An Energy Efficient Dynamic Schedule based Server Load Balancing Approach for Cloud Data Center

Rajesh Bose¹, Sudipta Sahana²*, and Debabrata Sarddar³

¹Ph.D. Scholar, Department of Computer Science & Engineering, University of Kalyani, West Bengal, India
²Assistant Professor, Department of Computer Science and Engineering, JIS College of Engineering, Kalyani, Nadia, W.B. India
³Department of Computer Science & Engineering, University of Kalyani, Nadia, West Bengal, India

bose.raj00028@gmail.com, ss.jisce@gmail.com, dsarddar@rediffmail.com

Abstract

Cloud computing has firmly installed itself as a highly evolved concept for hosting and providing hardware and software resources across networks and the Internet. With rapidly emerging markets, cloud service providers have come up against a significant hurdle. Aspiring to remain firmly competitive in the long run, cloud service providers have realized that maintaining energy efficient controls in place without compromising performance is an aspect that cannot be ignored. With ever expanding sphere of cloud computing, energy demands for supporting computing resources and networks keep on growing. The high rate of demand for crucial energy needs is the salient point that keeps featuring in the horizon of almost every cloud computing service provider. Servers powering cloud computing services need to be supplied constantly with energy to support end users. To help ensure reduced energy consumption, we have examined the application of dynamic time schedule based server utilization method in our work. In this paper, we have applied this approach to cut back generation of heat by attempting to avoid server overloads. Our method involves different power consumption patterns which help saving energy costs. Consequently, carbon emission rates are kept under control. Thus, achieving a green cloud computing model is possible without additional cooling systems which would, in turn, have required more power to operate.

Keywords: Data centers, energy consumption, server utilization, remote computing, carbon emission.

1. Introduction

The adverse long term effects of using energy-hungry devices have been well-documented and are understood almost universally. To mitigate harmful effects arising out of the production of energy to meet rising power demands, studies have been – and are still being made – to improve energy efficiency of almost every product and device. The concept of “Green Computing” has its roots in designing, engineering, manufacturing, consuming, utilizing, and finally disposing or recycling of computers and allied peripherals in a manner such that the environment is kept safe and protected. Across the world, companies and manufacturers are investing in the design of improved energy efficient computers and computing devices. Much importance has been attached to the reduction of use of dangerous and/or precious raw materials. Even more has been
attributed to progress made in recycling electronic devices and allied materials, collectively referred to as “e-waste”. The term “Green Information Technology” or “Green IT” is synonymous with “Green Computing”.

In order to achieve economic sustainability and bolster profit margins, Green Computing is usually the most preferred practice. Green Computing involves design, development and implementation of deploying production procedures which can not only function non-stop without adversely affecting our environment, but also utilize computer hardware which are energy efficient. Highly developed disposal and recycling procedures form key ingredients in Green Computing practices.

For operating high performance computing (HPC), web and enterprise applications which demand very high processing and computational assets, cloud computing offers infrastructure that is competitive not only in terms of operating expenses but as well as capital investments. Cloud computing offers a sustainable proportionate increase in productivity with every increase in the quantum of resources being harnessed. This has fueled a trend in the shift from the general approach of investing and owning often costly IT hardware resources, to subscription based models. Cloud computing offers on-demand subscription plans that range from renting out servers to plans which offer services on hire. With the aid of many commercial cloud computing offerings, users can work and collaborate on their work, either individually or jointly, from any part of the world. It is, however, corporations and enterprises who are able to often reap the rewards of shifting to cloud computing services. No longer do firms have to concern themselves with purchasing, installing, configuring and maintaining their own computing infrastructure. This, therefore, allows increase in savings in terms of often costly investments that can altogether be bypassed. Instead companies stand to gain from optimized cloud data center services, low cost, scalable computing resources, and quicker software deployment options.

