IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Vol.-6, Issue-12, Dec 2018 E-ISSN: 2347-2693

Energy-Aware Frameworks in Cloud Data Centers to Manage Workload and Diminish Power Consumption: A Survey

Rajesh P. Patel[1*], Ramji Makawana[2]

[1]Dept. of Computer Engineering, C. U. Shah University, Wadhwan, India
[2]Founder, e-Vsn Technology, Rajkot

*Corresponding Author: rpp.kirc@gmail.com, ramjimmakwana@gmail.com Tel: +91-9712823247

Available online at: www.ijcseonline.org

Accepted: 15/Dec/2018, Published: 31/Dec/2018

Abstract- Cloud Computing is a service model for enabling convenient, on-demand network access to a shared pool of configurable computing resources which can be rapidly provisioned and released. In cloud data centers various computing resources like servers, network devices and cooling systems which constantly evolve in size and in complexity so it consumes large amount of energy which increase extensive power consumption in data centers. As cloud data center resources are not optimized for their maximum utilization, they consume more power so it needs to consolidate virtual machines (VMs) of servers of data center which helps to optimize the usage of cloud resources hence reduce the energy consumption. By considering the optimized power consumption of various data center resources, the researchers have proposed various methodologies and algorithms to reduce power consumption in servers and network devices. In this paper, we have done insightful study of the modern techniques on data center's power model of servers, network components also on VM overload/under-load detection, VM selection and VM placement or consolidation of VMs which optimize the utilization of data center's servers for power model which and reduce energy consumption in data center.

Keywords: Server consolidation, VM Migration, Quality of Service, virtualized data center, Service Level Agreements, Highest Thermostat Setting, Energy efficient, virtual machine placement, migration, dynamic resource allocation, cloud computing, data centers

I. INTRODUCTION

Cloud computing that is providing computer resources as a service, is a technology Revolution offering flexible IT usage in a cost efficient and pay-per-use way namely networks, storage, servers, services and applications, without physically acquiring them [1]. This type of computing provides many advantages for businesses, shorter start-up time for new services, lower maintenance and operation costs, higher utilization through virtualization, and easier disaster recovery that make cloud computing an attractive option [2]. This technological development has enabled the realization of a new computing model, in which resources (e.g., CPU and storage) are provided as general utilities that can be leased and released by users through the Internet on-demand fashion [3].

Cloud computing which allows users to access services on demand. It provides pool of shared resources of information, software, databases and other devices according to the client request. Cloud computing services are related to software, platform, infrastructure, data and identity and policy management [4].

Cloud service delivery model in cloud environment states in three main types; Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) [5]. In IaaS, basic infrastructure layer services like storage, database management and compute capabilities are offered on demand. In PaaS, this platform used to design, develop, build and test applications. While SaaS is highly scalable internet based applications offered as services to the end user where end users can avail software or services provided by SaaS without purchasing and maintaining overhead [6]. In following figure: 1 Many data processing, social network services are software as service to be considered as workload which are handled by virtual machines. VMs run on public cloud infrastructure or private cloud infrastructure.

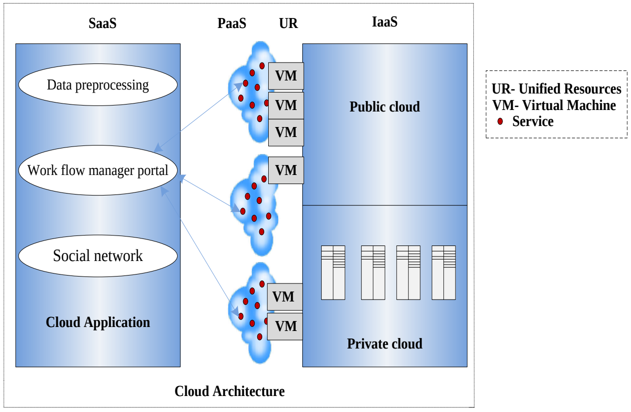

Figure: 1 Cloud Architecture

The four essential organization models in cloud computing are public cloud, private cloud, community model and hybrid cloud. To convey cloud computing services numerous computing services providers including Yahoo, Microsoft, IBM and Google are quickly sending data centers in various locations [5]. With a specific end goal to increase high efficiency and save power through expansions of IT the cloud computing has walked to the IT business. The cloud computing worldwide uptake has subsequently driven dramatic increments in datacenter power consumption. By the datacenters thousands of interconnected servers are composed and worked to give different cloud services [7].

