An Optimal Task Placement Strategy in Geo-Distributed Data Centers involving Renewable Energy

RAN WANG\textsuperscript{1, 2}, (Member, IEEE), YIWEN LU\textsuperscript{1}, KUN ZHU\textsuperscript{1, 2}, (Member, IEEE), JIE HAO\textsuperscript{1, 2}, (Member, IEEE), PING WANG\textsuperscript{3}, (Senior Member, IEEE), and YUE CAO\textsuperscript{4}, (Member, IEEE)

1College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China (e-mail: {wangran, luyiwen, zhukun, haojie}@nuaa.edu.cn)
2Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, China
3Department of Electrical Engineering and Computer Science, York University, Canada (e-mail: pingw@yorku.ca)
4Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, UK (e-mail: yuecao@northumbria.ac.uk)

Corresponding author: Dr. Kun Zhu (e-mail: zhukun@nuaa.edu.cn).

This work was supported by the Fundamental Research Funds for the Central Universities, NO. NS2017072.

ABSTRACT Nowadays, modern data centers are seeking for importing renewable energy together with conventional energy in order to be more environment-friendly and to reduce operation expenditures. Meanwhile, considering the fact that electricity prices and renewable energy generations are diverse in time and geography, a task scheduling strategy should be designed to ensure the efficient and economic operations of data centers. In this paper, an optimal task placement strategy is presented for geo-distributed data centers powered by mixed renewable and conventional energies with dynamic voltage and frequency scaling (DVFS) technique. We aim at minimizing the total electricity cost and making full use of the renewable energy so as to construct green and economic data centers. The optimal task placement problem is formulated as a mixed integer nonlinear problem (MINLP), in which the quality-of-service (QoS) constraint is restricted by an M/G/1 queuing model. To tackle the complexity of the MINLP, we first transform it into a tractable form, then develop an optimal server activation configuration (SAC) and task placement algorithm to solve it. The proposed algorithm can obtain the global optimal solution of the electricity minimization problem and meanwhile dramatically reduce the complexity of the problem solving. Finally, evaluations based on real-world traces exhibit impacts of different system parameters on the electricity cost and sever activation configurations, which prove the superiority of our proposed algorithm and provide us some illuminations on how to build cost-effective and eco-friendly data centers.

INDEX TERMS data centers; dynamic voltage and frequency scaling (DVFS); renewable energy; sever activation configuration (SAC); task placement.

I. INTRODUCTION

Data centers are always working in 24 hours day and night without shutting down, consuming enormous electricity which mainly comes from conventional energy such as coal, oil and natural gas, thus causing serious air pollution to the environment [1, 2]. There is an estimation that a data center with $5 \times 10^4$ servers may use over 100 million KWh per year [3], as much as the urban energy consumption for $10^5$ households. Under such circumstance, renewable energy powered data centers have attracted more attentions in recent years due to their low costs and environment-friendly properties. While renewable energy offers a cheaper and cleaner electricity supply, the integration of climate-dependent renewable energy into data centers also imposes great operation challenges because of its high inter-temporal variation and limited predictability. In addition, electricity price and renewable energy generation (REG) are diverse in time and geography domains, hence an optimal task placement strategy should be properly designed to dynamically coordinate the renewable energy generation and task placement of data centers so that the energy expenditure for operating the data centers is minimized while tasks can be executed in time. In recent years, a power management technique, namely dynamic voltage and frequency scaling (DVFS) is employed to dynamically adjust server’s operating voltage and frequency. For example, if user requests flood into front por-
tals, servers should improve operating frequencies to process more user requests within a short time. If user requests are not large, servers can operate at low frequencies for energy conservation. The DVFS technique certainly enables more flexible task scheduling; meanwhile it will further complicate the task placement process. As a consequence, optimal task placement in renewable energy powered data centers considering DVFS technique becomes a practical and important research problem.

### A. RELATED WORK

Some research works have studied task scheduling problem in data centers. For pure task placement problem, in [4], H. Xu et al. propose to make workload management temperature aware. The workload management problem is formulated as a joint optimization of request routing for interactive workloads and capacity allocation for batch workloads. Reference [5] considered the stochastic multi-stage job scheduling problem on salable resources in data centers to maximize the utilization of rented VMs over a certain period of time. X. Zhao et al. in [6] present a model-predictive control (MPC) based scheduling strategy called ThermoRing to reduce cooling costs in data centers. ThermoRing copes with thermal emergencies by controlling task allocations to computing nodes. Mateusz et al. in [7] target an online scheduling problem of work-flows consisting of interrelated tasks in a data center. A novel methodology named Minimum Dependencies Energy-efficient Directed Acyclic Graphs scheduling is then developed. In reference [8], Y. Wang et al. build an optimization framework of an interaction system of the smart power grid, jointly accounting for the service request dispatch and routing problem in the data center with the power flow analysis in power grid. Literature [9] proposes a temporal task scheduling algorithm (TTSA) to effectively dispatch all arriving work loads to private data centers and public clouds. In each iteration of TTSA, the cost minimization problem is modeled as a mixed integer linear program (MILP). L. Yu et al. [10] optimize the problem concerning joint workload and battery scheduling with heterogeneous service delay guarantees for data centers. An online operation algorithm to solve the problem based on Lyapunov optimization technique is then designed.