By itself, a cloud computing infrastructure is either a single or a cluster of data centers working together. As such, the high demands placed on the servers in each data center entail large amounts of energy to operate. It is commonplace to find a data center operating with a thousand racks which require 10 MW of electricity to stay in operations [15]. With ever increasing demand for energy, the associated costs also rise quickly. It becomes obvious, therefore, that the cost of meeting energy requirements alone is an important factor when it comes to operating a data center.

In April 2007, a report presented by Gartner estimated that of the total global CO₂ emissions, the Information and Technologies (ICT) industry alone generates about 2%. This was almost equal to the then emission rate of the aviation industry that almost completely depends on fossil fuels to power commercial jetliners [16]. As a result of the insatiable demand for energy, and the consequent rise in rate of CO₂ emissions of cloud computing installations worldwide, focusing on how to minimize the adverse effects on the environment has become a priority. Energy efficient solutions, therefore, are important in order to make certain that cloud computing model stays economically viable and environmentally stable. Currently, data centers continue to form the vital constituent of cloud computing infrastructure. Most solutions are, therefore, targeted at data centers for cutting down intake of energy resulting in indirect reduction of CO₂ emissions [20]. While these solutions are effective at curtailing energy consumption to a significant extent, there are no guarantees of overall reduction in CO₂ emissions. For example, a cloud data center that uses energy generated from a relatively inexpensive fuel source such as coal, goes on to enlarge its carbon emission footprint and, thus, offsets the cost advantage. Considering such factors, we propose a Carbon-Aware Green Cloud Architecture for cutting down the size of carbon footprint of a cloud computing system. Our model aims to do so in totality without compromising quality of service, viz., performance, responsiveness, and availability, offered by cloud providers.
The rest of our paper is organized in seven sections. Section 2 titled “Related Work” touches upon some of the previous works in this direction. “Proposed Work” is presented in Section 3 in which we discuss the proposed technique. We introduce our proposed Best Fit Algorithm in the subsequent section marked Section 4. Section numbers 5 and 6 titled “Flow Chart” and “Result Analysis”, respectively, follow, thereafter. Finally, we bring our study to a close in “Conclusion” which is submitted in Section 7.

2. Related Work

Green Cloud Computing has gained rapid acceptance worldwide. The user base is growing at a significantly steady rate. A host of research work has gone toward establishing Green Cloud in the real world. This has steadily pushed upwards the rate of energy use in data centers. Cavdar et al., [1,2] presented Green Grid for optimizing energy efficiency in running data centers. Green Grid proposes factors such as Power Usage Effectiveness (PUE) [7], Data Center Efficiency (DCE) metrics [10], Thermal Design Power (TDP) [2], etc. Of these, PUE stands out as the common consideration.

Wikipedia describes PUE as a measure of efficiency in how a data center utilizes the power that is made available for its operation. The range of values of PUE may vary anywhere between 1.0 to infinity. If the value of PUE is seen to approach 1.0, it is an indication that efficiency levels are at the 100% mark and, further, that all available power is being utilized by the computing hardware resources. Lately, few organizations have managed to realize the goal of very low PUE levels. Google, incidentally, has attained a PUE level of 1.13 [9].

A PUE value of 1.5 is interpreted to mean an energy consumption of 1 KWh by computer hardware in use within a data center. Energy spent in cooling, and in work not considered productive, e.g., dissipation of heat generated by processors, etc. is calculated to be as 0.5 KWh. There are data centers where PUE values are known to reach 3.0, and in certain cases even exceed that mark. Implementing appropriate and efficient designs, PUE values of around 1.6 can be achieved [5]. Calculations by Lawrence Berklely National Labs have shown that out of 22 data centers studied, PUE values ranged from 1.3 to 3.0 [8].

Combining neural network predictor for cutting down energy consumption, Truong Duy, Sato and Inoguchi et al., designed the green scheduling algorithm [3]. The algorithm allows a server to forecast load at time “t” to the time it actually takes for restarting, and then computes to arrive at the peak load figure. Accordingly, the server state number is selected based on the peak load. For example, let Nₙ be the number of servers that are required; and Nₒ servers that are in the “On” state. If Nₒ is less than Nₙ, then choose the server which is in the “Off” state and send the restart signal to that server only. If Nₒ is greater than Nₙ, select a server which is in the “On” state and signal it to shut down.