With the rapid growth of the cloud computing technology and the construction of large number of data centers, the high energy consumption issue is becoming more and more crucial. The performance and efficiency of data center can be expressed in terms of amount of supplied electrical energy [8].

Datacenter put away more than 20 million gigajoules of energy per year, producing four million tons of carbon-dioxide emissions into the environment. Production of CO2 in the environment results in global warming. Cost of energy is rising and depletion of natural resources increases concern for the environment [9]. In 2007 report: The United States Environmental Protection Agency (EPA) declared that Datacenters consumed 1.5% of the total electricity produced in the all over country that year, the same of the combined use of 5.8 million average U.S. domestic or a cost of approximately $4.5 billion dollars[10].

The rest of the paper is organized as follows. Section-II contains the survey of load balancing and virtualization. Section-III describes power usage and management strategies for data center resources. Section-IV, explains various dynamic virtual machine consolidation algorithms proposed by researchers. Section-V, describes survey of various proposed energy aware models and virtual machine consolidation methods. Finally Section-VII concludes the whole paper.

II.LOAD BALANCING AND VIRTUALIZATION

In cloud environment the services requested by the client is rectified by employing virtual machines present in a server. Each virtual machine has different capabilities and resource requirements, so it becomes more complex to schedule job and balance the work-load among nodes [11]. Load balancing is one of the central issues in cloud computing it is a mechanism that distributes the dynamic local workload evenly across the entire server in the whole cloud to avoid a situation where some servers are heavily loaded while others are idle or doing little work [12]. The trend towards server-side computing and the exploding popularity of Internet services has made data centers become an integral part of the Internet fabric rapidly. Data centers become increasingly popular in large enterprises, banks, telecom, portal sites, etc, [13]. As data centers are inevitably growing more complex and larger, it brings many challenges to the deployment, resource management and service dependability, etc. [14]. A data center built using server virtualization technology with virtual machines (VMs) as the basic processing elements is called a virtualized (or virtual) data center (VDC) [15]

Virtualization is viewed as an efficient way against these challenges. Server virtualization opens up the possibility of achieving higher server consolidation and more agile dynamic resource provisioning than is possible in traditional platforms [16] [17]. The consolidation of multiple servers and their workloads has an objective of minimizing the number of resources, e.g., computer servers, needed to support the workloads. In addition to reducing costs, this can also lead to lower peak and average power requirements. Lowering peak power usage may be important in some data centers if peak power cannot easily be increased [18]. Server consolidation is particularly important when user workloads are unpredictable and need to be revisited periodically. Whenever a user demand change, VMs can be resized and migrated to other physical servers if necessary [19].

III. POWER USAGE & MANAGEMENT STRATEGIES FOR DATACENTER RESOURCES

The Green Grid [20] a global association of IT companies and professionals trying to improve energy efficiency in data centre, they defined Power Usage Efficiency (PUE) as, measuring the efficiency of a data centre in terms of its electricity use. Its objective is to compare how much power is being used for useful Computing and how much is required for infrastructure.

A data centre’s PUE is the ratio of total power consumed by the facility to power used by the computing equipment.

PUE – Total Facility Power / IT Equipment Power

We have surveyed various power management strategies and task consolidation for datacenter resources like server, network component, virtual network etc. which are concluded all in following table.

Table 1: Power management strategies and task consolidation methodology

Scope	Strategy	Metrics	Reference	Description
Server	Consolidating tasks amongst virtual clusters	CPU usage & Network latency	Hsu et al. [21]	To ration CPU utilization Controlling CPU utilization below threshold point, consolidating cloud workload among virtual clusters.