As for task scheduling in renewable energy powered data centers, N. Hogade et al. in [11] design techniques for geographical load distribution that will minimize the energy expenditure for executing incoming tasks considering many aspects of the overall system. In [12], the authors propose a novel analytical model to calculate profit in large data centers without and with behind-the-meter renewable power generation. Then the derived profit model is adapted to develop an optimization-based profit maximization strategy for data centers. A robust workload and energy management framework for sustainable data centers is developed in [13]. To deal with the uncertainty from renewable energy generation, the resource allocation task is formulated as a robust optimization problem that minimizes the worst-case net cost. S. Chen et al. in [14] build a comprehensive framework covering the costs of server power, cooling power, and hardware maintenance. A joint optimization of the costs of electricity and server maintenance is then introduced. A. N. Toosi et al. in [15] propose a framework for load balancing of web application requests in geo-distributed data centers based on the renewable energy in each region. Another work can be found in [16], where a unified management approach is proposed allowing data centers to adaptively respond to intermittent availability of renewables under long-term quality-of-service (QoS) requirements.

For task placement with DVFS technique, a common way is to operate servers at high frequency when processing large size of user requests and operate servers at low frequency when their workload is small. Note that high frequency usually leads to high energy consumption and low frequency means long execution time, thus the electricity cost minimization and QoS requirements should be balanced. L. Gu et al. in [17] design an iterative searching algorithm to solve the complex user requests allocation problem. S. Wang et al. in [18] propose a DVFS-based task model describing the QoS requirements of tasks, without the prior knowledge of execution time of tasks, and then transform the task scheduling problem into minimizing the total energy consumption ratio. Z. Tang et al. [19] propose a DVFS-enabled energy-efficient work-flow task scheduling algorithm by merging the relatively inefficient processors through reclaiming the slack time, which could make use of the slack time recurrently. H. Lei et al. in [20] built a four-objective framework to optimize the utilization of renewable energy, completion time of tasks, total energy consumption and processing rate of tasks. An enhanced multi-objective co-evolutionary algorithm is proposed to solve the problem. C. Wu et al. in [21] propose a scheduling algorithm for the cloud data center with DVFS technique, which aimed at efficiently increasing resource utilization hence decreasing the energy consumption for processing jobs. Readers interested in task scheduling problem in data centers can refer to surveys [22] and [23] for a more comprehensive understanding.

### B. MAIN CONTRIBUTIONS

The above papers [4]-[23] optimized over the operations of data centers either with no DVFS-enabled servers where the main consideration was the number of activated servers, or with no renewable energy imported whose goal was to minimize the total energy consumption. Most literature normally modeled the system from the micro viewpoint. To conquer these weaknesses, we investigate the task placement problem in geo-distributed data centers considering QoS requirements and DVFS technique, aiming at minimizing the electricity cost and making full use of renewable energy. The main contributions of this paper are summarized as follows:

- We describe the dynamic task flow in each server as an M/G/1 queue which is then adopted to formulate the QoS constraint in data centers. To the best of our knowledge, this is the first time that the M/G/1 queue...
is employed to model task flow in renewable energy powered data centers.

- We formulate the optimal DVFS based task placement problem into a mixed integer nonlinear problem (MINLP) to determine the distribution flow of user requests to geo-distributed data centers. Factors such as the REG and electricity price, the number of activated servers in each data center, the amount of user requests each data center processes, the amount of conventional energy each data center consumes and at what frequency level each server operates are jointly considered.

- To deal with the high complexity and non-convexity of MINLP, we first transform the original problem into an equivalent but less complicated one. Then an optimal server activation configuration (SAC) and task placement algorithm is developed to efficiently solve the transformed problem in polynomial time, where a global optimal solution is achieved.

- Numerical results based on real-world traces evaluate the impacts of different parameters on electricity cost and server operating frequencies, providing some insights on how to design task placement policies in data centers. Additionally, the proposed task placement strategy can reduce the electricity cost considerably, providing us some insights on how to build economic and sustainable data centers.

The remainder of this paper is organized as follows: Section II introduces the particulars of the system model. In Section III, we show the mathematical formulation of the initial optimal task placement problem and then transform it into an easier-solving one, which drastically decreases the complexity of the original problem. In Section IV, an optimal SAC and task placement strategy is designed to solve the relaxed problem. Simulation results based on real-world data are presented in Section V. Finally we conclude the paper in Section VI.

II. SYSTEM MODEL

In this section, the particulars of the system operation are presented in details below.

