Proceedings of the Second International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2018)
IEEE Xplore Part Number:CFP18OZV-ART; ISBN:978-1-5386-1442-6

The Future of Cloud Computing: Opportunities, Challenges and Research Trends

Amanpreet Kaur, V.P. Singh

Computer Science and Engineering Department,
Thapar Institute of Engineering and Technology
Patiala, Punjab, India
akamanpreet261@gmail.com, vpsingh@thapar.edu

Sukhpal Singh Gill

Cloud Computing and Distributed Systems (CLOUDS) Laboratory
School of Computing and Information Systems,
The University of Melbourne, Australia
sukhpal.gill@unimelb.edu.au

Abstract

In a cloud computing environment, datacenter consists of number of servers, cooling and power delivery equipment's that require enormous measure of computational energy to drive complex frameworks. Due to the rising demand of the computation power, datacenter has become the hub for significant increase in the power consumption, heat dissipation and rise in temperature of the servers. Cloud datacenter's energy consumption has increased tremendously due to increase in the computation requirements of the user workload. Thus, saving energy has become an important concern to address. Researchers proposed different techniques to optimize the energy consumption. In this paper, we focus on different aspects of cloud computing for holistic management of cloud resources in an energy-efficient, reliable and sustainable manner. We recognized different opportunities, identified research challenges and propose possible future research directions for cloud computing.

***Keywords*- Energy Efficiency, Consolidation, Energy Consumption, Cloud computing**

I. Introduction

Cloud computing enables the on-demand provisioning of flexible resources (infrastructure/platform/software) as services on the basis of pay as you use. Cloud computing has revolutionized the information and communication technology industry due to its rising demand. The fast increment in the distributed computing has brought about foundation of vast scale server farms containing enormous number of complex servers. These server farms expend colossal measure of electrical vitality and emanate CO_2 in environment. Literature reported that there are two important aspects where energy is saved are computing and cooling. The scheduling of workload or allocating the virtual machines to save the energy. Cooling involves applying mechanisms in order to remove the heat and reduce the thermal dissipation of a physical machine.

A. The Next Generation of Cloud Computing

The significant piece of the cost in data centers is involved in two major areas: computing and cooling. Huge amount of power is consumed in operating the servers and cooling down these servers. The stress for saving the energy bills and to lessen the carbon footprint created the need for conserving energy. As a result, the methods to lessen energy necessity in distributed computing has dependably pulled in researchers [5]. With the regularly expanding power utilization and high pressing density of servers, both the warmth scattered in server farms and the temperature have expanded significantly. High temperature is unfortunate in the task of a server farm for a few reasons. It decreases the unwavering quality of the servers, decreases the reliability and performance of the system. Subsequently, in this paper, a concise study depicting energy preserving methods has been introduced. Indeed, the use proportion of data centre resources is just 30% [3]. Allocating the tasks to minimum number of the hosts is critical for sparing energy. Furthermore, critical level of aggregate energy consumption in cloud computing is wastage in VM migration [4]. With the quick development of cloud server farms in both amount and scale, the energy devoured by server farms, has immensely expanded. The power utilization of server farms has immense effect on the environment [1]. The extending use of server farms and their growing demand for energy, has made the examination of energy utilization basic.

B. Related Surveys and Our Contributions

Previously researchers Aruzhan et.al [23] and G.B.Hima et.al [24] have done reviews on achieving energy efficiency in data centers. The various techniques and green approaches are uncovered till now for saving energy only for servers. This research work presents the systematic review different aspects of cloud computing for holistic management of cloud resources in an energy-efficient, reliable and sustainable manner.

C. Paper Organization

The paper is organized as: Section 2 outlines the background of energy efficiency approach. Section 3

describes the major techniques used for energy efficiency and the related work done using the given techniques. Section 4 explores the classification based on techniques and scope where energy efficiency is achieved. Section 5 the open challenges faced in a cloud computing environment. Section 6 summarizes the paper with the future scope in the field of energy.

