Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue11 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

On Enabling Energy-efficiency in Geographically Distributed Data Centers through Data Deduplication

Muhammad Imran Khan Khalil
Department of Computer Science, University of Engineering & Technology Peshawar 25000, Pakistan

Fawad Ali
Department of Computer Science, University of Engineering & Technology Peshawar 25000, Pakistan

Syed Adeel Ali Shah
Department of Computer Science, University of Engineering & Technology Peshawar 25000, Pakistan Email: adeel@uetpeshawar.edu.pk

Izaz Ahmad Khan
Department of Computer Science, Bacha Khan University, Charsadda 24420, Pakistan

Alauddin
Department of Computer Science, University of Engineering & Technology Peshawar 25000, Pakistan

Amer Taj
Department of Computer Science, University of Engineering & Technology Peshawar 25000, Pakistan

ABSTRACT

Cloud Service Providers (CSPs) are increasingly adopting geographically dispersed data centers (DCs) to cater to growing user demands while minimizing latency. However, this distributed architecture poses significant energy consumption challenges. While research on cost-reduction techniques for geographically distributed DCs exists, a clear gap remains in exploring cost savings during workload processing, particularly when incorporating data deduplication techniques. This paper proposes Data Deduplication Geographical Load Balancing (DDGLB), a novel approach designed to reduce operating costs in geographically distributed data centers. DDGLB leverages a data deduplication algorithm, such as Knuth-Morris-Pratt (KMP), to compare incoming workloads with those previously processed within a specific DC. If a match is found, the global load balancer (GLB) refrains from forwarding the duplicate workload for processing. Instead, it returns the pre-computed results, achieving significant energy savings. Conversely, if no match is found, the

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue11 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

GLB allocates the new user request to a specific DC for processing. This allocation considers various optimization parameters, including the availability of historical workload traces, on-site renewable energy sources, dynamic energy prices, and the number of active servers at each location. Through comprehensive numerical evaluations, this study demonstrates that DDGLB can significantly reduce operating costs for geographically distributed DCs compared to existing user request distribution strategies.

Keywords: Data Deduplication; Energy Efficiency; Geographical Load Balancing; Green Computing; Energy Efficient Algorithms

Introduction

The widespread adoption of cloud computing services has significantly increased energy consumption in Information and Communication Technology (ICT) systems. Data centers (DCs), being significant energy consumers, play a crucial role in providing open-access interfaces for extensive data analysis [1,2]. The growing demand for cloud computing has led to expanding data centers, resulting in a substantial increase in power requirements to process user requests. Recent studies project a fourfold increase in annual data production by 2024, accompanied by a nearly 10% rise in the power consumption of cloud DCs over the past decade [3,4]. This surge has prompted rapid increases in energy prices to cover the costs of powering and cooling computing resources, emphasizing the need for designing energy-efficient and environmentally friendly workload allocation systems [5].

Cloud computing providers leverage geographical load balancing (GLB) to distribute workloads across geo-distributed data centers, reducing operating costs [6]. GLB considers regional variations in energy prices, contributing to the cost-effectiveness of data centers [7]. By utilizing the flexibility of geographically distributed data centers, GLB directs incoming workloads to areas featuring renewable energy and facilitates the seamless integration of green energy sources [8].

Recent research has focused on using green energy to power data centers, allowing them to reduce reliance on traditional power plants and decrease overall costs and environmental impact [7,8]. While renewable power sources like wind and solar can significantly diminish dependence on fossil fuels, they inherently exhibit less reliability than traditional grid electricity [9]. Green energy production is subject to variations based on the time of day, season, weather, and other factors. Efficiently utilizing green energy requires predicting production and balancing it with the workload in the data center [10].

Both academic researchers and cloud industry professionals have proposed various techniques to minimize energy consumption while processing incoming user requests, ensuring a high-quality service. These workload distribution strategies encompass data deduplication, energy derivatives, consolidation, and virtualization [7-10]. This research investigates how Cloud Service Providers (CSPs) managing multiple data centers in various locations can allocate incoming user requests and minimize energy costs. In contrast to prior work [1], [8-11], we formulate the problem of reducing energy costs in data centers as an integer linear optimization problem, considering

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue11 (2025)
Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

factors such as data deduplication algorithms, dynamic energy prices, and renewable energy sources.

Data deduplication has revolutionized cloud computing and is crucial in addressing various real-world challenges. It supports completing time-efficient tasks across multiple research areas and is particularly useful for identifying specific patterns within strings [12]. Real-world applications of data deduplication include plagiarism detection, intrusion detection systems, search engines, spam filters, digital forensics, DNA sequencing, spell checks, and digital bioinformatics. This research introduces an eco-friendly architecture for geographically distributed data centers [13]. The global load balancer directs incoming user requests to the most energy-efficient data center. Our primary objective is to leverage a data deduplication algorithm (e.g., KMP) to compare incoming workloads with previously processed documents in a DC. If a match is identified, the GLB avoids forwarding the incoming workload to a DC, presenting the user with results from the prior workload instead of conserving energy. In cases without a match, the GLB assigns the incoming user requests to a specific data center, considering factors such as expected delay, server utilization levels, the availability of green energy, and dynamic electricity prices.

We formulate the problem of reducing operating costs as an integer linear optimization program and devise an online algorithm employing a greedy approach. Extensive experiments using real-world data validate the efficiency of our proposed workload distribution strategy. In summary, the contributions of this research work are outlined as follows:

By focusing on a unified framework for geographic workload balancing in DC, we consider a systemic approach, integrating a range of workloads and energy efficiency factors into an in-depth model than that used in cutting-edge techniques. Compared to [3-11], have solved the load balancing problem by considering interactive workload and energy efficiency, which poses a challenge for developing energy efficient distributed workload strategies.

We decouple the discrete-time optimization variables and reformulate the dynamic problem as an integer linear optimization program. Using data deduplication technique, we develop a novel online algorithm called DDGLB to reduce DC operating costs while integrating renewable energy sources considering interactive loads. In particular, DDGLB uses a greedy online algorithm strategy and does not rely on future information such as dynamic energy prices or incoming requests from users.

Although most existing research works in this problem setting (e.g. [3], [4], [6], and [9]) incorporate energy efficient workload distribution considering independent and identical workload distribution strategy. Assuming the underlying system dynamics, we analytically determine the feasibility and optimality of DDGLB under a more general assumption that is easily adapted to the above problem setting and can support a broad scope of real-world user requests. We evaluate DDGLB using benchmark algorithms and demonstrate its usability and performance through a rigorous numerical evaluation using real-world data, including green power availability, average latency, current server utilization level, and incoming workload.

The rest of this research paper is structured as follows: Section 2 reviews the literature on GLB using data deduplication techniques. Section 3 formulates the GLB problem as an optimization problem, and Section 4 describes the problem setting. Section 5

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue11 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

proposes an online algorithm called DDGLB to solve the problem using a greedy approach. Section 6 evaluates the performance of DDGLB on real-world datasets. Section 7 concludes the paper.

Related Work

We organize the related work according to the following factors:

Energy Efficiency and Data Deduplication

A recent survey conducted by cloud service providers (CSPs) reveals that around 80% of cloud industries are considering the adoption of data deduplication and string-matching techniques to minimize duplicate data and enhance storage efficiency within their storage systems [10], [12-16]. Data deduplication, eliminating identical copies of the same data to save storage space, is coupled with versatile string-matching approaches used across various computing fields [13], [17]. CSPs focus on detecting patterns within incoming user requests, and data deduplication algorithms play a vital role in searching for unnecessary patterns in the repository after pattern recognition by an optical character recognition system [15].