A. SYSTEM ARCHITECTURE OF GEO-DISTRIBUTED DATA CENTERS

Generally, we model over a system with \( N \) geo-distributed data centers and \( M \) front portals deployed on different sites. For easier understanding, a data center system composed of 3 front portals and 4 data centers powered by mixed renewable and conventional energies is illustrated in Fig. 1, i.e., \( M = 3 \) and \( N = 4 \). Each front portal is responsible for collecting surrounding user requests and delivering them to appropriate data centers. A data center usually consists of great number of servers, with \( S_n \) and \( Pr_n \) denoting server number and electricity price at data center \( n \), respectively. Note that a multi-electricity market is considered in this paper, where electricity price presents regional diversities.

B. DYNAMIC VOLTAGE AND FREQUENCY SCALING TECHNIQUE

DVFS is essentially a power-saving technique, allowing dynamic adjustment of working frequency of a microprocessor via regulating the supplied voltage. As we all know, higher supplied voltage normally indicates higher frequency and larger power consumption meanwhile. Under such case, DVFS technique is utilized in data centers to set reasonable operating voltage and clock frequency based on the actual power consumption of the chip at the time, which ensures sufficient but not exceeding power supply. Generally, power consumption \( P \) is a function of the supplied voltage \( v \) and the working frequency \( f \) in the range of \( f_{\text{min}} \) to \( f_{\text{max}} \), with \( f_{\text{min}} \) and \( f_{\text{max}} \) denoting the minimum and maximum frequency, respectively. Thus,

\[
P = B v^2 f + P_{\text{static}}
\]

where \( B \) is a coefficient related to different processors and \( P_{\text{static}} \) is the static power consumption independent of \( f \), which is mainly caused by leakage currents in devices and circuits. Based on the existing research, \( P_{\text{static}} \) can be viewed as a constant occupied 10% ~ 60% of the maximum power consumption according to different architectures and technologies. In general, the relationship between the supplied voltage and operating frequency is \( v \propto f^\beta \), where \( \beta \) is always set as 1. Let \( \alpha = 2\beta \), then we have

\[
P = B f^{\alpha+1} + P_{\text{static}}.
\]

C. WORKLOAD MODEL

In this paper, we adopt the widely accepted assumption that user requests arrive at front portals in a Poisson process [24]. We denote the user request arrival rate at front portal \( m \) as \( \lambda_m \). As soon as user requests arrive at front portals, they will be distributed to processing servers in geo-distributed data centers with a predetermined probability, hence the user request arrival in a server can be also viewed as a Poisson process. Denote \( \lambda_{mn}^s (s \in [1, S_n]) \) as the request’s average arrival rate at server \( s \) in data center \( n \) from front portal \( m \).

In modern data centers, management controllers are in charge of the task dispatching to servers, where a local cache is equipped to buffer the waiting tasks. Without loss of generality, the processing time of a task satisfies a general distribution if the server frequency is given. Thus the dynamic state of a server’s task queue in data center networks can be modeled as an M/G/1 queue which has already been adopted by the previous literature [25, 26]. In addition, the service rate in a DVFS-enabled server is proportional to its frequency, i.e., \( \mu_n^s = r f_n^s \), where \( f_n^s \) is the operating frequency of server \( s \) in data center \( n \) and \( r \) is a scaling factor.

III. PROBLEM FORMULATION

A. THE POWER CONSUMPTION CONSTRAINT

Let \( P_n^s \) denote the power consumption of server \( s \) in data center \( n \). \( P_n^s \equiv 0 \) if a server is deactivated. Otherwise, \( P_n^s \)

When it is activated, i.e.,

\[ Y_n^s = \begin{cases} 1 & \text{if it is activated} \\ 0 & \text{otherwise} \end{cases} \]

is a joint function of processing frequency \( f_n^s \) and usage \( \rho_n^s \) when it is activated, i.e.,

\[ P_n^s = B(f_n^s)^{\alpha+1} \rho_n^s + P_{\text{static}}, \]

where \( \rho_n^s \) is the probability that server \( s \) in data center \( n \) is busy. According to the M/G/1 queuing theory, we have

\[ \rho_n = \frac{\lambda_n^s}{\mu_n^s} \]

where

\[ \lambda_n^s = \sum_{m=1}^{M} \lambda_{mn} \]

is the total requests distributed to server \( s \) in data center \( n \) from \( M \) front portals.