Reduction of energy usage in data centers was also the focus of work of Fumiko Satoh et al. [4]. Designing energy management for the future, the authors devised a system consisting of an optimized virtual machine allocation combined with sensor management function. The system was designed with a view to bring down the energy consumption in multiple data centers. Subsequent results show that this will save around 30% energy. The system was also employed to lessen carbon emission rates.

Proposing an energy-aware software layer, Rasoul Beik et al. produced a work in which they designed an architecture to calculate energy consumption in data centers. This architecture would also enable users to utilize available energy efficiently.

In producing their work on cloud computing metrics, Bhanu Priya et al. worked towards bringing greener cloud to fruition. Their paper focused variations in energy models for a more energy efficient cloud model. The aim of their work was to reduce power consumption and bring down CO₂ emission rates to result in a greener cloud. The authors concentrated on virtualization, work load distribution and software automation.
They also went on to discuss pay-per-use and self-service which were considered vital factors for reduction in consumption of energy.

Kliazovich and Pascal Bouvry [12] stated that expenses incurred as a result of maintenance of cloud data centers and operations performed on cloud are on the rise [12]. In their paper, these authors examined how work load could be distributed among data centers such that energy consumption can be evaluated at the packet level.

Kaur and Singh et al. proposed a model that would calculate the energy wasted by producing different gases. In the course of examining the various challenges involved with energy in cloud computing, they put forward their proposed model that contained several fields: Data, Analysis, Record, Put on guard, and restrain. These were combined with the concept of virtualization to give their proposed model the final shape. The authors suggested that this model would help pave the way for a greener cloud by using energy efficiently and, thus, foster a sustainable and safe environment.

Hosman and Baikie et al., [14] challenged the very idea of providing continuous energy source to data centers in the field of cloud computing. The authors argued that a steady conventional energy source could be substituted by harnessing the power of the sun. Use of solar power to supply energy needs of cloud data centers is a much-discussed topic. In this paper, the authors proposed a small cloud data center which would utilize a combination of three technologies. These technologies involve using platforms that require lesser power, energy efficient cloud computing frameworks, and DC power distribution systems.

Owuso et al., [17] conducted a survey to determine the technology that could be considered state-of-the-art in the field of cloud computing. The paper quite eloquently presented the argument on energy efficiency as being the most contentious cloud computing topics. Using energy in an efficient manner is discussed in great detail by the author.

Introducing key approaches such as use of virtual machines, power management schemes, recycling of materials, and telecommuting of green cloud computing, Yamini et al., [18] has made an interesting presentation. The paper aims at reducing high levels of energy consumption by task scheduling or amalgamating resources in green cloud computing environment. Results shown in the paper demonstrate that while direct and severe reduction in consumption of energy may not be within the realms of possibility as yet, significant savings in electricity bills can be possible in cases where data centers have extremely large and complex setups to support cloud services and operations.

Buyya in his work [19] warned of high levels of production of harmful gases. He has pointed out in his paper that as demands for cloud computing continue to surge upwards at a rapid rate, energy consumption levels as well as emission of toxic fumes produced as a result of supplying energy would also reach and cross unacceptable levels. The proliferation of unhealthy gases is one of the major issues which concern health care experts. It is also one of the critical factors that plays a leading role in driving up the operations costs of cloud computing. Discussing the aspects of cloud computing that are key elements when considering the quantum of total energy consumption, Buyya has presented a lucid observation on the associated empirical evidence. The various components of cloud computing that invoke green cloud computing, has been discussed in this paper.