II. Background

In general, energy efficiency strategies can be employed into following major areas: i) Servers ii) Storage iii) Memory iv) Network v) Cooling. Servers are the significant consumers of power, approaches for energy saving for servers include Dynamic Voltage Frequency scaling (DVFS), Server Consolidation, and Virtualization. Another significant energy consumer is networking infrastructure strategy for saving energy is turning off system components or placing them into rest mode, another solution is assigning the virtual network requests to a few physical network devices in case of low traffic. Cooling is other major aspect for achieving energy efficiency one of the approach for cooling is raised floor to change over warm air to cool air by expelling warmth to the outside.

A. Current Status of Cloud Computing

The evolution of energy efficiency as shown in *Fig. 1*, describes the advancement in existing strategies, new techniques are built in cloud computing to minimize the energy. This section explores the evolution of work done in energy over the years based on the parameters Quality of Service (QoS) and Focus of Study (FoS). Many Energy efficient algorithms (EEAs) improve the cloud environment along with improving the utilization, responsiveness, performance and other QoS parameters.

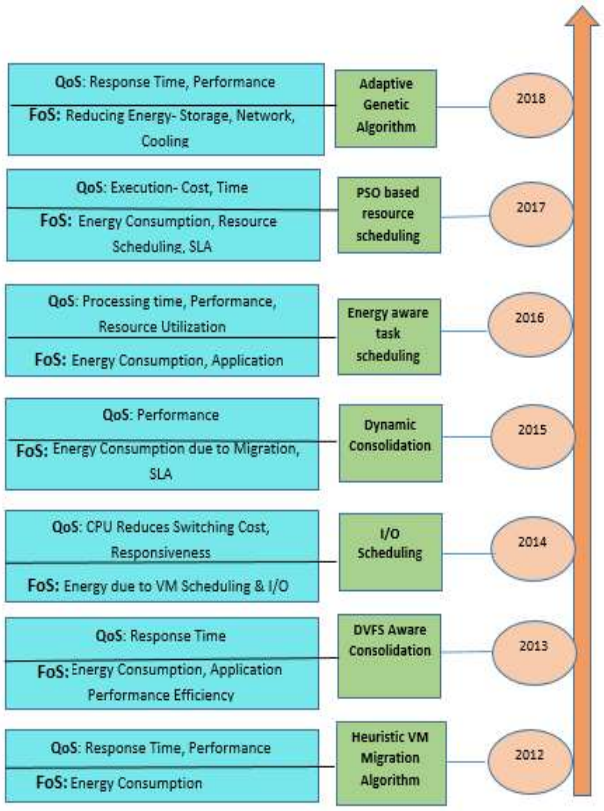

Fig. 1 Energy Efficiency Evolution

In 2018, Huda Ibrahim et al. [8] built up an Integer Linear Programming (ILP) model that minimizes the energy consumption in a Cloud. It focuses on dynamic workload scheduling technique. Energy efficiency and also the near optimal scheduling decisions are achieved by implementing an adaptive genetic algorithm. The algorithm finds the timetable for the underlying arrangement of tasks as received. Before introducing the new arrangement of tasks, the calculations experience the rundown of received undertakings and builds the list which depends on asked for and accessible capacities of assets equipped for executing each assignment.

In 2017, Singh et al. [9] proposed Particle Swarm Optimization (PSO) based resource provisioning and scheduling technique that aims to reduce energy

consumption and resource utilization along with execution cost, time and SLA as other parameters. In 2016, Leila Ismail et al. [10] derived an energy aware task scheduling strategy that takes into account the power consumption of the Cloud for energy-efficient resource utilization and increases the application efficiency.

In 2015, Subhadra Bose Shaw et al. [11] utilized the proactive and reactive hotspot detection technique to reduce the number of virtual machine migration as a result reduces the energy consumption in cloud data center. The concept is migration is performed after analysing that migration is required or not in case of hotspot detection. After taking the decision the VM will be shifted to a new host using a novel approach based on predicting the future load on the respective load. It performs when and where will VM will be migrated.