This study aims to expand existing data deduplication algorithms, explicitly focusing on minimizing energy costs during workload processing in geographically dispersed data centers. Various string-matching techniques have been evaluated, considering time complexity and pattern length [10]. Experimental assessments of online approximate algorithms for duplicate data have been conducted to reduce energy costs during user request execution [12]. A workload distribution algorithm that relies on data deduplication strategies within geographically distributed data centers has also been introduced [13]. While past studies concentrated on enhancing data storage efficiency, this study explores the novel problem of reducing operating costs in data centers during workload execution using data deduplication techniques.

Data Centers with Renewable Energy

Cloud service providers are expected to minimize energy utilization in DCs while ensuring the highest possible efficiency [4]. Opting for a new-generation green data center architecture represents a valuable long-term investment with economic and environmental advantages. Green data centers leveraging renewable energy substantially reduce electricity and fuel expenses and adopt measures to recycle electronic waste generated within the data center [9].

Renewable-powered data centers have garnered increased interest from both industry and academia [6], [8,9]. Previous research has explored the feasibility and benefits of using geographical load balancing for delay-sensitive workload execution with the integration of green power [5], [18–23]. While prototypes for workload distribution have showcased load balancer efficiency, investigations into implementing a delay-sensitive load balancer considering green power sources have primarily focused on interactive workloads without energy storage devices (ESDs) and future data [5], [18–23]. In contrast, our method is online, operating without advanced information, and can manage interactive workloads alongside renewable energy sources in the data center.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue11 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Energy Efficiency and GLB

Reducing operating costs for geo-distributed data centers is a critical area of current research [22], [24–32], [41–46]. This study focuses on optimizing operating costs in geographically distributed data centers while processing user requests. A comprehensive review of Global Load Balancing (GLB) for optimizing power utilization in data centers has been presented [24]. The "follow the renewables" GLB approach, requiring dynamic workload distribution in the presence of renewable power in data centers, has been discussed [25]. Investigations have been initiated into geographical load balancing and its influence on non-renewable energy (brown energy) consumption [26]. Algorithms enhancing the utilization of green power supply through geographical load balancing have been proposed and evaluated using numerical simulations based on real-world traces [26].

Further extending this research, online algorithms designed to optimize decentralized renewable energy sources over an extensive geographical location have been suggested [27]. Research illustrating methods to enhance the efficiency of incorporating renewable energy sources like solar and wind power has been presented [27]. A method for planning and organizing workloads in data centers to enhance sustainability, considering the costs associated with carbon emissions, has been proposed [28]. Proposing an algorithm for workload distribution in data centers considering cooling dynamics, ambient temperature, green energy variability, and workload deadlines has been initiated [22].

All preceding studies on geographically distributed data centers (DCs) have assumed that the centers derive energy exclusively from environmentally sustainable and renewable sources. This study makes the same assumption and proposes a novel data deduplication technique (DDGLB) based on an online greedy approach. DDGLB demonstrates energy efficiency, surpassing the cost-effectiveness of current workload distribution methods in dynamic system conditions, including fluctuating factors such as incoming user requests, the availability of renewable energy, the number of active servers, and dynamic electricity prices. The practical viability of our proposed system is emphasized by subjecting it to validation using renewable energy, incoming user requests, and dynamic electricity prices.

Research Methodology

This section outlines the research problem and presents the system model for distributing incoming workloads among geographically distributed data centers.

Problem Formulation

In this research, we consider a set of geographically distributed data centers, denoted by $i \in N$, each equipped with multiple homogeneous servers. Each data center denoted by i is powered by a combination of conventional brown energy sources, such as oil, coal, or gas, and sustainable green power sources.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue11 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

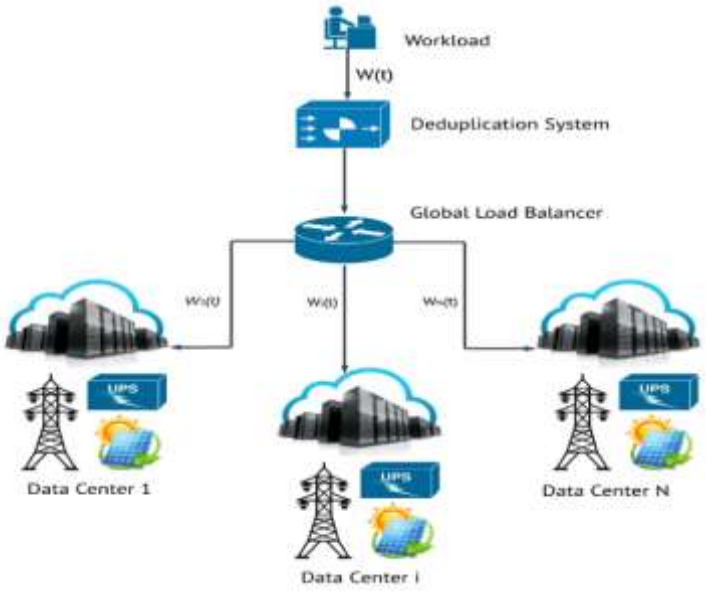


Figure 1. Workload processing model

When a workload is submitted to a data center i, the hash value is cross-referenced with the hash table to determine if it has been previously executed. If identified as a new workload, GSLB forwards it to the optimal data centers for further processing and maintains a record for future reference. In the case of a previously executed workload, the deduplication system identifies it and delivers the results to the end user without repeating the execution. GSLB ensures that users are directed to the optimal server or data center for accessing the required content or service, regardless of location. This optimization aims to enhance user experience by minimizing delay and ensuring the quality of service. The online routing decisions made by GSLB are designed to minimize power consumption and workload average latency and reduce the total operating costs of the data center. Please refer to Figure 1 for a visual representation.

Workload Model

In this study, the role of λi is to validate the downloadability of incoming workloads. If a user attempts to download the same file from multiple regions within a defined time slot, both requests are exempted due to generating the same hash value. Excluding such requests is crucial; otherwise, the deduplication system would treat the second request as a duplicate workload, resulting in its dismissal. Consequently, the workload model is designed to ignore duplicated workloads for downloading again. The incoming workload allocated to data center i at time t is denoted by $w_i(t)$. The following relationship is then derived:

$$\sum_{i=i}^{N} \lambda_i \, w_i(t) = W(t) \quad (1)$$

$$\sum_{i=i}^{N} \lambda_i = 1 \text{ where } \lambda_i \in [0, 1] \quad (2)$$

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue11 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

The variable λi signifies the downloadability factor for the incoming workload, assuming either 0 or 1 values. The constraints specified in equations (1) and (2) ensure that every workload W(t) is not meant for downloading; the system will disregard it under such circumstances. In each data center, the total number of available servers is denoted by S_i^{max}, and this value cannot exceed the maximum count of active servers, represented by S_i^{ac}. In other words, S_i^{max} must always be greater than or equal to S_i^{ac}, as indicated in equation (3).

$$0 \leq S_i^{ac}(t) \leq S_i^{max} \quad (3)$$

The Model of Renewable Energy Generation

In data centers, energy primarily originates from conventional sources and on-site green energy, exemplified by solar panels. This study delves into the application of solar panels for renewable energy production. The variable $R_i(t)$ represents the current renewable energy level in data center i at time t.

$$0 \leq R_i(t) \leq R_i^{max} \quad (4)$$

The generated quantity of renewable energy, $R_i(t)$, is non-negative and must consistently adhere to the constraint outlined in equation (4), ensuring it remains below or equal to the designated maximum capacity R_i^{max}. This constraint is crucial to enforce sustainable and optimal energy utilization within the data center, preventing the energy level from surpassing the system's maximum capacity.