Buyya et al., [21] examine the contribution of a third party concept consisting of two types of directories. Introducing green offer and carbon emission, the authors suggest that these two directories can involve both users and service providers. The authors detail how green services can be provided and utilized to lower carbon emission rates. Services can be scheduled according to least \( \text{CO}_2 \) emission rates. This is facilitated by green brokers that approach the green offers directory for the required services and time the services accordingly.
Virtualization techniques are the focus of the work of Beloglazov and Buyya et al. [22] to slow down and minimize energy use. The authors examine how virtual machines can be organized using the dynamic reallocation technique. They propose switching off the servers which are not in use. As a result, energy that would have been otherwise not spent gainfully is saved. Translated in terms of real world cloud computing data center scenarios, significant amount of monetary and environmental rewards can be reaped.

While most works concentrated on directly influencing the impact and rates of energy consumption, Nimje et al. thought of focusing attention on the aspect of security. Their thinking was to work towards green cloud environment by employing the concept of virtualization. Several methods have been deliberated upon in this paper to enhance security and lower power use. Virtualization offers simplicity in terms of design, delivery, deployment and management. Consequently, this also mitigates the effects of large workloads in cloud data centers. In this work, Nimje has proposed a hypervisor setting that doubles as a provider of a high level of protection as well as serve as a virtualization platform within green cloud computing framework.

3. Proposed Method

In our paper, we have focused our attention on managing different energy states. We propose a Data Center Manager (DCM) model to monitor the servers in a cloud data center. The DCM would be responsible for managing the data center by triggering different energy states. Further, the DCM would scrutinize each cloud server in real-time to ensure that proper load balancing takes place. Thus, chances of server overload leading to an increase in carbon emissions are either eliminated or kept to a minimum as far as possible. We believe that implementing this DCM model would also enhance overall serviceable life of the cloud servers managed.

In our DCM model, we define the field of maximum utilization for each connected server to be at 100. We, then, assign the maximum allowable utilization value to be 95. This, therefore, leaves a margin that acts as a cushion to fend off potential server overload situations. As a result, the need to install additional cooling equipment is obviated. Consequent upon which, an indirect savings in terms of energy costs can be derived.

Our proposed DCM model has been designed to perform the functions as enumerated below:

- To observe and probe server utilization factors individually. The DCM would also maintain a close watch on the sum total utilization factor of all connected working servers.
- In order to maintain CO₂ emission rates at the lowest possible level, the DCM would manage energy states of each connected server. At any given point of time, a server can be at running, sleep, or ready state. By triggering the appropriate energy state, energy can be consumed more efficiently and CO₂ emissions kept under tight control.
- DCM would maintain a queue of available servers in ready state. In doing so, our proposed DCM would also check for existence of at least one server in ready state. The purpose of maintaining a queue of such servers is to allow for allocation as and when a requirement arises.
- A transition buffer would be maintained by DCM. This would facilitate in transitioning workload from one server to the next based on utilization factors.
- To ease distribution of tasks to a different server, DCM would maintain an allocation buffer.

3.1. Server State

Through our research, we have analyzed the different states of consumption of power. We have examined the effects of these energy states and the overall impact that these lead
to in minimizing CO₂ emission rates. While focusing on bringing down emission rates, we have shown how the appropriate energy states can help achieve this goal. The three energy state levels are described as follows:

Running state: Based on CPU utilization of a server, maximum power is drawn in this state.

Ready state: A server continues to run its basic activities with its CPU processes at near idling settings. In this state, consumption of energy is lower than that of running state.

Sleep state: In this state, a server consumes only the amount of energy required to bring it out of this state.