In 2014, Peng Xiao et al. [12] explored the VM scheduling policy named Share Reclaiming with Collective I/O (SRC-I/O) in order to compensate the energy losses caused by I/O virtualization.

In 2013, Yongqiang et al. [13] proposed a scheme of dynamic resource manager that took advantage of server consolidation and dynamic voltage frequency scaling. The energy efficient resource management framework where incoming workloads are submitted to its corresponding application manager through dispatcher module. These workloads are than allocated in round robin fashion to their virtual machine. In 2012, R. Karthikeyan et al. [14] discovered an efficient VM migration algorithm by using heuristic strategy to reduce the energy consumption and carbon emission.

III. Areas to Explore: Opportunities

Conserving energy in cloud computing specially in data centers is a major issue for the researchers. In this work, we aim to explore the various strategies for energy optimization in cloud data centre. There are number of ways by which energy consumption by data centers in cloud can be lowered some of the major techniques to save the energy consumption can be classified as shown in *Fig. 2.*

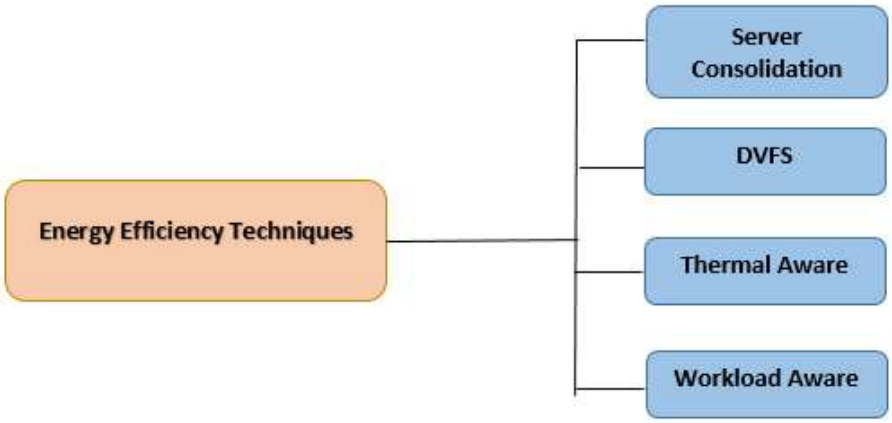

Fig. 2 Energy Efficiency Techniques

The techniques can be further classified based on other parameters. Most of the proposed scheduling algorithms aim to reduce the average energy consumption in the cloud center, other scheduling mechanisms target on reducing the high temperature of physical hosts while a few techniques are designed with a goal to reduce peak power consumption.

A. Server Consolidation

Aggregating the workload on fewer physical machines while turning off the rest of server machines. It is one of the energy efficient approach for achieving energy efficiency via virtualization or migration. To achieve energy efficiency low stacked PC frameworks are virtualized and run on couple of physical machines. Consolidation was done statically before, where low loaded virtual machines were manually migrated to one physical server. Dynamic consolidation permits to adjust the quantity of physical servers as per existing workload. It allows periodic reallocation of virtual machines to under loaded or normal hosts. It involves the detection of overloaded and under loaded hosts in the data center, which virtual machine to be migrated when to be migrated and where (physical machine) to be migrated [7].

Generally, there are two ways in which migration can be performed: regular migration and live migration. The main strategy includes moving a virtual machine from one host to other by delaying the initially utilized server, and continuing it on the target server while duplicating its memory substance from original server. The second technique plays out a similar usefulness

yet without stopping the server [6]. Dongyan Deng et al. [14] presented a energy efficient-oriented framework based on virtual machine framework. They introduced a VM placement policy called MAUD that take host list and VM migration list as input, algorithm considers the load balancing problem. The algorithm uses the difference between host utilization after accepting VM and the average utilization of data centre to optimize the energy.

