International Journal of Cloud Computing and Database Management

E-ISSN: 2707-5915
P-ISSN: 2707-5907
IJCCDM 2024; 5(2): 93-96
www.computersciencejournals.com/ijccdm
Received: 18-07-2024
Accepted: 26-08-2024

Sweta S Munnoli
Java Full Stack Developer Internship, Kodnest, BTM Layout, Bengaluru, Karnataka, India

Devi Venkatesh Gowtham
Aarnavi Research Technology, 21A/1, Ammasai Street, K.K. Pudur Sai Baba Colony, Coimbatore, Tamil Nadu, India

Corresponding Author:
Sweta S Munnoli
Java Full Stack Developer Internship, Kodnest, BTM Layout, Bengaluru, Karnataka, India

AI-powered data center power control for sustainable development

Sweta S Munnoli and Devi Venkatesh Gowtham

DOI: https://doi.org/10.33545/27075907.2024.v5.i2b.74

Abstract
The effects of cloud computing on energy consumption and environmental degradation are rising to the forefront as the industry grows in size. Half of data centers' operational expenses are attributable to energy use, according to statistics. The increasing demand for energy puts significant strain on the environment and necessitates a substantial amount of energy. People in the IT industry are paying special attention to the problem of cloud data centers' excessive energy usage. This issue likewise requires immediate resolution. Two factors are now contributing to excessive energy use. First, a low server usage ratio is caused by a resources scheduling method that prioritizes completion time, and little jobs consistently use a lot of resources. Secondly, data centers' existing cooling systems are built on the peak value approach, which results in an oversupply of cooling, higher operational costs, and massive energy waste. This study proposes a new paradigm for green cloud data centers, taking into account the reasons for a great deal of energy of data centers. As a means of assessing the efficacy of optimal elastic scaling for actual data center traces, this study delves into workload and information center modeling to aid in the prediction of workload as well as data center operation. Working with an analytical approach, we simulate the workload and information center operations using three different machine learning algorithms. In order to determine the optimal design and scaling activities, we first utilize a mathematical framework as a predictor to assess and test the set of optimization solutions. Only then will we apply it to the actual data center. The findings demonstrate that analysis-oriented machine learning may aid in determining optimal workload demand forecast values and assessing the necessary provisioning of scaling and resource capacity. Finding the best setup and solving the elasticity scaling boundary values are both done via machine learning. We propose an intelligent refrigerating engine and a scheduling control engine that use AI-related strategies to reduce energy usage. Also, we validate the framework's viability, construct the scheduling control engine, and design a platform for environmentally friendly cloud data centers. It shows that the framework is capable of making the cloud platform run efficiently and with little power consumption and that data centers can run with little power waste.

Keywords: Green cloud and data center, energy optimizing, deep learning, and sustainable development

1. Introduction
As a whole, the energy sector is shifting its focus to be more environmentally friendly and low-carbon. At the heart of the digital era, data centers house massive amounts of mathematical demand. Improved energy efficiency, reduced energy waste, and carbon emissions may be achieved by the use of green low-carbon energy conservation management in data centers. Consequently, we suggest an AI-fitting control-based intelligent management approach for the dynamic energy utilization of data center networks [1]. To mitigate the extraordinary effects of climate change, sustainable development is obviously essential. Funding is a critical component in starting sustainability projects, particularly for SMEs. Finding and assessing viable initiatives is the first and most important stage in providing funding for sustainable development [2]. The digital and intelligent evolution of the nuclear business has led to the exponential growth of nuclear power plant data, which is now a valuable asset. Though nuclear power facilities may benefit from lifetime data management systems, they currently do not exist [3]. Machine learning, artificial intelligence, Recognizing speech, and OpenAI are just a few examples of the state-of-the-art technologies made possible by today's technological advancements. The voice-activated personal assistant is a well-known example of such technology. Among the many applications of machine

learning and voice recognition, this is among the most common [4]. The recent success of large-scale language models (LLMs) has attracted multi-hundred billion dollar investment in data centers focused on artificial intelligence (AI) during the next three to five years, thanks to their better capabilities across key sectors. Consequently, worries about AI's impact on the environment and the amount of energy it consumes are on the rise. Disruptive dynamic energy usage behavior, however, is an often-overlooked problem that is just as difficult and important as artificial intelligence models and infrastructure performance [5].