To illustrate how our proposed DCM model is able to save energy, we examine the durations of the three energy states that a server is subject to in a 24-hour cycle. Let the values $S_{\text{Running}}$, $S_{\text{Ready}}$, and $S_{\text{Sleep}}$ represent running, ready and sleep states represented in watts per hour. In our model, a change in energy state between these would be invoked based on demand and in accordance to the service to be provided to a user. A server can be in any of these three states at any given point in a day. Expressed in terms of time values, let $T_{\text{Running}}$, $T_{\text{Ready}}$, and $T_{\text{Sleep}}$ be the durations for $S_{\text{Running}}$, $S_{\text{Ready}}$, and $S_{\text{Sleep}}$, respectively, in hours. Thus, the sum total value of $T_{\text{Running}}$, $T_{\text{Ready}}$, and $T_{\text{Sleep}}$ would be exactly equal to 24 hours. Expressed in terms of energy consumed by a server in a day, the formula is given as $T_{\text{Running}} \times S_{\text{Running}} + T_{\text{Ready}} \times S_{\text{Ready}} + T_{\text{Sleep}} \times S_{\text{Sleep}} = E_{\text{total}}$ watt-hours. Most authors, according to the literature survey conducted, have concentrated on load balancing without taking into account the power consumption factor. In all those cases, those studies have overlooked the fact that a server might completely be in a running state for 24 hours. Hence, the energy consumed in such cases amounts to $T_{\text{Running}} \times 24$ hours. The consumption of energy, in other words, remains higher than it would be if our proposed model is implemented. Considering these factors, we can say that using our proposed technique a cloud data center would be able to save $(T_{\text{Running}} \times 24) - E_{\text{total}}$ watt-hours.

### 3.2. Transition Procedure

In a process we present as being a transition procedure in our work, we set the maximum amount of time that a server can be in the running state continuously to be 2 hours. A limit of 6 hours is set for a combination of running and ready states. Following expiry of the limit, the server is sent into a sleep mode to cool down. It is when a server is sent from the running to the sleep state directly that our proposed DCM has a vital role to play. At the very edge of nearing the end of the running time limit of a server, all processes in operation is suspended. The process stack is copied after the server is paused; and a command to execute a process is sent to it to send the server from the paused to state to sleeping mode for cooling. If the time limit runs out, the server state is switched directly from ready to the sleep state. On an on-demand basis, the server is triggered to wake up to a ready mode from the sleep state. This is made possible by our proposed DCM.

![Figure 1. Transition Cycle](image-url)
3.3. Server Utilization Factor Check

In this stage, our proposed DCM continuously scans the utilization factors of all connected servers. If a server utilization factor is detected to be lower than 5% for more than 5 minutes, the DCM is designed to then copy the process stack and transfer the same to the best fit server found. The current server is switched to the ready state to conserve energy.

Architecture

![Architecture for the Proposed Method](image-url)

Figure 2. Architecture for the Proposed Method
4. **Proposed Algorithm**

We present a two-stage algorithm design for our proposed DCM model. In the first stage, which we have termed “Best Fit Algorithm”, the server which is best-suited for the task at hand is identified. The second stage of the algorithm walks through the steps on how to insert a new process and assign the best fit server for the task.

**Example Scenario**

To help explain the workings of the two-stage algorithm, we present a scenario as an example. In this scenario, we have two servers – Server1 and Server2, concurrently functioning in the running state. The utilization factors of each are 60 and 80. According to our proposed “Best Fit Algorithm” technique, our model would select Server2 assuming that a new process with utilization factor of 5 is required to be handled.

4.1. **Best Fit Algorithm**

1. Scan utilization factor of each running server individually.
2. Subtract the utilization factors obtained from the maximum utilization factor ceiling. The set limit is 95.
3. Check the utilization factor required for the process that has newly arrived.
4. Compare this process utilization factor with the resultant values obtained after subtraction.
5. The server that has the maximum load and can yet accommodate the new process is selected as the “Best Fit” server.

4.2. **Insertion of new process**

1. A request for service to the DCM is sent by a node.
2. DCM places this request in its allocation buffer queue.
3. The request is analyzed by DCM to compute the required utilization factor requirement.
4. Compare the utilization factors of the servers in the running state, with the computed utilization factor.
5. If the factors match, the “Best Fit” server is signaled by the DCM to begin processing the task.
6. In case a “Best Fit” server is not found, DCM would then invoke a new server from its queue of servers in the ready state.
5. Flow Chart

Start

Check the utilization factor of each server

Subtract each utilization factor from the maximum utilization factor of 95. Say the result for one server is $X_1$ for the next server $X_2$; so on and so forth

Check the required utilization factor for newly arrived process. Say this factor is $F_1$

Compare $F_1$ with $X_1, X_2$ to $X_n$

Server that can accommodate $F_1$ and has the maximum load is selected as the Best-Fit server

Stop

Best Fit Algorithm
Yes
Assign task to best fit server

No
Invoke new server from ready queue by DCM

Is \( R_1 \) equal to server utilization factor

DCM analyzes request to find out the utilization factor requirement say \( R_1 \).