The Model for Quality of Service (Delay)

When assessing system performance, it is vital to account for network and queuing delays [35]. Network delay, denoted as $L_i(t)$, represents a constant delay encountered by incoming workloads from the global load balancer upon reaching the data center at time t. Queuing delay is formulated using a queuing system. The mean waiting time, denoted as $d_i^Q(t)$, is determined by $\frac{1}{\mu_i S_i^{ac}(t) - w_i(t)}$. This study assumes that servers in active mode in DCs consistently execute incoming user requests, equivalent to 1 [36]. Additionally, servers within a data center are presumed to be homogeneous, featuring a uniform service rate μ_i [37]. Therefore, the average delay experienced by workloads at data center DC can be calculated using the following formula:

$$d_i(t) = L_i(t) + \frac{1}{\mu_i S_i^{ac}(t) - w_i(t)}$$

where $\mu_i S_i^{ac}(t) > w_i(t)$. At time t, the delay cost for DC i is provided by:

$$d_i(t) = \beta\left[L_i(t) + \frac{1}{\mu_i S_i^{ac}(t) - w_i(t)}\right]$$

A delay cost model is utilized to quantify delay performance in monetary terms. Moreover, the cost per unit delay value, denoted as beta, is assumed to remain constant [38,39]. Ensuring the satisfaction of end users' delay requirements necessitates imposing a constraint to prevent the average delay from exceeding the maximum allowable delay. This constraint is vital for maintaining the system's performance within acceptable limits and meeting necessary specifications. Implementing this restriction enables the control and regulation of workload delay, ensuring it stays within the specified acceptable range, as outlined below:

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue11 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

$$0 \leq d_i(t) \leq D^{max} \quad (5)$$

The Model for Power Consumption

The two primary sources of power consumption in data centers (DCs) are IT equipment and cooling systems. Cooling systems are essential for maintaining DCs at optimal temperatures for effective server performance [40]. Additionally, they contribute significantly to the overall power consumption to meet their cooling requirements. The power consumption of cooling systems is assessed about the total power used in a DC, incorporating the power usage effectiveness (PUE) measure [41]. PUE compares the total incoming power of a DC, including cooling systems and other computing devices, with the power used exclusively by IT equipment. This measure offers crucial insights into the effectiveness of the cooling system and the overall power consumption of the DC. By regularly monitoring and optimizing PUE, DC operators can enhance energy efficiency, reduce environmental impact, and cut operating expenses by minimizing unnecessary power usage in cooling systems. As per [42], the industry's average PUE is approximately 1.5. The power usage of IT equipment in a DC at a given time can be expressed as follows:

$$P_i^{IT}(t) = S_i^{ac}(t)\left[P_i^{in}(t) + P_i^{ac}(t)\, U_i(t)\right]$$

The average server utilization in DC i at time t, denoted by $U_i(t)$, is calculated as the ratio of the workload $w_i(t)$ to the product of the number of active servers $S_i^{ac}(t)$ and the uniform service rate μ_i. One method to represent the total power usage in DC i at time t is as follows:

$$P_i(t) = PUE_i(t)\,.S_i^{ac}(t)\left[P_i^{in}(t) + P_i^{ac}(t)\frac{w_i(t)}{S_i^{ac}(t)\mu_i}\right]$$

We considered that data centers generate renewable energy, and the significance of the capital cost in the overall cost structure of renewable power generation is emphasized. In this context, the marginal cost of renewable power generation is considered negligible. This viewpoint shifts focus away from ongoing operating costs and onto the initial outlay required to set up and maintain the infrastructure for producing renewable energy. We utilize a weighted average approach to simultaneously depict the electricity cost across all data centers. This approach considers the electricity cost linked to renewable and non-renewable power sources.

$$C(t) = \sum_{i=1}^{N} q_i(t)\,[P_i(t) - R_i(t)]$$

where $P_i(t)$ signifies the quantity of electricity acquired at time t and datacenter i, while $R_i(t)$ denotes the renewable energy produced by data center DC at time t.

Geographical Load Balancing Optimization Problem

The primary objective of GLB is to minimize the operating cost incurred by each DC at each time step by strategically selecting geographically dispersed DCs based on their power consumption, delay performance, and renewable energy generation capabilities. The optimization problem is formulated as a minimization challenge concerning the combined electricity costs and delay across all data centers and time intervals.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue11 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

$$\text{Minimize} \quad \sum_{t=1}^{T}\sum_{i=1}^{N} \lambda_i \left[C_i(t) + d_i(t)\right] \quad (6)$$

Subject to;

$$\sum_{i=1}^{N} \lambda_i w_i(t) = W(t) \qquad \forall t \in [1,T] \quad (6a)$$

$$\sum_{i=1}^{N} \lambda_i = 1 \quad \lambda_i \in [0,1] \qquad \forall t \in [1,T] \quad (6b)$$

$$0 \le S_i^{ac}(t) \le S_i^{max}(t) \qquad \forall i \in [1,N] \quad (6c)$$

$$0 \le d_i(t) \le D^{max} \qquad \forall i \in [1,N] \quad (6d)$$

$$0 \le R_i(t) \le R_i^{max} \qquad \forall i \in [1,N] \quad (6e)$$

Constraints (6a) and (6b) guarantee that $w_i(t)$ is allocated to a data center i at the time t, (6c) restrict the total servers in active mood within a data center, (6d) enforce a delay constraint, and (6e) ensure that the quantity of renewable energy sources remains non-negative and within an upper bound.

Solution of Optimization Problem

This section provides a solution to the optimization problem introduced in Section B. The primary objective of this optimization problem is to optimally select the most appropriate DC at each time interval, considering average delay and energy utilization, to minimize the overall cost in a geo-distributed data center. The optimal solution to this optimization problem employs the mixed-integer linear programming (MILP) technique, particularly well-suited for modeling and resolving data center selection problems in this context [36]. The objective function and constraints can be explicitly expressed mathematically by formulating the optimization problem as a MILP. This approach allows us to identify the optimal data center selection criteria, considering energy consumption and average delay across various periods while minimizing total expenditures. The optimization problem is reformulated to account for the delay constraint and determine the number of active servers.

$$\mu_i S_i^{ac}(t) - w_i(t) \ge \frac{1}{D^{max} - L_i(t)} \quad \forall i \in [1,N]$$

By arranging;

$$S_i^{ac}(t) \ge \frac{w_i(t)}{\mu_i} + \frac{1}{\mu_i[D^{max} - L_i(t)]} \qquad \forall i \in [1,N] \quad (7)$$

The optimization problem in (6) can be rearranged as follows:

$$\text{Minimize} \sum_{t=1}^{T}\sum_{i=1}^{N} \lambda_i \left[C_i(t) + d_i(t)\right]$$

Subject to (6a), (6b), (6c), (6d), (6e), and (7)

Initially, we address the optimization problem (6) and (6a)–(6e) to ensure that each incoming workload is allocated to a single data center. Subsequently, the number of servers essential for the operation of the DC is determined utilizing equation (7). A

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue11 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

branch-and-bound approach, elaborated in [36], can be applied to pinpoint the optimal solution for the optimization problem.

Data Deduplication Geographical Load Balancing (DDGLB)

The optimization problem outlined in Algorithm 1 faces two inherent obstacles. Firstly, integrating future data, such as electricity prices and user demands, into the optimization process for an optimal solution is impractical [36]. Secondly, even with access to future data, determining the optimal solution for numerous periods and data centers becomes computationally infeasible. For instance, a brute-force approach necessitates an overwhelming number of computations, and the branch-and-bound strategy must navigate through an excessive number of nodes, rendering both methods impractical for larger values of T and N [36], [41]. Consequently, achieving an ideal solution is not viable.