2. Related Work

This research [6] suggests an AI-powered automated approach to find viable sustainability initiatives that might need funding. In order to find possible sustainability projects for funding services, the suggested technique use web crawlers and text mining tools, such as Natural Language Processing (NLP), to scour the web, examine text data, and assess the information quantitatively. The suggested approach was put into action and evaluated objectively. With an accuracy rate of 87%, the AI-enhanced system can detect and rank possible sustainability initiatives. Financial professionals and decision-makers might use the study's findings to refine their current approaches to locating prospective projects to fund by making better use of the information readily accessible online. The article explores the revolutionary impacts of AI on precision agriculture, commonly called AgriTech or AgTech, in [7]. When artificial intelligence (AI) meets agricultural techniques, it changes the game by making better use of data analytics, automation, and machine learning to manage resources and make better decisions. By using real-time data from artificial intelligence technologies like remote sensing, satellite images, and Internet of Things (IoT) devices, precision agriculture allows for the maximization of agricultural yields with the minimization of waste. Improving production while decreasing environmental effect is possible with the use of AI-driven insights that allow for exact suggestions for irrigation, fertilizer, and insect management. Artificial intelligence also helps with fast-track genetic research, improves agricultural health assessments, and encourages climate-resilient crop varieties. The paper delves into the possibilities of artificial intelligence in areas such as sustainable practices, genetic editing, predictive analytics, and worldwide food security. Increased output is achieved via the use of AI-powered robots and automation in the agricultural sector. To realize this dream, studies have investigated the vast possibilities of organic agricultural practices and plant-based raw resources in [8]. Beyond their obvious practical uses, plants possess a vast reservoir of latent potential. The research explores several methods in which various plant components, such as branches, leaves, stems, and other "waste" materials, might be used to gain access to renewable resources, reduce environmental impact, and enhance self-sufficiency. Numerous case studies from different parts of the world show how this potential may benefit communities and the environment. The report also delves into the novel use of blockchain technology to strengthen India's agricultural ecosystem by making it more open and resistant to disruptions. Envision climate-smart strategies driven by blockchain technology, with secure and transparent transactions and data from sensors guiding precision agriculture. This goal seeks to enhance the resilience of the Indian agricultural industry via water-efficient irrigation, ecologically friendly pest management, and robust traceability mechanisms. Using a lifecycle node and business flow perspective, this paper [9] summarizes the current state of the literature on nuclear power plant lifecycle managing information in China, examines the demands and challenges of this field, and builds a data management framework. First, encouraging the full digital transformation of the nuclear power industry while standardizing nuclear power plant lifecycle data is a major undertaking. Second, creating data models and categorizing lifecycle information according to data types is another major task. Third, to guarantee the security of nuclear power plant lifecycle data, scientific data quality control rules and measurement standards must be developed. In order to examine how people in developing economies see food waste, the authors of the pioneering research [10] combined artificial intelligence analysis with participatory visual approaches (photovoice). In this research, twenty-six college students shared their thoughts and feelings about food waste in the home using written and visual means. One striking finding was that 32% of people were shocked by how much food they wasted on a daily basis, whereas 28% were quite unconcerned. To effectively evaluate the enormous visual and textual collection, AI-powered computer vision as well as language processing algorithms was used. A more nuanced understanding of consumer mindsets, actions, and the social and cultural factors that contribute to waste was obtained via the use of a mixed-methods approach. Concern for little waste, aesthetic and practical considerations, societal standards, and physical constraints were the main points. The dynamics of consuming practices were uncovered and critical awareness was raised via the participatory process.