Server	DCD	CPU, RAM ,Network ,Disk	D. Meisner [22]	PowerNAP - Min power, min performance loss
Server	Dynamic Voltage & frequency scaling	Server voltage	Horvath et al. [23]	Dynamically adjust the server voltages to minimize the total system power consumption, while also meeting end-to-end delay constraints in a multi-tier web service environment
Network	VM assignment & Network traffic	Network traffic	Lin Wang[24]	It reduces the number of active switches and balance traffic flows, depending on the relation between power consumption and routing, to achieve energy conservation.
Network	Routing	Idle network devices and link	Mingwei Xu et al. [25]	Throughput-guaranteed power-aware routing. Idea is to use as little network power as possible to provide the routing service, without significantly compromise on the network performance. The idle network devices and links can be shut down or put into the sleep mode for power saving.
Network	Virtual network implant	Bandwidth	Botero et al. [26]	Allocates the set of virtual networks to the reduced number of physical network elements, then unutilized switches and links are turned off
Server and Network	Server load consolidation with network idle logic	CPU and switch idle mode	Tran Manh Nam et al. [27]	Idle logic approach allows reducing power consumption of network devices by rapidly turning off sub-components when no activities are performed, and by re-waking them up when the system receives new activities
Server and Network	VM consolidation, green routing, sleep mode	Traffic rate	Fang et al. [28]	Optimizes VM placement and traffic flow routing by sleep scheduling network elements

IV. DYNAMIC VIRTUAL MACHINE CONSOLIDATION

In order to reduce the energy consumption we need to consolidate the workload of VMs which requires to identify the overloaded and under loaded servers in data centers then selection of virtual machines from overloaded servers and last have to place them properly on others servers. For under load and overload detection various methods were proposed. Following are the methods for overload/under load detection.

A. Overload detection algorithm

- Static Threshold policy(ST)
- Median Absolute Deviation(MAD)
- Local Regression(LR)

Static CPU Utilization Threshold Algorithms: CPU utilization threshold is used to differentiate overloaded and non-overloaded host. It compare the recent CPU utilization of server with the predefined threshold if utilization exceeded it finds host overloaded [29]. An example of static CPU utilization threshold based algorithms is the averaging threshold-based algorithm (THR) [30].
The algorithm calculates the mean of the N current CPU utilization capacity and compares it to the specified threshold. It detects overload if the mean of the n last CPU utilization measurements is higher than the specified threshold.

For the dynamic and unpredictable workloads the fixed values of utilization threshold are not suitable. The algorithm should automatically adjust the utilization threshold depending on workload of the applications.

Adaptive Utilization Threshold Based Algorithms
Based on statistical analysis of historical data of the VMs these algorithms provide auto-adjustment of the utilization thresholds. Set the upper CPU utilization threshold depending on the strength of deviation of the CPU utilization. If deviation is higher, lower the threshold value. If deviation is higher then it will cause SLA violations which reduce SQA parameter.

Example algorithms are Median Absolute Deviation (MAD) and Inter quartile Range (IQR) [31]:

Median Absolute Deviation (MAD): In statistics, the median absolute deviation (MAD) is a robust measure of the variability of a univariate sample of quantitative data. In case of the standard deviation, the distances from the mean are squared, so on average, large deviations are weighted more heavily, and thus outliers can heavily influence it. On the other hand, in case of MAD, the magnitude of the distances of a small number of outliers is irrelevant.

Inter quartile Range (IQR): It is another measure of statistical dispersion. It is also called the mid spread or middle fifty as it equals the difference between the third and first quartiles in descriptive statistics. The IQR is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that separate parts are called the first, second, and third quartiles; Adaptive utilization threshold algorithms are more robust than static CPU utilization threshold algorithms in case of dynamic environments. But, unfortunately, they provide poor prediction of host overloading.