To overcome these limitations, we introduce the Data Deduplication Geographical Load Balancing (DDGLB) algorithm, a technique based on a greedy algorithm. The fundamental concept of our proposed algorithm is to dynamically distribute the incoming workload at each time step, considering current electricity prices, available renewable energy, and existing constraints.

Algorithm Design

We present an online algorithm based on the DDGLB approach, as outlined in this section, to tackle the optimization challenge discussed earlier. The DDGLB approach is elucidated through Algorithm 1, which utilizes an online greedy algorithm. The aim is to minimize the operational costs of the data center in each time slot, denoted as t.

Algorithm 1: DDGLB – Minimization of Operating Cost

1. **for** $t \in [1, T]$ **do**
2. Observe the incoming workload $W_{in}(t)$ for time t □ i $\in [1 - N]$

Part I: Checking for Data Deduplication

3. Compute Longest Prefix Suffix (LPS)
4. $k = 0, j = 0$
5. **while** $j <$ length. $W_p(t)$ **do**
6. **if** $W_{in}^{k}(t) == W_{p}^{j}(t)$ **then**
7. $j++,\ k++$
8. **if** $k =$ length. $W_{in}(t)$ **then**
9. return $j - k$
10. $k = LPS[k - 1]$
11. **elseif** $j <$ length. $W_p(t)$ & $W_{in}^{k}(t) != W_{p}^{j}(t)$
12. **if** $k > 0$ **then**
13. $k = LPS[k - 1]$
14. **else**
15. $j++$
16. return -1

Part II: Selection of DC i*

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue11 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

17. Compute $S_i^{in}(t), S_i^{ac}(t), R_i(t), P_i(t)$ to select the i^*

$$i^* = \min_{\arg i \in [1,N]} \sum_{i=1}^{N} \lambda_i [C_i(t) + d_i(t)]$$

Subject to limitations (6a), (6b), (6c), (6d), (6e) and (7)

18. Allocate the incoming workload to the data center i^*
19. Update $w_i(t), S_i^{ac}(t)$ and $S_i^{in}(t)$ ∀ $i \in [1, N]$
20. **end for**

DDGLB is structured into two sections. In Section I (lines 3–16), the content of $W_{in}^k(t)$ undergoes comparison with previously processed documents, $W_p^j(t)$, utilizing the Knuth–Morris–Pratt (KMP) data deduplication algorithm. Upon a successful match, GSLB refrains from directing $W_{in}^k(t)$ to a specific data center i^* for the execution of workload for conserving energy, and instead displays the user the outcome of $W_p^j(t)$. The incoming workload $W_{in}^k(t)$ is assessed for all data centers every time t. The longest prefix suffix and the already processed workload $W_p^j(t)$ are determined, with their values assigned to j and k, respectively (lines 3 and 4).
In Section II of DDGLB (Lines 17–20), when no match is identified between the incoming workload and the processed workload, Global-LB allocates the $W_{in}^k(t)$ to a DC i^*, considering various parameters for optimization such as server load in active mode, cost of bandwidth, dynamic prices of electricity, and the current renewable energy level. Bandwidth cost, current renewable energy level, number of active and inactive servers, and energy consumption in data center are computed for all data centers at time t in line 17. The algorithm selects the data center i^* based on the minimum bandwidth costs and electricity, while adhering to a set of constraints (6a) to (6e) and (7). Subsequently, the $W_{in}^k(t)$ is allocated to DC i^* for processing in line 18. Finally, the algorithm updates the active servers, inactive servers, and incoming workload for all data centers a for the upcoming user requests.

Experimental Setup

The experimental setup for assessing the efficiency of DDGLB involves a thorough investigation under both simulated and real-world traces, focusing on onsite green energy production and workload in geographically distributed data centers. The following sections detail the numerical evaluation:
Incoming Workload
Features of Data Center
Dynamic Electricity Prices
Onsite Renewable Energy Generation
Benchmark Algorithms

Incoming Workload

The tests conducted in this research study utilize batch workloads derived from real Wikipedia user query data. The dataset incorporates Coordinated Universal Time (UTC) for 10% of all user queries on Wikipedia from January 1, 2020, to January 30,

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue11 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

2020 [41]. To facilitate analysis, we create $T = 720$ time slots, averaging the data over hourly intervals to align with fluctuations in electricity prices. Assuming single workload consumes 10% process power of a server in active mode. Moreover, we evenly distribute the workloads across multiple servers. The traces depicted in Figure 2, illustrate the daily trends of normalized batch workloads in the real-world scenario.

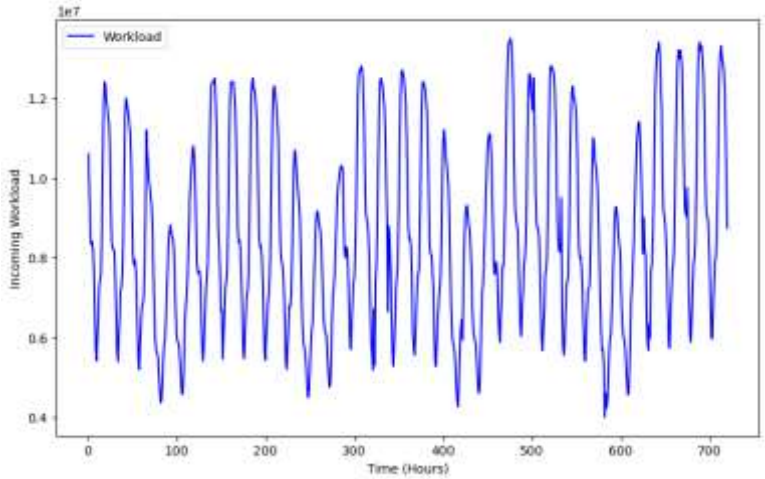


Figure 2. Incoming hourly workload

Features of Data Center

We consider a Global Server Load Balancer (GSLB) overseeing three geographically separated data centers (DCs). The assumed locations for these data centers are Illinois, USA; Utica, New York; and Ontario, Canada, respectively [7,8]. GSLB and DC bandwidth costs are adjusted to align with the power price and the geographical distances between the cities [18], [35]. Each data center is anticipated to host a maximum of 350 operational servers. The power consumption rates for individual servers are preconfigured at 100 watts and 120 watts in active and inactive modes, respectively [8], [36]. The processing speed of a single server is $\mu_i = 1$ [11]. It's important to note that we assume homogenous settings for the DCs to specifically measure the impact of various parameters for optimization, such as dynamic prices of electricity, availability of renewable energy, and bandwidth cost.