3. Proposed Work

A. Datacenter Model

Client applications' workload needs across all three levels of cloud services (AaaS, PaaS, and laaS) are described by the proposed analytical model, with the laaS layer reflecting the total demands on data center infrastructure resources. Computing resources including processor, memory, network, storage, virtual machine, and container are provided via the framework as a service layer. We mapped out its properties using statistical analysis and created the analytical model based on our prior work [40] that analyzed the Google cloud data centers with workload needs. The (W) cloud workload is constructed from (J_s) submitted jobs and (T_s) submitted tasks within a time frame T_w. Increasing or reducing the amount of cloud resources requested by workload demands across a time period denoted with time window segments T_w influences the demands J_s and T_s. Some jobs will conclude with a successful completion (T_ff), while others will terminate (T_ne) owing to failure, eviction, death, or loss of direction.

In the cloud, resources are made available in reaction to demand for submitted tasks, and they are released once the jobs J_e along with tasks T_e are finished. An instance is Alibaba's version of a task. A vector of network resource needs (R) is required by one task in each batch of tasks (T_j) included in each submitted job j. The resource component R_ij of activity i in job j is where all the data center resources, whether provided or deprovisioned, are

recorded. The distributed resource vector for jobs with integer index i is represented by RE_i.

B. Machine Learning Modeling Approach

The workload and information center models are constructed in this part using deep learning approaches. The data center's provisioned resources and workload needs may be predicted with the use of machine learning models that mimic data center while workload behavior. One kind of deep machine learning technique is the Long Short Memory (LSTM) method, which is a subset of the RNN family; another kind is the Convolutional Neural Network (CNN), and the analytical method is compared to all three. In order for modeling to achieve its intended purpose of workload demand forecasting, the ideal data center layout must correspond to the requirements of a functioning data center. The suggested optimization component, which will be covered later on, will also be validated and evaluated using it. Using unique viewpoint features and correlations, the three ML algorithms simulate the data center and workload. In order to deal with the complexity of the model's nonlinear behavior, it is recommended to build the neural network having deep layers [13].

Machine learning models trained with NN rely on feed-forward thick layers of fully connected Neural Networks, with their hidden layers piled on top of one another. All of the neurons in the layer below it are linked to all of the neurons in the layer above it. Overfitting may occur if the total amount of layers is increased. Nevertheless, when dealing with non-convex target issues and high-density characteristics, it is possible to improve the model's accuracy by adding more layers using an appropriate relation architecture. This model does not transmit any location data. The input weight for every layer is updated by interactions among layers and a number of factors. No specific settings are implemented. A 32-neurod input layer with 32 characteristics forms the first of five layers that make up the model. Subsequent levels are "Dense" layers with 64, 128 and 64 neurons, respectively. Our last step is to apply a "Dense" layer using one neural network for the regression result. In order to organize the logs according to the expected properties, they are all generalized. Take workload as an example; all the other features are input variables, and total demand is the output attribute. All of the other models are based on the same premise.

To improve model accuracy and simplify parameter estimation, convolutional neural networks (CNNs) use a convolution operation to establish correlations between characteristics and extract model features. After the input layer changes the input shape from a row of (32×1)(32×1) vectors to a (8×4)(8×4) array, a new layer is added to the previous model using the filter "Conv2D" for a convolution operation. The "MaxPooling2D" layer follows, and it's responsible for downsampling the input representation through maximizing the value along the feature axis of the convolution filter window. Subsequently, a "Flatten" layer is used, which augments the output data size of the batch and adds an additional channel dimension [12]. As a last step, "Dense" layers of 64, 32, and 1 are applied. By training log features on localized regions of the log traces, convolution may maintain the spatial connection between the input row vector of the log [14].

Thirdly, Long Short-Term Memory (LSTM) RNNs rely on self-loops between layers of neurons to store memories. Information may be preserved across neurons over time via these self-loops. For quick pattern recognition, this preserves particular patterns in neural network decision processes. The LSTM system can handle sequences of varying lengths, keep track of data dependencies across time, remember the order of the sequence, and communicate parameters among all of them. An LSTM's four gate types—a "forget" gate that removes superfluous data, a "store" gate that allows relevant data to be stored from the current input, a "update" gate that selectively stores data to be used in the "update cell state," and finally, a "output" gate that returns an altered version of the cell state—form the basis of the scheme. The model's architecture begins with a 32-neurod "LSTM" layer, then moves on to three, four, and a half "Dense" layers. The model is then trained using 32 neurons from the "LSTM" layer, 16 from the "Dense" layer, and 1 from the "Inverse" layer.