Regression Based Algorithms: This algorithms estimate the future expectation of CPU utilization. The algorithm based on local regression produces better results than its robust modification. It outperform than adaptive and CPU threshold utilization algorithms. They account performance matrices like energy, SLA violation, SLA violation per active host and performance degradation due to migrations. Example algorithms include Local Regression algorithms (LR) [32] such as Loess method and Local Regression Robust (LRR) [33], which is a modification of LR robust to outliers.

B. VM Selection Algorithms

Virtual machine selection many methods are proposed which more focused on migration time which are

- Minimum Migration Time (MMT)
- Random Choice Policy (RC)
- Maximum Correlation Policy (MC)

Techniques That Use Fixed Criteria

They employ a fixed criterion for decision-making so they are not suitable in dynamic environments. Some examples of those techniques are presented.

The Minimum Migration Time Policy (MMT)

In this method minimum migration time parameter is considered for VM selection. The migration time is calculated as the amount of RAM used by VM divided by the network bandwidth available for the host [31], [34], [35]. The minimization of the VM migration time is more important than the minimization of the correlation between VMs allocated to a host and the resources

The Random Choice Policy (RC): VM selection for the migration is based on uniformly a uniformly distributed discrete random variable X d= U(0, |Vj |), whose values index a set of VMs Vj allocated to a host j. [31], [34].

The Maximum Correlation Policy (MC): This policy works on correlation between the resources usages by workload running on an oversubscribed server, by higher the probability of the server overloading. To estimate the correlation between CPU utilization with VMs, the multiple correlation coefficients are considered. The multiple correlation coefficients are used in multiple regression analysis to assess the quality of the prediction of the dependant variable. It corresponds to the squared correlation between the predicted and the actual values of the dependant variable. It can also be interpreted as the proportion of the variance of dependant variable explained by the independent variables [30] [31].

MMT [38] and MC [31] employ a fixed criterion for decision-making so they are not suitable for decision-making in dynamic environments.

Techniques That Apply Multiple Criteria for Selecting VMs: The VM selection task is considered as a dynamic decision making task. An example of these techniques is VM selection using Fuzzy Q-Learning (FQL).

VM Selection using Fuzzy Q-Learning (FQL): Online dynamic decision making approach for VM selection. It is able to integrate multiple VM selection criteria to benefit from all advantages and possible synergetic contributions of them in long term learning. It is used to choose VM selection strategies from a set of possible strategies to get the best results in VM selection as compare to individual VM selection approach which approve the energy performance trade-off [37].

C. VM Placement Algorithms

VM Placement plays a very crucial role for energy saving because to place the virtual machine at proper server is very essential. Consolidation of virtual machine balance the servers workload in data center and put the under loaded servers on idle mode. Main traditional approach for VM placement is

- Bin packing problem
- Best Fit Decreasing (BFD)
- Constrain programming
- Stochastic integer programming
- Genetic algorithm

This sub-section provides a discussion and comparison among different VM placement algorithms [39]:

A Bin packing approach: The Bin packing approach consists of a series of items having sizes specified in the interval (0, 1]. The items need to be packed into least possible number of bins with capacity one. It is used in resource allocation algorithm, each item as a Virtual Machine (VM) to be packed in minimum number of bins, each considered as a Physical Machine (PM). The bin packing problem is NP hard Heuristic algorithm which is useful in dynamic VM placement where the demand is highly variable. It is useful when all physical machines have the same amount of processing capabilities and physical machines. It will give optimal solution on given amount time. It can be modeled with constraints also. In Multi-dimensions bin packing algorithm, the dimensions are corresponding amount of memory and number of processing units. [39][4].We can even model bin packing algorithms with constraints.

Constraint programming: In such kind of approach we already have the input data information means we have the information about all virtual machines need. We can extend the number of constraints which take too much time to find the optimal solutions. [40]

Stochastic integer programming: It is mathematical optimization approach in which the future demands and prices of resources are uncertain but their probability distributions are either known or can be estimated. This is the best technique to be used in the case where we have two or more uncertain parameters on which the cost depends. Some VM placement techniques use this approach to predict the suitable VM-PM mapping [41].