A binary variable \( Y_n^s \in \{0,1\} \) is used to denote whether a server is activated or not, i.e., \( Y_n^s = 1 \) if it is activated and \( Y_n^s = 0 \) otherwise, thus

\[ P_n^s = Y_n^s (B(f_n^s)^{\alpha} \frac{\lambda_n^s}{\mu_n^s} + P_{\text{static}}). \]

### B. THE POWER SUPPLY AND DEMAND BALANCE CONSTRAINT

The power demand \( D_n^s \) of server \( s \) in data center \( n \) should be equal to its power consumption \( P_n^s \) as we assume the power usage effectiveness (PUE) is 1 in ideal status for convenience, i.e.,

\[ D_n^s = P_n^s = Y_n^s (B(f_n^s)^{\alpha} \frac{\lambda_n^s}{\mu_n^s} + P_{\text{static}}). \]

A data center can obtain electricity from both public power grid and its own renewable energy plants. Denote the electricity bought from public power grid in server \( s \) as \( C_n^s \) and REG in data center \( n \) as \( R_n \). In order to satisfy all the user requests, we have the requirement that the electricity supply should not be less than the demand, i.e.,

\[ \sum_{s=1}^{S_n} C_n^s + R_n \geq \sum_{s=1}^{S_n} D_n^s. \]

### C. THE WORKLOAD BALANCE CONSTRAINT

User requests reach the front portal \( m \) at a rate \( \lambda_m \). Then they are allocated to servers for executing. Requests dispatched to data center \( n \) are the summation of requests executed by all servers in it, i.e., \( \sum_{s=1}^{S_n} \lambda_n^s \). Therefore, total requests to be executed in all servers should be equal to those arriving at all front portals, i.e.,

\[ \sum_{m=1}^{M} \lambda_m = \sum_{n=1}^{N} \sum_{s=1}^{S_n} \lambda_n^s. \]

### D. THE QOS CONSTRAINT

As aforementioned, the request processing procedure of server \( s \) in data center \( n \) is regarded as an M/G/1 queuing model with mean arrival and service rate as \( \lambda_n^s \) and \( \mu_n^s \), respectively. In the steady state, the expected delay \( t_n^s \) at each server is the summation of average service time \( \frac{1}{\mu_n^s} \) and average waiting time \( \frac{\lambda_n^s}{(1-\rho_n^s)} \) of the tasks that queue at server \( s \) in data center \( n \) [27], i.e.,

\[ t_n^s = \frac{1}{\mu_n^s} \frac{\lambda_n^s}{(1-\rho_n^s)}. \]

While tasks’ average service time \( \frac{1}{\mu_n^s} \) is equal to their average size \( \frac{1}{\sigma_n^s} \) divided by the server’s computing speed \( c_n^s \), i.e.,

\[ \frac{1}{\mu_n^s} = \frac{1}{\sigma_n^s} = \frac{1}{k \mu_n^s} \]

in which \( c_n^s = k \mu_n^s \), thus \( k = \sigma_n^s = \frac{1}{\mu_n^s \rho_n^s} \).
\( E(s_{\text{sz}}^n) \) is a known factor indicating the expectation of task size. Hence we have
\[
t_n^* = \frac{Y^*}{rf_n^*} + \frac{\lambda_n^* s_{\text{sz}}^n}{2(1 - \frac{\lambda_n^*}{r_n^*})(kr f_n^*)^2},
\]
where \( s_{\text{sz}}^2 = V(s_{\text{sz}}^n) + [E(s_{\text{sz}}^n)]^2. \) \( V(s_{\text{sz}}^n) \) is the variance of task size which is easy to obtain. To guarantee the QoS requirement, delay \( t_n^* \) at any activated server should not violate the maximum delay \( T \) which gives rise to the following constraint
\[
\frac{Y^*}{rf_n^*} + \frac{\lambda_n^* s_{\text{sz}}^n}{2(1 - \frac{\lambda_n^*}{r_n^*})(kr f_n^*)^2} \leq T.
\]

E. AN MINLP FORMULATION

Note that each server in data center \( n \) has the same electricity price \( Pr_n \), the overall electricity cost \( EC \) can be obtained by summing the conventional electricity cost derived from all servers across all geo-distributed data centers, i.e.,
\[
EC = \sum_{n=1}^{N} \sum_{s=1}^{S_n} C_n^s Pr_n.
\]

Taking all constraints presented above, the optimal task placement problem with electricity cost minimization objective is formulated as an MINLP where \( Y_n^*, f_n^*, \lambda_n^* \) and \( C_n^s \) are decision variables. Hence we have
\[
\begin{align*}
\min & \quad \sum_{n=1}^{N} \sum_{s=1}^{S_n} C_n^s Pr_n \\
\text{s.t.} & \quad \sum_{m=1}^{M} \lambda_m = \sum_{n=1}^{N} \sum_{s=1}^{S_n} \lambda_n^s, \\
& \quad \frac{Y_n^*}{rf_n^*} + \frac{\lambda_n^* s_{\text{sz}}^n}{2(1 - \frac{\lambda_n^*}{r_n^*})(kr f_n^*)^2} \leq T, \\
& \quad \sum_{s=1}^{S_n} C_n^s + R_n \geq \sum_{s=1}^{S_n} D_n^s, \\
& \quad D_n^s = Y_n^*(B(f_n^s)^{\alpha} + P_{\text{static}}), \\
& \quad 0 \leq f_n^s \leq f_{\text{max}}, Y_n^* \in \{0, 1\}.
\end{align*}
\]

Due to the high complexity of the problem (time complexity is \( O((2^{\sum_n S_n})^N) \)), the following proposition is proposed to simplify the original MINLP.