Dynamic Electricity Prices

We utilize the hourly energy pricing data, expressed in dollars per megawatt-hour (MWh), for the geographically distributed data centers. This data is obtained from publicly available government websites [36]. The hourly electricity prices from different regions are illustrated in Figure 3. The electricity prices are considered for the period between October 17, 2022, and November 18, 2022. Our observations indicate that Ontario, Canada's electricity costs occasionally decrease during periods of less power demand, high renewable and conventional energy supply.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue11 (2025)

Online ISSN
3007-3197

Print ISSN
3007-3189

http://amresearchreview.com/index.php/Journal/about

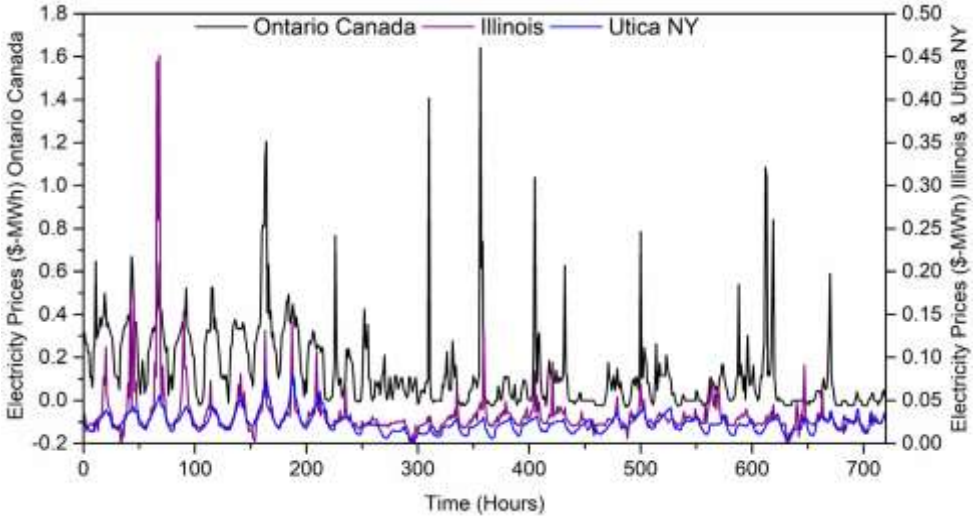


Figure 3. Electricity prices (hourly)

Onsite Renewable Energy Generation

Renewable energy manifests in diverse forms and exhibits considerable variability, even within a single session, such as an hour. Energy storage devices (ESDs), prevalent in every data center, provide consistent green energy [31]. Accordingly, our projection anticipates that 18% of the total energy consumption across data centers will be derived from on-site renewable energy generation within a single period ($t = 1$ hour) [41]. Figure 4 illustrate the renewable energy generation of the three geographically distributed data centers.

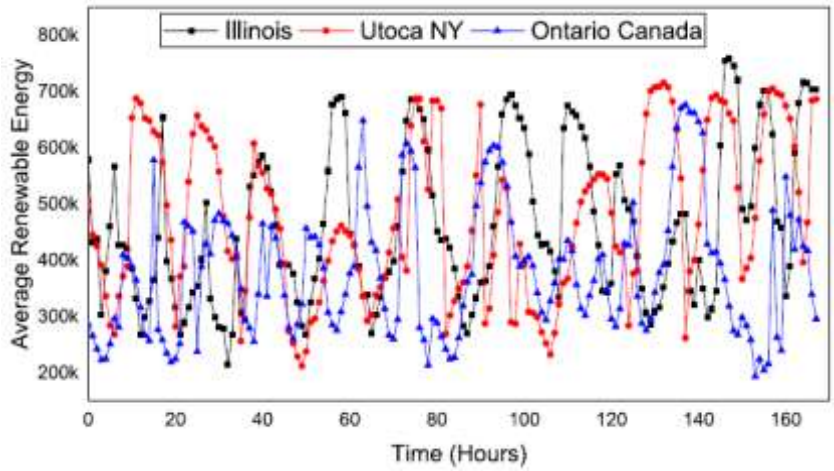


Figure 4. Renewable energy generation

Benchmark Algorithms

To establish benchmarks for the performance of DDGLB, we compare its outcomes to three baseline algorithms, either newly proposed or closely aligned with existing practices.

Benchmark Algorithm 1 (BA1) [1]: This initial benchmark algorithm adopts a temperature-independent approach, integrating incoming workload processing and data center selection. It initially permits the data center to handle an interactive

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue11 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

workload, aiming to reduce the overall operating cost. The optimization of user request routing relies on available capacity, considering only the fluctuation of electricity costs. Cooling energy is calculated with a constant Power Usage Effectiveness (PUE) of 1.2. Despite its simplicity, this approach is commonly employed in modern Internet businesses and facilitates an implicit assessment of prior work.

Benchmark Algorithm 2 (BA2) [8]: In this second baseline algorithm, we utilize the proposed workload allocation technique namely GreenGLB [40] to demonstrate its effectiveness compared to the proposed algorithm. GreenGLB leverages variations in power rates among various locations by routing incoming workloads. However, this study excludes batteries and data deduplication techniques. Incorporating data deduplication algorithms, GreenGLB can more equitably compare energy costs, as its objective is to reduce the overall operating cost in every timeslot based multiple optimization parameters.

Benchmark Algorithm 3 (BA3) [41]: Aimed at reducing the total energy cost of geographically distributed DSs. They introduce a string-matching-based geographical load balancing (SMGLB) for workload processing. The primary goal is to compare incoming workload contents with previously processed documents in a data center using a string-matching algorithm (Boyer Moore). A successful match prevents the GSLB from directing the incoming workload to a DC for execution. Instead, it immediately provides the user with the outcomes of the previously completed task, aiming to conserve energy.

Results and Analysis

This section presents and discusses the experimental findings conducted under the parameters outlined in Section 5.

Optimizing Energy Cost using Data Deduplication Technique

It is noteworthy that prior research studies, which prioritized reducing energy costs, mostly disregarded data deduplication techniques for workload processing. The proposed algorithm, DDGLB minimizes energy usage and operating expenses by utilizing data deduplication technique, global load balancing, renewable power sources, and interactive workloads. First, we assume that the incoming workload is sent to GSLB to compare its contents to previously processed documents in a data center using a data deduplication technique (i.e., KMP). The exact match precludes GSLB from distributing the incoming workload for processing to a specific data center. Moreover, GSLB propagates the results of the completed tasks to the user to minimize operating cost of the data center. As content matching is essential to DDGLB performance for a fair comparison, we select the incoming workload as an optimization parameter in our baseline models. If GSLB detects no match, the delay cost is the same in all benchmark algorithms.

Figure 5, Figure 6, Figure 7, and Figure 8 illustrate the average energy costs in geographically distributed data centers (DCs) for DDGLB and three benchmark algorithms, BA1, BA2, and BA3. Notably, DDGLB demonstrates lower average operating costs compared to the three benchmark algorithms. Specifically, BA2 incurs

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue11 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

the highest costs among all four algorithms, primarily due to the absence of data deduplication techniques.