Compare the servers utilization factors available with the utilization factor requirement, \( R_1 \).

Request is placed in DCM allocation buffers queue

Node requests DCM for service

Start

Providing new process insertion

Stop
6. Result Analysis

The number of users consuming cloud-based infrastructure services is on the increase. Under such circumstances, the primary goal of our proposed work is to provide an economically sustainable model that is also eco-friendly. Owing to the incessant workloads that are being placed on cloud data center servers, emissions of CO$_2$ are on the rise as progressively more power plants are pressed into service. Most existing research works have overlooked the effects of consumption of energy by servers in an always running state. Without due consideration to the running state of the servers, authors of those works have devoted their efforts to fine-tuning load balancing techniques. By focusing on toggling between energy states in a systematic manner, we have shown how wastage of energy can be reduced. In Fig. 3, we have shown a graph comparing the power consumed using an existing approach with that of our proposed one. In Fig. 4, we compare carbon emission rates of existing approach vis-à-vis that of our proposed model.

Figure 3. Power Consumption with Respect to Time
The purpose of this paper is to create an IT service selection prototype that narrows down on the impact of selection and combination of services on environment. The world is witnessing a surge in demand for cloud-based services in IT services industry. Largely, this demand has been seen as substitute for on-site business applications; and in many cases, even to the extent of enhancing the potential of such applications. However, along with these developments, governments, corporations, institutions and users are insisting on IT practices and hardware which are safe and do not waste our environment. We have, in this paper, put forth a technique to manage power and apportion server workloads through time-scheduled method. Our primary objective for green cloud data centers is reducing the carbon emission rate. We have shown how this can be achieved by means of monitoring and controlling server overload instances, and balancing energy states where power consumptions are regulated based on server workloads. A systematic and object-oriented power management policy can significantly reduce the carbon emissions generated otherwise. We have introduced DCM which has been designed to play a key role in monitoring a data center. It accomplishes this task using the best fit algorithm which we have examined in this paper. The algorithm enables the DCM to manage servers within a data center in a way such that overload situations can be avoided. Thus, operation of servers in a data center becomes more efficient as the DCM is able to dynamically and flexibly manage power demands according to processing requirements.

### Reference


Authors

Rajesh Bose, he is currently pursuing Ph.D. from Kalyani University. He is an IT professional employed as Senior Project Engineer with Simplex Infrastructures Limited, Data Center, and Kolkata. He received his degree in M.Tech. in Mobile Communication and Networking from WBUT in 2007. He received his degree in B.E. in Computer Science and Engineering from BPUT in 2004. He has also several global certifications under his belt. These are CCNA, CCNP-BCRAN, and CCA(Citrix Certified Administrator for Citrix Access Gateway 9
Sudipta Sahana, he is an assistant professor of a renowned engineering college of West Bengal. For more than 4.5 years, he has worked in this region. He has passed his M.Tech degree in Software Engineering and B.Tech Degree in Information Technology from West Bengal University of Technology with a great CGPA/DGPA on 2010 and 2012 respectively. He is recently working in Ph.D. in the domain of “Cloud Computing”. He is a member of the Computer Science Teachers Association (CSTA), and also a member of International Association of Computer Science and Information Technology (IACSIT).

Debabrata Sarddar, he is an Assistant Professor in the Department of Computer Science and Engineering, University of Kalyani, Kalyani, Nadia, West Bengal, INDIA. He has done Ph.D. at Jadavpur University. He completed his M.Tech in Computer Science & Engineering from DAVV, Indore in 2006, and his B.E in Computer Science & Engineering from NIT, Durgapur in 2001. He has published more than 75 research papers in different journals and conferences. His research interest includes wireless and mobile system and WSN, Cloud computing.