Proposition 1. For the given amount of requests, the minimum energy consumption in a data center is attained when user requests are uniformly delivered to all activated servers.

Proof. Note that the energy consumption function (3) is increasing with respect to working frequency \( f_n^* \), it’s reasonable to have
\[
f_n^*(\lambda_n^*) = \frac{k(T \lambda_n^* + 1) - \sqrt{k^2(T \lambda_n^* - 1)^2 + 2T \lambda_n^* s_{\text{sz}}^n}}{2Tkr},
\]
which is derived from inequation (5) for energy saving while ensuring the QoS requirements. Thus, the minimum power demand \( D_n^s \) can be rewritten as a function of \( \lambda_n^s \). Let \( S_n^a \) be the number of activated servers in data center \( n \), the total energy demand can be calculated as \( \sum_{s=1}^{S_n^a} D_n^s(\lambda_n^s) \).

Since \( D_n^s(\lambda_n^s) \) is strictly convex in range \((0, +\infty)\), applying Jensen’s inequality we have \( \sum_{s=1}^{S_n^a} D_n^s(\lambda_n^s) \geq D_n^s(\frac{\sum_{s=1}^{S_n^a} \lambda_n^s}{S_n^a}) \).

\( S_n^a \), where equality holds if and only if \( \lambda_n^1 = \lambda_n^2 = \ldots = \lambda_n^{S_n^a} \).

Therefore, it is proved that the minimum energy consumption of a data center is achieved via uniformly scheduling all requests to all activated servers.

Denote \( h_n = \sum_{s=1}^{S_n^a} \lambda_n^s \) as the requests allocated to data center \( n \). According to Proposition 1, the minimum energy consumption in data center \( n \) for the given \( S_n^a \) is equivalent to
\[
D_n(S_n^a, h_n) = Bf_n^a(S_n^a, h_n)^{\alpha} + h_n + P_{\text{static}}.
\]

where
\[
f_n^a(S_n^a, h_n) = \frac{k(T \cdot h_n + S_n^a)}{2Tkr \cdot S_n^a} - \sqrt{k^2(T \cdot h_n - S_n^a)^2 + 2T \cdot h_n S_n^a s_{\text{sz}}^n}.
\]

Consequently, problem (3)-(8) is reformulated as
\[
\begin{align*}
\min & \quad \sum_{n=1}^{N} \sum_{s=1}^{S_n} C_n^s Pr_n \\
\text{s.t.} & \quad \sum_{m=1}^{M} \lambda_m = \sum_{n=1}^{N} \sum_{s=1}^{S_n} \lambda_n^s, \\
& \quad \frac{Y_n^*}{rf_n^*} + \frac{\lambda_n^* s_{\text{sz}}^n}{2(1 - \frac{\lambda_n^*}{r_n^*})(kr f_n^*)^2} \leq T, \\
& \quad \sum_{s=1}^{S_n} C_n^s + R_n \geq \sum_{s=1}^{S_n} D_n^s, \\
& \quad D_n^s = Y_n^*(B(f_n^s)^{\alpha} + P_{\text{static}}), \\
& \quad 0 \leq f_n^s \leq f_{\text{max}}, Y_n^* \in \{0, 1\}.
\end{align*}
\]

It’s obvious to observe that the complexity of the new formulation is \( O((S_n^a + 1)^N) \), which is much smaller than that of the former one with \( O((2^{\sum_n S_n})^N) \), owing to two new variables introduced which reduce the solution space significantly.

IV. ALGORITHM DESIGN

In this section, we design efficient algorithms to solve problem (11)-(15). Based on the previous analysis, for the given \( S_n^a \), problem (11)-(15) is convex and can be solved in polynomial time. Under such circumstance, a double-case algorithm will be designed to solve it. The corresponding solution \( SAC = [S_1^a, S_2^a, ..., S_n^a] \) is named as the server activation configuration. According to the system design, the \( SAC \) is composed of two parts as shown in Fig. 2: \( SAC_{RE} \) representing the number of activated servers powered by renewable energy (RE) and \( SAC_{CE} \) representing the number of activated servers powered by conventional energy (CE).

The optimal \( SAC \) and task placement strategy is shown in Algorithm 1 aiming at solving problem (11)-(15) which further includes two cases. In Algorithm 1, all data centers are
Algorithm 1 Optimal SAC and Task Placement Strategy

Input: \( M, N, \lambda_m, P_r, S_n, f_{\min}, f_{\max} \)

Output: \( SAC, EC \)

1. Sort data centers in descending order of REG
2. for \( n = 1 \) to \( N \) do
3. \[ \{ S_n^{RE, max}, h_n^{RE, max} \} = \arg \max h_n^{RE} \text{ s.t. } D_n(S_n^{RE, max}, h_n^{RE, max}) = R_n, 0 \leq S_n^{RE, max} \leq S_n, 0 \leq h_n^{RE} \leq \sum_{m=1}^{M} \lambda_m, f_{\min} \leq f_n \leq f_{\max} \]
4. end for
5. if \( \sum_{m=1}^{M} \lambda_m \leq \sum_{n=1}^{N} h_n^{RE, max} \) then
6. Case I: Data centers are powered by RE.
7. else
8. Case II: Data centers are powered by RE and CE.
9. end if

In Case I (Algorithm 2), we compare all user requests \( \sum_{m=1}^{M} \lambda_m \) with total maximum requests \( h_{\text{total}} \) that data centers can process from one with the most REG, then the optimal \( SAC \) set and minimum number of activated data centers \( N_{\text{min}} \) is obtained.