4. Results and Discussion

A. Testbed

This study presents a green cloud architecture that is enabled by AI and introduces the scheduling control engine's associated algorithms in depth. That is, to ensure the RL-based decision approach is accurate. The suggested schedule control engine is implemented in the CloudSim cloud computing simulations environment in this research. The Java-based CloudSim is a tool for simulating cloud computing environments. It offers cloud computing capabilities that facilitate schedule simulation and resource management in the cloud. Researchers may avoid the hassle of real deployment and do algorithm tests on a single workstation while simulating large-scale cloud clusters. Hence, the CloudSim simulation setting was used for Experiment 1's studies. The assumption is that the arrival of the request mission follows a Poisson distribution. Table I displays the characteristics of the simulation environment used in Experiment 1. Fifty physical machines and fifty virtual machines make up the total. With a MIPS (Million Instructions Per Second) of 2,500, every physical computer has the same configuration. Since requests are dynamic and subject to change in the real world, system stability and timeliness are of the utmost importance. When we set up our cloud data center platforms at Inspur data center, we put it through its paces in a real-world setting to make sure the model is reliable. Two management nodes and seven data nodes make up Inspur's huge data center. 253 terabytes of data may be stored. Like a little data center, it serves a purpose. The results of the experiment conducted at the Inspur large data center prove that our model is available. After configuring the environment as described above, we begin by running a data center task simulation using the specified parameters on the simulation platform. In order to forecast the quantity of tasks at the subsequent time point, the training of the sample data set yields a trained prediction system. Each one sample size is a 6-dimensional vector, and it is used to forecast the number of jobs at the following point in time using data from the preceding 5 phases.

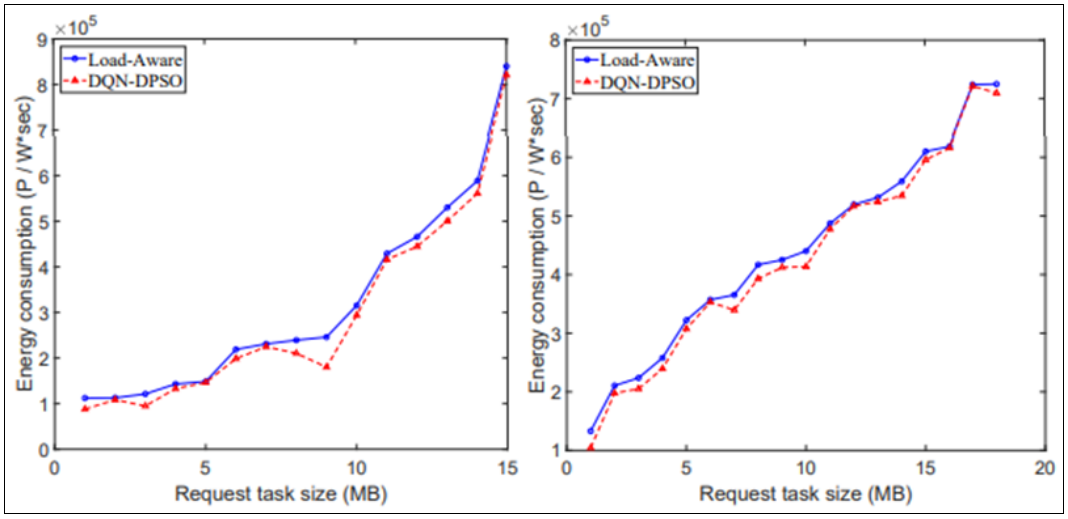

Fig 5: Experimental results for the performance evaluation of the energy consumption in Cloudsim

The first experiment's results are shown in Fig. 5. In this case, the x-coordinate represents the request's size and the y-coordinate the median energy use required to complete the operation. By sorting the task sizes in ascending order while performing two tests, we can easily see the energy consumption that corresponds to the size of the requested work. Compared to the load-aware resource method, our approach optimizes energy use more effectively, as shown in the figure. Then, we load five computing workloads into the data center's Testbed, observe disk access, service time, and CPU utilization rate, and see how well the model works.