Chun et al. [15] propose a hybrid server farm plan which utilizes heterogeneous stages to spare power. During low utilization phase, the employed technique exchanges the workload running on a high-performance host to a low-performance host and switch off the higher power servers. Performing server consolidation alongside workload migration in an energy efficient way.

B. Energy Efficiency

Dynamic Voltage Frequency Scaling Scheduling (DVFS) is an energy optimization dynamic technique for managing the power. DVFS is mainly done to lower the power consumption. DVFS is basically the adjustment of power and frequency settings of the computing devices in order to optimize the resource allotment for tasks and if resources are not required then maximize the power savings. Due to reduction in clock frequency of the processors less voltage is supplied. DVFS technique is used for virtual machines hosted by physical machines along with the algorithm or scheduling mechanism to reduce the energy. DVFS technique manages the power consumption of multicore processors, DRAM memories and other components. DVFS system and workload planning can be joined in two ways: (1) workload scheduling, and (2) slack recovery. In the schedule generation, tasks graph is (re)scheduled on DVFS-empowered processors in a worldwide cost function including both energy saving and make span to meet both energy and time limitations in the meantime. In slack reclamation, which fills in as post preparing method on the yield of planning calculations, DVFS procedure is utilized to limit the vitality utilization of undertakings in a timetable created by a different scheduler [10-11].

Patricia Arroba et al. [16] explored the dynamic voltage frequency scaling DVFS policy that takes into account the trade-off between the energy consumption and performance and a novel consolidation algorithm which is frequency aware while allocating cloud workload. The algorithm helps in boosting up the consolidation and reduces the number of active hosts.

Zhuo et al. [17] addressed the problem of energy consumption by proposing a DVFS enabled heuristic scheduling algorithm. First calculates the initial order of tasks and obtain the make span and deadline constrains using heft algorithm. From energy utilization obtain and merges the inefficient processors by reclaiming the slack time and redistributes tasks on it. Sharma et al. [18] presented an adaptive algorithm that minimizes energy consumption using a feedback loop which controls the recurrence and voltage levels to keep the momentary use of servers limited. The algorithm is implemented inside the Linux kernel for DVFS enabled processors, algorithm also adheres to the SLA.

C. Thermal Aware Scheduling

Allocation of workloads according to the temperature of physical machines in order to optimize the energy consumption as a way of reducing the cooling cost and the average temperature of the server. While scheduling the workload the operating system decides on which server the workload will be executed based on temperature history of the server. Thermal aware scheduling aims to avoid the creation of hotspots, performance degradation and reliability. To determine the temperature over time for server's various techniques have been proposed. One of the technique for performing thermal aware scheduling is thermal-aware monitoring and profiling. Thermal-aware monitoring includes frameworks to record and survey the heat dissipated from data centers. Thermal profiling is keeping the record of the characteristics of heat dissipated from servers, microchips, and computational workload. Thermal Aware Scheduling is done in 3 ways: i) Reactive Approach, ii) Proactive Approach and iii) Mixed Approach.

In reactive approach, the scheduling of the workload is done after thermal anomaly has occurred while in proactive scheduling is done before occurrence of any thermal anomaly. The thermal-aware scheduler uses the thermal profiles and predictions to place the workload across the data centre in order to lower the overall heat [l]. Ying-Jun Chen et al. [19] authors designed a thermal aware virtual machine migration manager that transfers the load from overheated physical machines to normal ones, by determining the temperature and resource utilization of the physical machine. Uses the proactive approach to save power; that employs heat transfer and migration time as criteria for VM selection policy and load balancing as VM allocation.