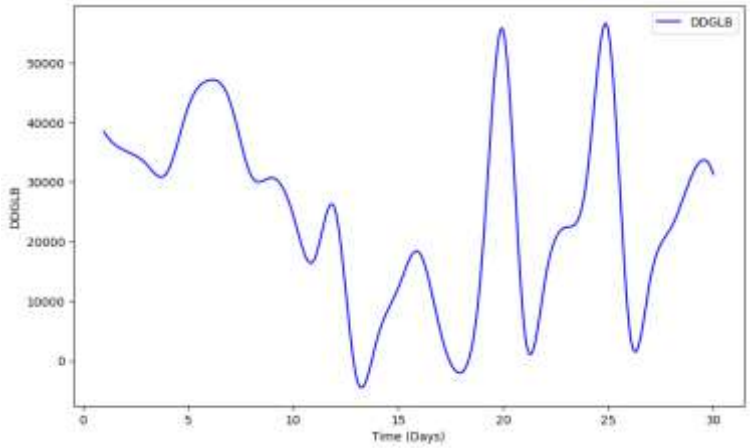


Figure 5. Average operating cost of DDGLB

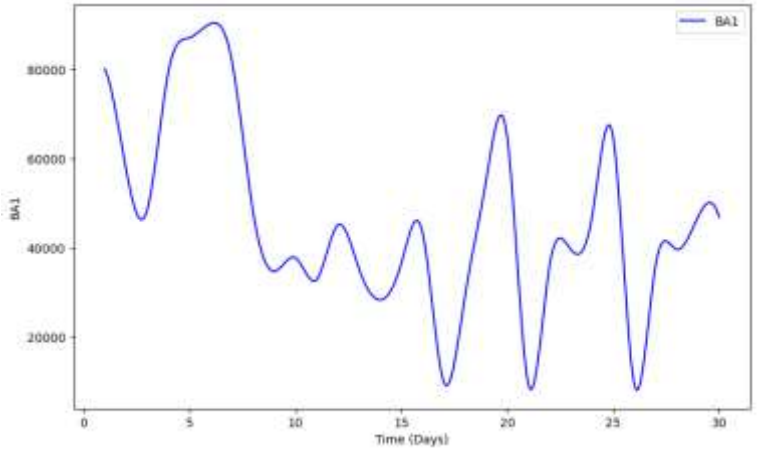


Figure 6. Average operating cost of BA1

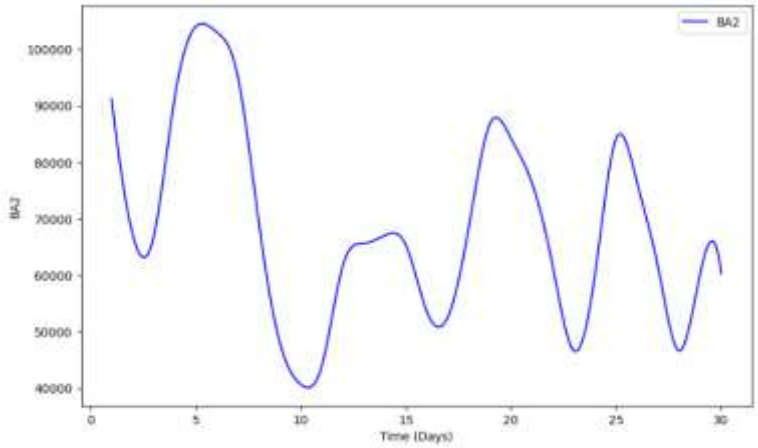


Figure 7. Average operating cost of BA2

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue11 (2025)

Online ISSN
3007-3197

Print ISSN
3007-3189

http://amresearchreview.com/index.php/Journal/about

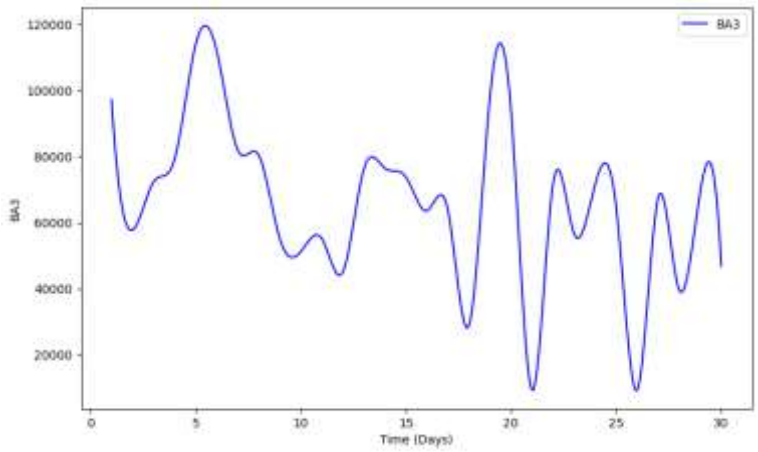
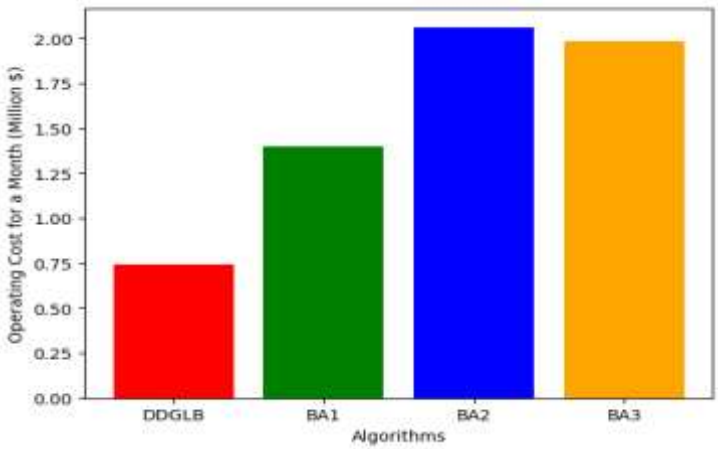


Figure 8. Average operating cost of BA3

Furthermore, DDGLB significantly decreases the data center's total operating expenses using the various optimization parameters such as, data deduplication algorithm, the use of renewable energy, dynamic electricity pricing, and geographical load balancing. Figure 9 and Table 1 depict the operating cost for a month in thousands for both the benchmark algorithms and our proposed workload distribution strategy, i.e., DDGLB. As shown in the figure, DDGLB surpasses the performance of all benchmark algorithms. Specifically, by comparing DDGLB and BA2, data deduplication can help lower the total energy cost. Furthermore, BA3 needs to be more knowledgeable about renewable energy sources and only attempt to process the user requests when electricity prices are lower. However, it considers the changing power price over time. It performs poorly and uses much renewable power as a result. This highlights the importance of task allocation in DCs with onsite green energy production. Lastly, by comparing BA1 and BA3, we can see how the user requests may raise the use of green power sources while lowering DC electricity costs.

Table 1. DDGLB improvement over benchmark algorithms

Comparison parameter	DDGLB improvement over		
	BA1	BA2	BA3
Energy cost	37.94%	49.85%	44.57%



Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue11 (2025)
Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Figure 9. Comparison of DDGLB with benchmark algorithms

6.2. Impact of Data Deduplication on Energy Storage Device Cost

To determine the impact of Energy Storage Device (ESD) costs on operational cost savings, we define S^i_{max} and V for each data center and assess DDGLB for different ESD loss values (ω_i). The outcomes are displayed in Fig. 14. We observe operational cost savings decrease as energy storage device costs rise. Please note that DDGLB does not use energy storage methods when ω_i is high. However, even in this instance, task load allocation based on data deduplication and GLB still yields cost savings over BA1 and BA2.

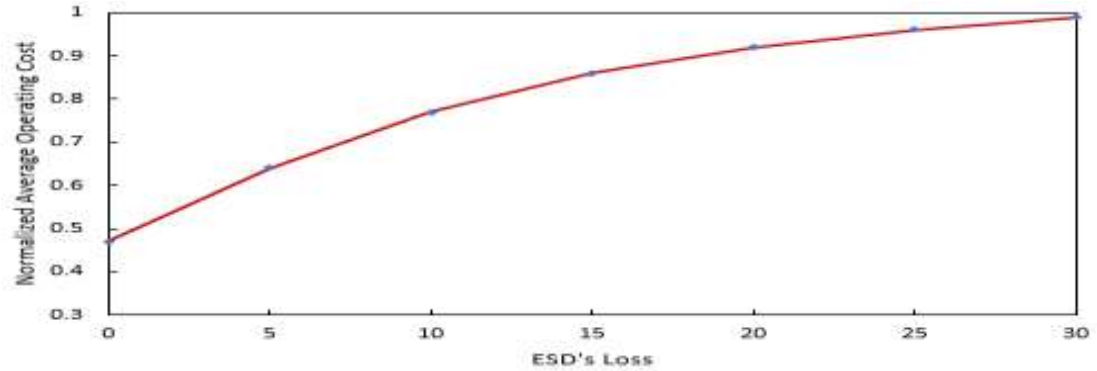


Figure 10. Impact of data deduplication and ESDs on operating cost

Optimizing Bandwidth Cost

We now compare our proposed algorithm DDGLB to the benchmark methods mentioned above for workload distribution and assess their bandwidth needs. Assuming that bandwidth cost is minimal and BA1 regularly allocates the user request to the nearest DCs. As seen in Figs. 6.1 and 6.2, the benchmark algorithms have higher running costs on average. In Fig. 6.1, DDGLB considers the difference in bandwidth pricing between GSLB and data centers, delivering the most significant potential decrease in overall operating expenditures. It is crucial to remember that even if BA1 has a higher energy cost and a lower bandwidth cost than BA3, their operating expenses are nonetheless equivalent. BA2 has the highest energy expenses despite using the least bandwidth.