5. Conclusion

Incorporating AI into cloud data centers and proposing an AI-enabled green cloud architecture are the main points of this article. Also included in this work are clever refrigeration engines and schedule control engines. The rational scheduling of resources has one goal: to increase the pace of resource usage while simultaneously decreasing the need for physical machines and energy consumption. On a related note, our work enables smart refrigerating system adjustments that take resources and the environment into account, which in turn reduces the cost of running and maintaining cloud data centers.

6. References

1. Li Y, Mughees M, Chen Y, Li YR. The unseen AI disruptions for power grids: LLM-induced transients; c2024.
2. Zulu ML, Pillay Carpanen R, Tiako R. Optimization of Internet of Things (IoT) for smart grid energy management using artificial intelligence (AI) techniques to reach SDG7. In: Proceedings of the 2024 32nd Southern African Universities Power Engineering Conference (SAUPEC); c2024. p. 1-6.
3. Effectiveness. J Math Comput Appl.
4. Alshahrani R, Yenugula M, Algethami H, Alharbi F, Goswami SS, Naveed QN, *et al*. Establishing the fuzzy integrated hybrid MCDM framework to identify the key barriers to implementing artificial intelligence-enabled sustainable cloud system in an IT industry. Expert Syst Appl. 2024;238:121732.
5. Talha MA, Rahman Rashed R, Ahmad S, Rofique R, Bhuyan PH, Howlader S, *et al*. IoT-based real-time monitoring and control system for distribution substation. In: Proceedings of the 2023 10th IEEE International Conference on Power Systems (ICPS); c2023. p. 1-6.
6. Sethy NK, Yenugula M, Goswami SS, Bhola A, Behera DK. Selection of ideal IoT-based overhead conductor for optimizing the performance of a small hydropower project; c2023.
7. Manukonda KR. Assessing the applicability of DevOps practices in enhancing software testing efficiency; c2022.
8. Kumar B, Anand V, Dhanabal R, Karthikeyan P, Jeyakiran R. A frequency regulation of an isolated distributed generation with electric vehicle and battery management system integration using artificial neural network technique. In: Proceedings of the 2024 Second International Conference on Smart Technologies for Power and Renewable Energy (SPECon); c2024. p. 1-6.
9. Yenugula M. Optimizing load balancing for green cloud via efficient scheduling. Int J Adv Acad Stud. 2023;4:224-30.
10. Behrooz H, Lipizzi C, Korfiatis GP, Ilbeigi M, Powell M, Nouri M. Towards automating the identification of sustainable projects seeking financial support: An AI-powered approach. Sustainability; c2023.
11. Debnath B, Kumar Basu S. AI-powered precision agriculture: Reshaping farming for efficiency, sustainability, and global impact. Int J Agric Sci; c2023.
12. SP, Sanghvi M, Josephine Williams M, Tandon A, Rawat G, Som A, LM. Optimizing food production with a sustainable lens: Exploring blockchain technology in raw plant materials and organic techniques in achieving sustainable development goals. In: Proceedings of the 2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies; c2024. p. 1-9.
13. Yenugula M, Kodam R, He D. Performance and load testing: Tools and challenges. Int J Eng Comput Sci. 2019;1:57-62.
14. Ren J, Yang P. Lifecycle data management of nuclear power plant: Framework system and development suggestions. Chinese J Eng Sci; c2022.
15. Gul K, Fasih S, Morande S, Ramish M. Participatory visual methods and artificial intelligence-driven analysis for sustainable consumption insights. Sustainability; c2024.