Genetic algorithm: This approach works on natural selection of suitable solution from all possible solutions. This heuristic can be called as bin packing extended with additional constraints. It requires better computing power and resources as compared to as compared to bin packing. GA is also useful for specifying VM-VM and VM-PM interference constraints [42].

Mi et al.[43] proposed an adaptive self reconfiguration based Genetic Algorithm Based Approach (GABA) which provides optimal solutions online. Request forecasting module is used to catch up with the changing workload.

V. PROPOSED ENERGY AWARE MODELS AND VARIOUS CONSOLIDATION METHODS

Gaurav Chadha et al[44] LIMO, a lightweight runtime system which dynamically manages the number of running threads of an application for maximizing performance and energy-efficiency. It uses DVFS and variable active core count to run an application efficiently. LIMO observes the progress of threads' along with the usage of shared hardware resources to determine the best number of threads to run and the voltage and frequency level. When there is no resource constraint and many threads can run at lower frequency. But when hardware or software (e.g., lock variables) resources limit parallel performance, very less number of threads are kept active at a higher frequency. LIMO gives an average of 21% performance improvement and a 2x improvement in energy-efficiency on a 32-core system over the default configuration of 32 threads for a set of concurrent applications from the PARSEC suite, the Apache web server, and the Sphinx speech recognition system.

Jordi Guitart et. al[45] proposed an global overload control strategy which is based on control based on secure socket layer (SSL) connection differentiation for secure web applications that brings together dynamic provisioning of platform resources and admission control. When workload increases, the resources are assigned to an application on request while number of new SSL connections will be limited and controlled to avoid the server's performance degradation by admission control mechanism. Servers having self management facilities that adapt to accede load to the assigned resources by using an admission control mechanism. Authors have implemented global resources manager for Linux hosting platform which distributes the available resources among the tomcat application servers running on it. Servers having self management facilities that adapt exceed load to the assigned resources by using an admission control mechanism.

The objective is to handle the overload hence system can remain working in the presence of overload even when the incoming request rate is too much compare to system's capacity. It should also maintain the response time at acceptable levels.

Anton Beloglazov et. al[46] proposed a heuristics energy aware approach which provision data center resources in order to make it sustainable and eco friendly to client applications which improves energy efficiency of data center also maintain Quality of service. Author has proposed (a) architectural principles for energy efficient management of clouds (b) resource allocation policies and scheduling algorithms which consider QoS and power usage characteristics of the device. They have develop autonomic and energy-aware mechanisms for self managing changes in the state of resources effectively and efficiently to satisfy service obligations and achieve energy efficiency. They have validated our approach by conducting a performance evaluation study using the CloudSim toolkit. The outcome shows that cloud computing has immense potential as it offers significant cost savings and demonstrates high potential for the improvement of energy efficiency under dynamic workload scenarios. This work also explores open research challenges in energy-efficient resource management for virtualized Cloud data centers to provide advancements of the state-of-the-art operational Cloud environments.

Nadjia Kara et. al[47] proposes genetic algorithms strategies for optimization in IVR (Interactive Voice Response) (IVR) virtualization task scheduling and computational resource sharing. IVR allows automatic human-computer interactions, via a voice commands or telephone keypad. Its key function is to provide end-users with self-service voice information. In proposed architecture three layers communicate via three planes: service, composition and management. This work more focused on resource management at the substrate layer of management plane. Before a service provider can make an IVR application available to its end-users, he should develop such an application by discovering and (eventually) composing existing substrates. It then activates the application, a phase which includes the instantiation of the substrates required to run the application. It is only after this that the end-users can interact with the application.