In Case II (Algorithm 3), some data centers have to be powered by conventional energy to process the user requests. First, all data centers are sorted in the ascending order by the electricity prices and we obtain initial \( SAC \) by activating the remaining servers from data center with the lowest electricity price, and the computation capacity \( h_n^{CE, min} \) of each data center with minimum energy consumption can be calculated in line 5. Then we compare the remaining user requests \( h_r \) with each data center’s minimum computation capacity \( h_n^{CE, min} \). If the former is smaller than the latter, the remaining user requests can be processed by this data center and the maximum number of activated data centers \( N_{\text{max}} \) is calculated in lines 6-9. Otherwise, more data centers must be utilized. Keep doing so until all user requests are satisfied and the initial \( SAC \) is attained. For the next step, we iteratively update \( SAC \) by deactivating servers from data centers with high electricity price to minimize the total electricity cost. First, with initial \( SAC \) derived aforementioned, an initial computation capacity \( h_n \) and electricity cost \( EC \) can be calculated by solving problem (11)-(15). Then we reduce the number of activated servers from the most expensive data center with a bisection method as shown in lines 14-26. For each new \( SAC' \), its corresponding minimum electricity cost \( EC' \) is achieved. If there is a comparatively lower electricity cost, we set it as new \( EC \) and seek to deactivate more servers in this data center as indicated in lines 19-21. Once the new cost \( EC \) exceeds the temporarily optimal cost, no more servers shall be deactivated thus we should recover some
overly deactivated servers in lines 22-24. Finally, with algorithm converging, the optimal \( SAC \) setting and the minimum electricity cost \( EC \) is attained and the problem (11)-(15) is solved.

Define the maximum server number in each data center as \( MS \), then the complexity of our algorithms is \( O(N\log MS) \) which mainly arises in Case II (Algorithm 3), where we iteratively deactivate data centers via a bisection method in lines 14-26. As to a data center \( n \) with \( S_n^a \) activated servers in the initial \( SAC \), \( \log_2 S_n^a \) iterations are implemented. When it comes to the worst case, all data centers are checked, leading to \( \sum_{n=1}^{N} \log_2 S_n^a = O(N\log MS) \) iterations in all. Most importantly, problem (11)-(15) is a convex problem given a \( SAC \), therefore we can get a global optimal solution in polynomial time.

V. EVALUATIONS AND ANALYSES

In this section, we perform our evaluations in MATLAB on an Intel workstation with 4 processors clocking at 3.3 GHZ and 8 GB of RAM. An optimization toolbox \( fmincon \) is utilized to solve (11)-(15) given a \( SAC \).

A. EVALUATION SETTINGS

Evaluation parameters involved are shown in Table 1. As a server’s power level is usually hundreds of Watts, the scaling factor \( r \) and coefficient \( B \) can be calculated accordingly.

**TABLE 1: Parameters involved in evaluations**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of front portals</td>
<td>4 – 8</td>
</tr>
<tr>
<td>Number of user requests</td>
<td>80 – 100000</td>
</tr>
<tr>
<td>Number of data centers</td>
<td>3 – 24</td>
</tr>
<tr>
<td>Number of servers</td>
<td>(3 – 9) ( \times ) ( 10^4 )</td>
</tr>
<tr>
<td>Server delay (s) [17]</td>
<td>0.01</td>
</tr>
<tr>
<td>Server operating frequency set (GHz)</td>
<td>[0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0, 2.2, 2.4, 2.6, 2.8]</td>
</tr>
<tr>
<td>Scaling factor ( r )</td>
<td>( 6 \times 10^{-10} )</td>
</tr>
<tr>
<td>Coefficient ( B )</td>
<td>( 1.8 \times 10^{-26} )</td>
</tr>
<tr>
<td>Proportion of static power consumption</td>
<td>0.1 – 0.6</td>
</tr>
<tr>
<td>Size of user requests (MI)</td>
<td>4000 – 10000</td>
</tr>
<tr>
<td>Total request rate (units/s)</td>
<td>( (0.7 – 1.3) \times 10^5 )</td>
</tr>
</tbody>
</table>

Unless stated otherwise, the other basic parameters are set as follows:

1) Data center parameters: Three data centers located in Brussels, Flanders, Limburg and four front portals in Belgium are tested. Server number of these three data centers are 15000, 30000 and 10000, respectively.