Moore et al. [24] developed two temperature aware workload placement policies: Zone based discretization and minimize-heat-recirculation. The first policy uses the information about steady state hot spots and cold spots in the data center for. Second policy minimize the total amount of heat that

recirculates before returning to the CRAC units and maximizes the potential utilization of each server. Yousri et al. [25] implemented the thermal aware scheduler that maps the virtual machine request to a physical machine with respect to the temperature of the host. Uses the thermal and power model for migrating the virtual machines according to temperature and utilization of the servers.

D. Workload Aware Scheduling

Present day server farms commonly have an expansive number of servers and thus, the choice about allocating the workload on particular servers influences the heat dissemination and power-utilization. Thus, workload aware scheduling is scheduling of the incoming workload on the basis of nature of the workload on the appropriate resources. Inappropriate arrangement results in incredible expand in the temperature of the data centre which will additionally build the warmth dispersal of the physical machines and furthermore increment the cooling necessities. Thus, workload-scheduling strategies have been proposed which put the workloads on available servers with the objective of power saving, lessening the temperature and the cooling necessities.

Ehsan Pakbaznia et al. [20] employed a short-term workload forecasting technique to predict the incoming workload to decide on the number of on servers and placement of workloads while simultaneously adjusting the supplied cold air temperature. Achieves the power savings by performing dynamic resource provisioning. R.K.Jena [21] paper centers to optimize energy and time using workload planning utilizing clonal section algorithm. The clonal algorithm is and adaptive based on clonal section theory as the new request of resources arrives, CSA is executed by the system to adjust the placement of resources. The algorithm optimally schedules user tasks to data centers randomly and each user task is assigned to the processing element of each allocated data centre.

IV. Holistic Management Aspects: A Comparison

Based on the above discussed literature, *Table 1* presents the comparisons of different holistic management aspects using different criteria such as year, algorithm, environment, scope, technology and Service Level Agreement (SLA).

Table 1: The comparisons of different holistic management aspects

Year	Holistic Management Aspect	Algorithm	Environment	Scope	Technology	SLA Agreement
2018	Workload aware scheduling	Adaptive Genetic Algorithm	Dynamic	Server, Storage, Network, Cooling	Single cloud data centre	No
2017	Energy aware scheduling	Scheduler	Homogenous	Server	CloudSim	Yes
2016	Server Consolidation	Underload decision algorithm	Dynamic	Server	CloudSim	Yes
2016	Workload aware scheduling	Genetic Algorithm	Dynamic	Server, Storage	Single cloud data center	Yes
2015	DVFS aware scheduling	Dynamic consolidation algorithm	Dynamic	Severs	CloudSim	Yes
2014	Energy aware scheduling	Scheduler-RESCUE	Heterogeneous	Server	Private Cloud	Yes
2013	DVFS + Server Consolidation	Dynamic Resource Management	Heterogeneous	Server, Network	Own testbed	Yes
2012	Thermal aware scheduling	Task Scheduling algorithm	Heterogeneous	Server, Cooling	Real Data Centre Environment	Yes

V. Open Challenges

Generally, there are a few complex issues advancing in the distributed computing condition in which noteworthy commitments can be made given that appropriate consideration is paid to them. There are various research issues to resolve in cloud computing which are as follows: There are following open challenges [1-7] in various different aspects of holistic management:

- *Server Consolidation:* There is significant increase in the utilization of used servers, degrading their performance as whole workload is concentrated on these servers. It may degrade the response time and maximize the transition costs.
- *DVFS:* The confinement of DVFS, is that a diminishment in frequency likewise lessens the performance of the circuit which consequently, affect the system performance. Thus, DVFS need to be used wisely, to maintain the performance.
- *Thermal Aware:* Monitoring the accurate inlet temperature of the servers, ambient temperature continuously is a tedious job, hence thermal aware scheduling need appropriate mechanisms for determining the temperature.
- *Workload Aware:* Predicting the nature of the workload according the history is quite cumbersome.