F. Farahnakian et. al[48] developed consolidation algorithm which is dynamic in nature to minimize the number of active physical servers in a data center in order to reduce energy cost. The algorithm utilizes the method of k-nearest neighbor regression algorithm which predicts the resource usage in each server for overutilization and underutilization purpose. The KNN also used to predict future resources need based on historical data about resource usage. By using prediction method it improves dynamic VM consolidation performance in data center which migrate VMs appropriate with the recent and future requests. VM selection consider MMT (minimum migration time) policy and VM allocation consider modified power aware best fit decreasing with SLA parameter.

Y. Wu et. al [49] proposed an optimized simulated annealing algorithm for virtual machine consolidation to influence the cost and performance. The configuration of the system, the function to obtain new configuration, the function for obtaining new configuration, the objective function for optimization problem are used in this simulated annealing algorithm. This proposed work to decrease SLA violation and reserve some resources of physical machine in order to react on decreasing random demands in nearest future. It has potential to replace First Fit Decreasing or combine with FFD to generate a better VM placement. SAVPM algorithm works on temperature based parameter, which used random search method to choose PM solution until it finds suitable one for feasible solution. It provides better solution when capacity index is larger and search space is smaller. SAVMP can generate better VM assignment then FFD, with 0-25 % more energy saving FFD.

OpenStack is one of the largest open-source cloud computing middleware development communities. The Yu Huanle1 et. al[50] proposed Openstack cloud platform resource scheduling mechanism. The work has improved the Openstack cloud platform stability. The framework divides the physical hypervisor host into multiple logical resource pools. The primary scheduling policies implement performance isolation and security isolation effectively, and the on-going optimization policies optimize resource utilization, ensure load balancing, improve the user experience by monitoring system load change and migrating virtual machines during system runtime. The results state that the multi-resource pool resource optimization scheduling optimization framework this article is good for OpenStack resource optimization, performance isolation and security isolation, and has a good extension.

In this work Moreno Marzolla et al. [51] presented V-Man, a full decentralized algorithm for consolidating VMs in cloud datacenters. V-Man is derived from simple gossip protocol which can operate on any arbitrary initial allocation of VMs on the cloud does not require central coordinator which iteratively producing new allocations that quickly converge towards the one maximizing the number of idle hosts. V-Man executed periodically to maximize to achieve efficiency, scalability and robustness to failures.

V-man is executed periodically to identify a new arrangement of existing VM instances hence the number of idle server is improved. After new allocation is identified then migrate the VM monitors Xen1, OpenVZ2 and VMware3. It produced optimal VM placements in few rounds, also showed good scalability. This is also robust which can cope with computing nodes added or removed in cloud data centers.

N. Kord et. al[52] proposed a Minimum co-relation coefficient (MCC) for virtual machine placement in virtualized datacenters. This approach is related to service level agreement (SLA) and lower energy consumption using power aware best fit decreasing algorithm and minimum coefficient concept for VM assignment. Here VM selection performance based on fuzzy Analysis Hierarchy Process (AHP) with depend statistical analysis on historical data.

Seyed Saeid et. al [53] proposed a virtual machine selection approach with fuzzy Q-learning which takes online decision to improve energy efficiency in cloud environments. For better result**,** an adaptive and predictive mechanism can be efficient to make decisions about which criterion can achieve best result in current state.

It finds an optimal strategy to applying multiple criteria for selecting VMs during a dynamic VM consolidation procedure. This distributed dynamic VM consolidation is divided into four decision making tasks: (1) Host overload detection (2) VM selection (3) Host under load detection (4) VM placement. The multi agent system architecture employs cooperative learning strategy to increase learning convergence rates and as a result better dynamic decision. The performance is outstanding compare to static approach also it can be implemented in a real life cloud environment with open source cloud management framework like Openstack neat and snooze.

Reinforcement learning based dynamic consolidation method (RL-DC) was proposed by F. Farahnakian et.al [54] to minimize the number of active host according to the current resource requirement. Reinforcement learning is a machine learning approach which is used here where agent precepts the surrounding environment and selects an action at each state. Agent learns from past knowledge which takes decision about when to switch to the sleep or active mode. RL-DC needs prior information about workload to dynamically adapt to environment. Agent collects feedback about quality of action. RL method takes intelligent decision to switch a host into the active or sleep power mode. The learning agent collects host power mode detection policy through Q-learning of RL. In Q-learning provides a self optimizing controller design without a prior knowledge of the environment.