Time Complexity of DDGLB

The time complexity of the DDGLB algorithm in the worst case is $O(n + m)$ in both scenarios, for example, when $W^k_{in}(t)$ does not exist in $W^j_p(t)$ and when $\mathbf{W_{in}(t)}$ occurs in $W^j_p(t)$. Moreover, $O(n + m)$ represents linear time complexity, where n is the length of the text and m is the length of the pattern.

Conclusion

With the growth of Cloud Service Providers (CSPs), large-scale commercial data centers (DCs) have rapidly increased. Several CSPs often use geographically distributed DCs to ensure service reliability and the quality of service of client support. Optimizing energy cost is becoming a contemporary problem for every data center due to the enormous amount of data being handled. Numerous studies have

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue11 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

employed dynamic energy pricing and geographical load balancing to manage incoming workloads, aiming to minimize the operating costs of DCs. However, many challenges still need resolution. This rationale explains why most earlier attempts to handle incoming workloads focused on reducing these expenses. In this study, we designed a Data Deduplication Geographical Load Balancing (DDGLB) algorithm based on the data deduplication technique to reduce the operating cost for geo-distributed DCs in a dynamic electricity market. DDGLB has two goals. Firstly, we compared the contents of incoming workload to previously processed workload in a DC using a data deduplication algorithm (i.e., KMP). To conserve energy consumption, a successful match stops Global Server Load Balancer (GSLB) from forwarding the incoming user requests to a data center for execution and presenting the outcome to the user. If no match is found, GSLB will assign the workload to a data center while accounting for several optimization parameters, including the server load, the current level of renewable energy, and the dynamic price of electricity to minimize the DC's running costs. Comprehensive analyses utilizing actual data demonstrate that DDGLB significantly reduces energy costs compared to current workload allocation techniques.

References

Hogade, N.; Pasricha, S.; Siegel, H.J. Energy and network aware workload management for geographically distributed data centers. IEEE Trans. Sustain. Comput. **2021**, 7, 400–413.

Jiang, D.; Wang, Y.; Lv, Z.; Wang, W.; Wang, H. An energy-efficient networking approach in cloud services for IIoT networks. IEEE J. Sel. Areas Commun. **2020**, 38, 928–941.

Shekhawat, V.S.; Gautam, A.; Thakrar, A. Datacenter workload classification and characterization: An empirical approach. In Proceedings of the IEEE 13th International Conference on Industrial and Information Systems (ICIIS), Rupnagar, India, **2023**; 1–7.

Kumari, A.; Tanwar, S.; Tyagi, S.; Kumar, N.; Obaidat, M.S.; et al. Fog computing for smart grid systems in the 5G environment: Challenges and solutions. IEEE Wirel. Commun. **2023**, 26, 47–53.

Albreem, M.A.; Sheikh, A.M.; Alsharif, M.H.; Jusoh, M.; Yasin, M.N.M. Green Internet of Things (GIoT): Applications, practices, awareness, and challenges. IEEE Access **2021**, 9, 38833–38858.

Chakravarthy, V.D.; Amutha, B. A novel software-defined networking approach for load balancing in data center networks. Int. J. Commun. Syst. **2022**, 35, e4213.

Ahmad, I.; Khalil, M.I.K.; Shah, S.A.A. Optimization-based workload distribution in geographically distributed data centers: A survey. Int. J. Commun. Syst. **2020**, 33, e4453.

Khalil, M.I.K.; Ahmad, I.; Almazroi, A.A. Energy efficient indivisible workload distribution in geographically distributed data centers. IEEE Access **2019**, 7, 82672–82680.

Jeba, J.A.; Roy, S.; Rashid, M.O.; Atik, S.T.; Whaiduzzaman, M. Towards green cloud computing an algorithmic approach for energy minimization in cloud data centers. In Research Anthology on Architectures, Frameworks, and

Online ISSN 3007-3197
Print ISSN 3007-3189

Integration Strategies for Distributed and Cloud Computing; IGI Global: Hershey, PA, USA, **2021**; 846–872.

Shakarami, A.; Ghobaei-Arani, M.; Shahidinejad, A.; Masdari, M.; Shakarami, H. Data replication schemes in cloud computing: A survey. Cluster Comput. **2021**, 24, 2545–2579.

Khalil, M.I.K.; Ahmad, I.; Shah, S.A.A.; Jan, S.; Khan, F.Q. Energy cost minimization for sustainable cloud computing using option pricing. Sustain. Cities Soc. **2020**, 63, 102440.

Javadpour, A.; Abadi, A.M.H.; Rezaei, S.; Zomorodian, M.; Rostami, A.S. Improving load balancing for data-duplication in big data cloud computing networks. Cluster Comput. **2022**, 25, 2613–2631.

Chong, F.T.; Heck, M.J.R.; Ranganathan, P.; Saleh, A.A.M.; Wassel, H.M.G. Data center energy efficiency: Improving energy efficiency in data centers beyond technology scaling. IEEE Des. Test **2013**, 31, 93–104.

Bird, S.; Aijit, A.; Othman, A.M.; Wenjin, H.; Kerop, J.; et al. Distributed (green) data centers: A new concept for energy, computing, and telecommunications. Energy Sustain. Dev. **2014**, 19, 83–91.

Koutitas, G.; Demestichas, P. Challenges for energy efficiency in local and regional data centers. J. Green Eng. **2010**, 1, 1–32.

Ernst, M.; Hogue, R.; Hollowell, C.; Strecker-Kellog, W.; Wong, A.; et al. Operating dedicated data centers–is it cost-effective? J. Phys. **2014**, 51, 062053.

Dhaya, R.; Kanthavel, R.; Venusamy, K. Dynamic secure and automated infrastructure for private cloud data center. Ann. Oper. Res. **2021**, 1–21.

Guo, Y.; Gong, Y.; Fang, Y.; Khargonekar, P.P.; Geng, X. Energy and network aware workload management for sustainable data centers with thermal storage. IEEE Trans. Parallel Distributed. Systems. **2013**, 25, 2030–2042.

Lei, H.; Wang, R.; Zhang, T.; Liu, Y.; Zha, Y. A multi-objective co-evolutionary algorithm for energy-efficient scheduling on a green data center. Comput. Oper. Res. **2016**, 75, 103–117.

Yu, L.; Jiang, T.; Cao, Y. Energy cost minimization for distributed internet data centers in smart microgrids considering power outages. IEEE Trans. Parallel Distrib. Syst. **2014**, 26, 120–130.

Gu, C.; Li, Z.; Liu, C.; Huang, H. Planning for green cloud data centers using sustainable energy. In Proceedings of the IEEE Symposium on Computers and Communication (ISCC), Messina, Italy, 2016; 804–809.

Chen, T.; Marques, A.G.; Giannakis, G.B. DGLB: Distributed stochastic geographical load balancing over cloud networks. IEEE Trans. Parallel Distrib. Syst. **2016**, 28, 1866–1880.