2) REG parameters: We assume there are solar panels installed in geo-distributed data centers. All REG traces are based on the real-world data [28] of the real-time estimations of actual solar-PV generation of Brussels, Flanders and Limburg in August, 2017.

3) Electricity price: Electricity price traces are obtained from NYiso [29]. We adopt the locational based marginal electricity prices of central New York Control Area on August 15th, 2017.

(4) User request information: User requests are considered to arrive at front portals in Poisson distribution. The average request rates at four portals are 30000, 15000, 15000 and 20000 units/s, respectively. Request size is uniformly varied between 4000 MI and 10000 MI (million instructions).

B. RESULTS AND DISCUSSIONS

In this section, we conduct experiments to evaluate the impacts of different system parameters (static power consumption, server number, user requests, etc.) on electricity cost and server frequencies, and assess the performance of the proposed optimal task placement strategy.

1) The Impact of Static Power Consumption

In some literature, the static power consumption is usually ignored [20, 30]. However, the real fact is that static power consumption has a strong impact on data center management and the electricity cost. In this test, the proportion of static power consumption is set from 0.1 to 0.6. We select REG at 12:00 as 1150.25, 1290.1, 1290.99 MW/h and electricity price as 42.93, 20.27, 55.30 $/MW for the three data centers, respectively. Results concerning the impact of static power consumption on electricity cost corresponding to data centers with and without solar panels are depicted in Fig. 3. It’s observed that electricity cost rises when static power consumption increases, thus we should think carefully whether it is necessary to activate more servers for more requests. Furthermore, data centers installed solar panels can save electricity bills significantly.

![Electricity cost with respect to different static power consumption](image)

**FIGURE 3:** Electricity cost with respect to different static power consumption

2) The Impact of REG and Electricity Price

Based on the proposed strategy, user requests are more prone to be processed by data centers with sufficient
REG and low electricity price. To better present the results, we consider 6 data centers in this test with user requests arriving rates as 30000, 30000, 30000, 20000 respectively at 4 front portals, and server numbers are 15000, 30000, 10000, 20000, 5000, 25000 for 6 data centers, respectively. Unless explicitly specified, the proportion of static power consumption is set as 0.2. We select electricity price at 12:00 as 42.93, 20.27, 55.30, 30.93, 25.27, 50.30 $/MW and REG as 341.4, 1290.1, 83.5, 634.7, 322.9, 923.8 MW/h for the 6 data centers, respectively. Results concerning the impact of REG and electricity price on server activation configuration and electricity cost are shown in Figs. 4 and 5 respectively. In Fig. 4, it is clear that more servers are activated in data centers with large REG, thus they are enabled to process more user requests with lower costs. In Fig. 5, user requests are mostly dispatched to data centers with small electricity price for executing. However, as indicated by the blue bars, lower electricity price together with heavier workload doesn’t definitely lead to low electricity cost, considering the variation of cost savings from the renewable energy.

3) The Impact of Active Server Number

In this test, we evaluate the impact of active server numbers on the electricity cost under different proportions of static power consumption. The active server number varies from 30000 to 90000 and the proportions of static power consumption is set as 0.2 and 0.5 respectively. Other parameters are kept the same as the settings in V-B1. Results corresponding to electricity cost of data centers with and without solar panels are depicted in Fig. 6. When the proportion of static power consumption is 0.2, it is shown that electricity cost decreases first and then increases when the active server number grows. This is because when user requests are distributed to more servers, servers on each data centers process less requests with lower frequencies which leads to lower electricity cost. However, when the active server number increases over an inflection point (80000 in this experiment), more activated servers consume more static power consumption, and the server frequencies are far away from the economic interval, thus leading to the increment of electricity cost. This phenomenon can also be seen when the static power proportion of servers is set as 0.5, where the optimal active server number is around 45000. The results also verify that there exists an optimal server activation configuration point to process all the user requests before the deadline.
4) Impact of Total User Requests
In this test, we evaluate the impact of total user requests on the electricity cost. The results corresponding to data centers with and without solar panels are shown in Fig. 7. It is obvious to observe that electricity cost increases when total user requests expand. In addition, the electricity cost are more sensitive to the user requests when they are at high levels. This is because the power consumption of servers increases faster when servers run at higher frequencies.

5) The Variation of SAC, Requests Processed and Frequency Traces in 24 Hours
In this test, we evaluate the proposed optimal task placement strategy during 24 hours in a day adopting real REG and electricity price traces mentioned in Section V-A. Results concerning electricity cost, SAC and number of requests processed during 24 hours with different REG and electricity prices at each time slot are depicted in Fig. 8. In Fig. 8 (a), the trend of electricity cost follows the trend of electricity price except in time slots 11 – 18 when electricity price is relatively high and renewable energy is sufficient to process user requests. Due to the fact that REG tends to be high in time slots 7 – 21, all renewable energy is utilized to process user requests as shown in Fig. 8 (b). Then the remaining user requests are dispatched to time slots with low electricity prices, such as time slots 1 – 6, 7 – 15 and 20 – 24 to ensure that tasks can be processed within their deadlines. In addition, when electricity price is high, the number of servers activated by conventional energy is small in time slots 16 – 18, while large in the remaining time slots. From Fig. 9, it’s obvious to observe that the frequency trace has similar trend with the electricity cost trace, from which we may come to the conclusion that servers indeed work at high frequencies in dealing with large user requests owing to DVFS technique we utilized. Such observation further proves that our optimal task placement strategy based on DVFS technique can decrease the electricity cost in a multi-electricity market.