VI. Summary and Future Directions

In this research, we explored the issues in cloud computing environment more specifically pertaining to energy related. Analysed various algorithms employed using energy-efficient techniques in cloud data centers. Mostly research proposals are mostly focused on energy-saving approaches for servers. For the sustainability of cloud computing reducing the power consumption has become an important issue due to rise in power cost and rise in carbon emission. Researchers have applied various mechanisms to achieve the energy efficiency while maintain the SLA violations. This research effort presents major energy efficient approaches in cloud. Perhaps, it aims to emphasize the need of the efficient technique where energy efficiency of data centers can be achieved. In the future, the above-mentioned techniques will be applied in a synergistic way to provide much energy savings in a holistic way. The challenge is to develop a coordination framework that permits consistent combination of various approaches.

The further future directions can be:

- *Data Security:* Access to the physical security arrangement of server farms is not provided to the service providers. They rely upon the framework supplier to get full information security. In a virtual private cloud condition, the service providers can determine the security setting remotely, and don't know precisely those are implemented.
- *SLAs:* Monitoring the expected level of services; as a agreement done between the consumers and service provider. Taking care of QoS attributes is an integral part of SLA like performance, response time.
- *Fault Tolerance:* Fault Tolerance is technique that permits a system to keep performing when one of its part fails or it can be defined as capacity of a system to react nimbly to an unexpected equipment or programming break down.
- *Data Filtering:* Data from various geographically distributed data centres is quite big. Data filtering will involve taking out information that is useless to a reader or information.
- *Peak Temperature among Servers:* Temperature is another important parameter for both physical servers and virtualization solutions. Variance in the on chip temperature and the resultant occurrence of hot spots degrades the performance of processors, increases the energy consumption. Thermal management strategies are required to uniformly distribute the temperature.
- *Data Storage:* One of the problem resultant from storing vast data is data synchronization.
- *Total Processing Resource Wastage by the Physical Machine:* Resource utilization has linear relationship with the energy consumption. Thus a metric is required for measuring the utilization of resources.
- *High Level of Power Consumption by the Servers:* In efficient or non-energy aware scheduling techniques lead to increase in power consumption among the servers which degrades the server's performance, reliability.
- *High Energy Demand in Cooling Servers:* High power consumption lead to creation of

hot spots and increase in server temperature. Thus requiring cooling methodology for cooling the data centres.

References

[1] Karthikeyan, R., and P. Chitra. "Novel heuristics energy efficiency approach for cloud Data Center." *Advanced Communication Control and Computing Technologies (ICACCCT), 2012 IEEE International Conference on*. IEEE, 2012.

[2] Gill, Sukhpal Singh, and Rajkumar Buyya. "A Taxonomy and Future Directions for Sustainable Cloud Computing: 360 Degree View." arXiv preprint arXiv:1712.02899 (2017).

[3] Dzmitry Kliazovich, Pascal Bouvry, Yury Audzevich, Samee Ullah Khan, "GreenCloud: a packet-level simulator of energy-aware cloud computing data centers,"2010 IEEE Global Telecommunications Conference, 2010, pp. 1-5, doi: 10.1109/GLOCOM.2010.5683561.

[4] Singh, Sukhpal, and Inderveer Chana. "EARTH: Energy-aware autonomic resource scheduling in cloud computing." Journal of Intelligent & Fuzzy Systems 30, no. 3 (2016): 1581-1600.

[5] Singh, Sukhpal, Inderveer Chana, Maninder Singh, and Rajkumar Buyya. "SOCCER: self-optimization of energy-efficient cloud resources." Cluster Computing 19, no. 4 (2016): 1787-1800.

[6] Singh, Sukhpal, and Inderveer Chana. "A survey on resource scheduling in cloud computing: Issues and challenges." Journal of grid computing 14, no. 2 (2016): 217-264.

[7] C. Hsu, S.Chen, C.Lee, H. Chang, K. Lai, K. Li, C. Rong, "Energy – Aware Task consolidation Technique for cloud computing", published in IEEE 3rd International Conference on Cloud Computing Technology and Science, pp.115-121, 2011.