Rahman, A.; Liu, X.; Kong, F. A survey on geographic load balancing-based data center power management in the smart grid environment. IEEE Commun. Surv. Tutor. **2013**, 16, 214–233.

Shuja, J.; Gani, A.; Shamshirband, S.; Ahmad, R.W.; Bilal, K. Sustainable cloud data centers: A survey of enabling techniques and technologies. Renew. Sustain. Energy Rev. **2016**, 62, 195–214.

Liu, Z.; Lin, M.; Wierman, A.; Low, S.H.; Andrew, L.L.H. Greening geographical load balancing. ACM SIGMETRICS Perform. Eval. Rev. **2011**, 39, 193–204.

Online ISSN 3007-3197
Print ISSN 3007-3189

Lin, M.; Liu, Z.; Wierman, A.; Andrew, L.L.H. Online algorithms for geographical load balancing. In Proceedings of the International Green Computing Conference (IGCC), San Jose, CA, USA, **2012**; 1–10.

Abdelghany, M.A.; Mohsenian-Rad, H.; Alizadeh, M. Wholesale electricity pricing in the presence of geographical load balancing. In Proceedings of the Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, **2017**; 653–658.

Paul, D.; Zhong, W.-D.; Bose, S.K. Energy efficiency aware load distribution and electricity cost volatility control for cloud service providers. J. Netw. Comput. Appl. **2016**, 59, 185–197.

Toosi, A.N.; Qu, C.; de Assunção, M.D.; Buyya, R. Renewable-aware geographical load balancing of web applications for sustainable data centers. J. Netw. Comput. Appl. **2017**, 83, 155–168.

Lu, X.; Kong, F.; Liu, X.; Yin, J.; Xiang, Q.; et al. Bulk savings for bulk transfers: Minimizing the energy-cost for geo-distributed data centers. IEEE Trans. Cloud Comput. **2017**, 8, 73–85.

Goiri, Í.; Haque, M.E.; Le, K.; Beauchea, R.; Nguyen, T.D.; et al. Matching renewable energy supply and demand in green datacenters. Ad Hoc Netw. **2015**, 25, 520–534.

Qu, C.; Calheiros, R.N.; Buyya, R. Mitigating impact of short-term overload on multi-cloud web applications through geographical load balancing. Concurr. Comput. Pract. Exp. **2017**, 29, e4126.

Saleh, A.Z.M.; Rozali, N.A.; Buja, A.G.; Jalil, K.A.; Ali, F.H.M.; et al. A method for web application vulnerabilities detection by using Boyer-Moore string matching algorithm. Procedia Comput. Sci. **2015**, 72, 112–121.

Cantone, D.; Faro, S. Fast-Search: A new efficient variant of the Boyer-Moore string matching algorithm. In Proceedings of the International Workshop on Experimental and Efficient Algorithms, Berlin, Germany, 2003; 47–58.

Peng, Y.; Kang, D.-K.; Al-Hazemi, F.; Youn, C.-H. Energy and QoS aware resource allocation for heterogeneous sustainable cloud datacenters. Opt. Switch. Netw. **2017**, 23, 225–240.

Khalil, M.I.K.; Shah, S.A.A.; Taj, A.; Shiraz, M.; Alamri, B.; et al. Renewable-aware geographical load balancing using option pricing for energy cost minimization in data centers. Processes **2022**, 10, 1983.

Hampton, H.; Foley, A. A review of current analytical methods, modelling tools and development frameworks applicable for future retail electricity market design. Energy **2022**, 260, 124861.

Khan, F.; Jan, M.A.; Alam, M. Big data in healthcare: A survey. In Applications of Intelligent Technologies in Healthcare; Springer: Cham, Switzerland, 2019; 143–152.

Qureshi, A.; Weber, R.; Balakrishnan, H.; Guttag, J.; Maggs, B. Cutting the electric bill for internet-scale systems. In Proceedings of the ACM SIGCOMM 2009 Conference on Data Communication, New York, NY, USA, 2009; 123–134.

Rao, L.; Liu, X.; Xie, L.; Liu, W. Coordinated energy cost management of distributed internet data centers in smart grid. IEEE Trans. Smart Grid **2011**, 3, 50–58.

Online ISSN 3007-3197
Print ISSN 3007-3189

Khalil, M.I.K.; Shah, S.A.A.; Khan, I.A.; Hijji, M.; Shiraz, M.; et al. Energy cost minimization using string matching algorithm in geo-distributed data centers. Comput. Mater. Contin. **2023**, 75, 6305–6322.

Khalil, M.I.K.; Mubeen, A.; Taj, A.; Jan, N.; Ahmad, S. Renewable and temperature aware load balancing for energy cost minimization in data centers. J. Comput. Biomed. Inform. **2023**, 6, 1–8.

Luo, S.; Zhang, G.; Wu, C.; Khan, S.U.; Li, K. Boafft: Distributed deduplication for big data storage in the cloud. IEEE Trans. Cloud Comput. **2020**, 8, 1199–1211.

Jawad, M.; Qureshi, M.B.; Khan, M.U.S.; Ali, S.M.; Mehmood, A.; et al. A robust optimization technique for energy cost minimization of cloud data centers. IEEE Trans. Cloud Comput. **2021**, 9, 447–460.

Ali, S.M.; Jawad, M.; Khan, M.U.S.; Bilal, K.; Glower, J.; et al. An ancillary services model for data centers and power systems. IEEE Trans. Cloud Comput. **2020**, 8, 1176–1188.

Sajid, S.; Jawad, M.; Qureshi, M.B.; Khan, M.U.S.; Ali, S.M.; Khan, S.U. A conditional-constraint optimization for joint energy management of data center and electric vehicle parking-lot. In Proceedings of the 10th International Green and Sustainable Computing Conference (IGSC), Alexandria, VA, USA, October **2019**.

Muhammad, D.; Ahmad, I,; Khalil, M.I.K.; Khalil, W., Ahmad, M.O. A generalized deep learning approach to seismic activity prediction. Applied Sciences, **2023**, 13(3), p. 1598.

Khalil, T.; Jan, S.; Ghoniem, R.M.; Khalil, M.I.K. Multi-CNN fusion framework for predictive violence detection in animated movies. CMC-Computers, Materials, and Continua, **2025**.

Rahman, Z.U.; Khalil, M.I.K.; Nawaz, A.; Jan, N.; Ahmad, S. Analysis and clustering of Pakistani music by lyrics: A study of Coke studio Pakistan. Journal of Computing and Biomedical informatics. **2024**, 7(1), 281–296.

Khan, A.I.; Ullah, F.; Abrar, M.; Shah, S.; Taj, F.; Khalil, M.I.K. Ransomware early detection model using API-calls at runtime by random decision forests. International Review of Basic and Applied Sciences. **2022**, 10(2), 349–359.

Khalil, M.I.K. Improve quality of service and secure communication in mobile adhoc networks through group key management. International Review of Basic and Applied Sciences. **2013**, 1(3), 107–115.

Khalil, M.I.K.; Mubeen, A.; Taj, A.; Jan, A.; Ahmad, S. Renewable and temperature aware load balancing for energy cost minimization in data centers: A study of BRT, Peshawar. Journal of Computing and Biomedical informatics. **2023**, 6(1), 211–223.

Khalil, M.I.K.; Ullah, A.; Taj, A.; Khan, I.A.; Ullah, F.; Taj, F, Shah, S. Analysis of critical risk factors affecting software quality: A study of software houses in Peshawar, Pakistan. International Review of Basic and Applied Sciences. **2022**, 10(2), 338–348.

Akhtar, M. U.; Khalil, M.I.K. Link prediction techniques in complex networks. International Review of Basic and Applied Sciences. **2018**, 6(8), 60–64