.L., 2020. Modelling and optimisation of microgrid configuration for green data centres: A metaheuristic approach. Future Gener. Comput. Syst. 108, 742–750.

Sajid, S., Jawad, M., Qureshi, M.B., Khan, M.U.S., Ali, S.M., Khan, S.U., 2019. A conditional-constraint optimization for joint energy management of data center and electric vehicle parking-lot. In: 2019 Tenth International Green and Sustainable Computing Conference (IGSC). IEEE, pp. 1–6.

Shao, H., Rao, L., Wang, Z., Liu, X., Wang, Z., Ren, K., 2013. Optimal load balancing and energy cost management for internet data centers in deregulated electricity markets. IEEE Trans. Parallel Distrib. Syst. 25 (10), 2659–2669.

Shehabi, A., Smith, S., Sartor, D., Brown, R., Herrlin, M., Koomey, J., Masanet, E., Horner, N., Azevedo, I., Lintner, W., 2016. United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory.

Sun, J., Chen, M., Liu, H., Yang, Q., Yang, Z., 2020. Workload transfer strategy of urban neighboring data centers with market power in local electricity market. IEEE Trans. Smart Grid 11 (4), 3083–3094.

Temiz, M., Dincer, I., 2022. A unique bifacial PV and hydrogen-based cleaner energy system with heat recovery for data centers. Appl. Therm. Eng. 118102.

Wang, P., Cao, Y., Ding, Z., 2020. Flexible multi-energy scheduling scheme for data center to facilitate wind power integration. IEEE Access 8, 88876–88891.

Yao, J., Liu, X., Zhang, C., 2013. Predictive electricity cost minimization through energy buffering in data centers. IEEE Trans. Smart Grid 5 (1), 230–238.

Zhang, Y., Wilson, D.C., Paschalidis, I.C., Coskun, A.K., 2021a. A data center demand response policy for real-world workload scenarios in HPC. In: 2021 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, pp. 282–287.

Zhang, Y., Wilson, D.C., Paschalidis, I.C., Coskun, A., 2021b. HPC data center participation in demand response: an adaptive policy with QoS assurance. IEEE Trans. Sustain. Comput..

Zhang, W., Zavala, V.M., 2021. An electricity market clearing formulation for harnessing space-time, load-shifting flexibility from data centers. arXiv e-prints, arXiv–2105.

Zhou, Q., Lou, J., Jiang, Y., 2019. Optimization of energy consumption of green data center in e-commerce. Sustain. Comput.: Inf. Syst. 23, 103–110.