VI. CONCLUSIONS
In this paper, we focus on the electricity cost minimization problem in data centers powered by mixed renewable and conventional energies. An optimal task placement strategy problem considering the spatial and temporal diversity of REG and electricity price, as well as DVFS technique and M/G/1 queuing model is first formulated as a complicated MINLP. Afterwards, we transform the original MINLP into a more easier-solving form. In this strategy, user requests are distributed to data centers with high REG first, then the remaining user requests are dispatched to those with low electricity price for economic operation. The optimal SAC and task placement algorithm is proposed to obtain a global optimal solution. Numerical evaluations based on real-world traces have tested the impacts of different system parameters on electricity cost and server activation configurations.
Simulation results show that user requests arriving at front portals are more likely to be processed in abundant REG and low electricity price data centers and the idea of importing portals are more likely to be processed in abundant REG and the idea of importing forms are more likely to be processed in abundant REG and the idea of importing forms are more likely to be processed in abundant REG.

**ACKNOWLEDGEMENT**

We would like to thank the Foundation of Graduate Innovation Center in NUAA (with grant KFJJ20171607) for providing us the testing environment.

---

**REFERENCES**


DR. RAN WANG (M’17) is currently an assistant professor at College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics (NUAA), and Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, P.R. China. He received his B.E. in Electronic and Information Engineering from Harbin Institute of Technology (HIT), P.R. China in July 2011 and Ph.D. in Computer Science and Engineering from Nanyang Technological University (NTU), Singapore in April 2016. He was a research fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore from October 2015 to August 2016. He has authored or coauthored over 30 papers in top-tier journals and conferences. He received the Nanyang Engineering Doctoral Scholarship (NEDS) Award, Singapore and innovative and entrepreneurial Ph.D. Award of Jiangsu Province, China in 2011 and 2017, respectively. His current research interests include intelligent management and control in smart grid, network performance analysis and evolution of complex networks, etc.

DR. KUN ZHU (M’15) is currently a Professor in the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China. He received his Ph.D. degree in 2012 from School of Computer Engineering, Nanyang Technological University, Singapore. He was a research fellow with the Wireless Communications, Networks, and Services Research Group in University of Manitoba, Canada. His research interests include resource allocation in 5G, wireless virtualization, and self-organizing networks. He has served as TPC for several conferences and reviewer for several journals.

DR. JIE HAO (M’16) received her BS degree from Beijing University of Posts and Telecommunications, China, in 2007, and the Ph.D. degree from University of Chinese Academy of Sciences, China, in 2014. From 2014 to 2015, she has worked as post-doctoral research fellow in the School of Computer Engineering, Nanyang Technological University, Singapore. She is currently an Assistant Professor at College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China. Her research interests are wireless sensing, visible light communication and etc.

DR. PING WANG (M’08, SM’15) received the Ph.D. degree in electrical engineering from the University of Waterloo, Canada, in 2008. She is currently an Associate Professor with the Department of Electrical Engineering and Computer Science, York University, Canada. Before that, she was with Nanyang Technological University, Singapore. Her current research interests include resource allocation in multimedia wireless networks, cloud computing, and smart grid. She was a co-recipient of the Best Paper Award from the IEEE International Conference on Communications in 2007 and the IEEE Wireless Communications and Networking Conference in 2012. She has been serving as an Associate Editor for several journals including the IEEE Transactions on Wireless Communications, the EURASIP Journal on Wireless Communications and Networking, and the International Journal of Ultra Wideband Communications and Systems.

DR. YUE CAO (M’16) received the Ph.D. degree from the Institute for Communication Systems (ICS), 5G Innovation Centre (5GIC), at University of Surrey, Guildford, UK in 2013. He was a Research Fellow at the ICS until September 2016, and Lecturer in Department of Computer and Information Sciences, at Northumbria University, Newcastle upon Tyne, UK until July 2017, and currently the Senior Lecturer since August 2017. His research interests focus on Intelligent Mobility. He is the Associate Editor of IEEE Access and International Journal of Vehicular Telematics and Infotainment Systems.

MS. YIWEN LU is currently pursuing her Master Degree in Software Engineering at College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics (NUAA), P.R. China. She received her B.E. Degree in network engineering from Qingdao University in June 2016, during which she received various national scholarships. Her current research interest includes intelligent management and control in smart grid and data centers.

VOLUME *, *

11

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2876361, IEEE Access