[8] Ibrahim, Huda, Raafat O. Aburukba, and Khaled El-Fakih. "An Integer Linear Programming model and Adaptive Genetic Algorithm approach to minimize energy consumption of Cloud computing data centers." Computers & Electrical Engineering (2018).

[9] Gill, S. S., Buyya, R., Chana, I., Singh, M., & Abraham, A. (2018). BULLET: Particle Swarm Optimization Based Scheduling Technique for Provisioned Cloud Resources. Journal of Network and Systems Management, 26(2), 361-400.

[10] Ismail, Leila, and Abbas Fardoun. "Eats: Energy-aware tasks scheduling in cloud computing systems." *Procedia Computer Science* 83 (2016): 870-877.

[11] Shaw, Subhadra Bose, and Anil Kumar Singh. "Use of proactive and reactive hotspot detection technique to reduce the number of virtual machine migration and energy consumption in cloud data center." Computers & Electrical Engineering 47 (2015): 241-254.

[12] Xiao, Peng, et al. "Energy-efficiency enhanced virtual machine scheduling policy for mixed workloads in cloud environments." Computers & Electrical Engineering 40.5 (2014): 1650-1665.

[13] Gao, Yongqiang, et al. "Service level agreement based energy-efficient resource management in cloud data centers." Computers & Electrical Engineering 40.5 (2014): 1621-1633.

[14] Deng, Dongyan, Kejing He, and Yanhua Chen. "Dynamic virtual machine consolidation for improving energy efficiency in cloud data centers." Cloud Computing and Intelligence Systems (CCIS), 2016 4th International Conference on. IEEE, 2016.

[15] Chun, Byung-Gon, et al. "An energy case for hybrid datacenters." ACM SIGOPS Operating Systems Review 44.1 (2010): 76-80.

[16] Arroba, Patricia, et al. "DVFS-aware consolidation for energy-efficient clouds." Parallel Architecture and Compilation (PACT), 2015 International Conference on. IEEE, 2015.

[17] Zhou, Zhou, et al. "Minimizing SLA violation and power consumption in Cloud data centers using adaptive energy-aware algorithms." Future Generation Computer Systems (2017).

[18] Sharma, Vivek, et al. "Power-aware QoS management in web servers." Real-Time Systems Symposium, 2003. RTSS 2003. 24th IEEE. IEEE, 2003.

[19] Chen, Ying-Jun, et al. "Using Thermal-Aware VM Migration Mechanism for High-Availability Cloud Computing." Wireless Personal Communications 97.1 (2017): 1475-1502.

[20] Pakbaznia, Ehsan, Mohammad Ghasemazar, and Massoud Pedram. "Temperature-aware dynamic resource provisioning in a power-optimized datacenter." Design, Automation & Test in Europe Conference & Exhibition (DATE), 2010. IEEE, 2010.

[21] Jena, R. K. "Energy Efficient Task Scheduling in Cloud Environment." Energy Procedia 141 (2017): 222-227.

[22] Kulseitova, Aruzhan, and Ang Tan Fong. "A survey of energy-efficient techniques in cloud data centers." ICT for Smart Society (ICISS), 2013 International Conference on. IEEE, 2013.

[23] Bindu, GB Hima, and J. Janet. "A statistical survey on vm scheduling in cloud workstation for reducing energy consumption by balancing load in cloud." Networks & Advances in Computational Technologies (NetACT), 2017 International Conference on. IEEE, 2017.

[24] Moore, Justin D., Jeffrey S. Chase, Parthasarathy Ranganathan, and Ratnesh K. Sharma. "Making Scheduling" Cool": Temperature-Aware Workload Placement in Data Centers." In USENIX annual technical conference, General Track, pp. 61-75. 2005.

[25] Mhedheb, Yousri, et al. "Load and thermal-aware VM scheduling on the cloud." International Conference on Algorithms and Architectures for Parallel Processing. Springer, Cham, 2013.