Relocation of Virtual Machines need careful planning otherwise it will increase the run-time overheads and increase the consumption of energy. The relocation problem proposed by Yufan Ho et. al[55] is a modified bin packing problem and proposed a new consolidation algorithm with bounded cost of relocation. With compare to other algorithm like First Fit and Best fit it works better. Result gives trade-off between server consolidation quality and relocation cost. The information saving, transferring, and restoring invite huge runtime overheads. It does not relocate a large portion of the virtual machines because of the huge overhead, and extra power consumption, in relocating them.

The work trade about 1% in server consolidation quality for a reduction about 50% in relocation cost, when compared with other well known bin packing algorithms. In dynamic bin packing model, Objective is not only need to reduce the number of bin used, but also the amount of virtual machines. We need to relocate from one bin to another. These concerns complicate the bin packing problem greatly small fraction of virtual machines, and still maintain performance guarantee after the relocation.

VM consolidation problem for MapReduce enabled computing clouds is proposed by Zhe Huang et. al [56] to reduce energy consumption. In this Work two resource allocation methods and also related service level agreement models designed for MapReduce and non-MapReduce instances. The VM consolidation problem is modelled as integer nonlinear optimization problem in MapReduce enabled cloud to optimize the solution. MapReduce jobs are used by homogeneous MapReduce VM instances that have identical hardware resource. The SLA model for MapReduce instances reserve I/O bandwidth from physical servers and distribute it evenly among all the MapReduce instances. This work proved that rapid improvement in VM consolidation is achieved by configuring the MapReduce instances them with the non-MapReduce instances. The consolidation process for the MapReduce instances is achieved by fine tuning the number of MapReduce instances created in order to find a good balance between the OS overhead (i.e.,) and the quantization error so that the aggregated I/O bandwidth required is minimized.

To solve the high energy consumption problem, an energy-efficient virtual machine consolidate algorithm named Prediction-based VM deployment algorithm for energy efficiency (PVDE) was presented by Zhou et.al [57]. To classifies the hosts in the data center the linear weighted method was utilized and predict the host load. They performed high performance analysis. In their work, the algorithm reduces the energy consumption and maintains low service level agreement (SLA) violation when compared to other energy saving algorithms in the experimental result.

Li et.al [58] presented an elaborate thermal model to address the complexity of energy and thermal modeling of realistic cloud data center operation that analyzes the temperature distribution of airflow and server CPU. To minimizing the total datacenter energy consumption the author presented GRANITE - a holistic virtual machine scheduling algorithm. The algorithm was evaluated against other existing workload scheduling algorithm IQR, TASA and MaxUtil and Random using real cloud workload characteristics extracted from Google datacenter trace log. The GRANITE consumes less total energy and reduces the critical temperature probability when compared with existing that demonstrated in result.

A new scheduling approach named Pre Ant Policy was introduced by Hancong Duan et.al [59]. Based on fractal mathematics their method consists of prediction model and on the basis of improved ant colony algorithm (ABC) a scheduler. To trigger the execution of the scheduler by virtue of load trend prediction was determined by prediction model and under the premise of guaranteeing the quality-of-service, for resource scheduling the scheduler is responsible while maintaining energy consumption. The performance results demonstrate that their approach exhibits resource utilization and excellent energy efficiency.

In order to elevate the trade-off between energy consumption and application performance Rossi et al. [60] presented an orchestration of different energy savings techniques. They implemented Energy-Efficient Cloud Orchestrator-e-eco and by using scale-out applications on a dynamic cloud in a real environment the infrastructure test were carried out to evaluate e-eco. Their evaluation result demonstrates that e-eco was able to reduce the energy consumption. When contrasted to the existing power-aware approaches the e-eco achieved the best trade-off between performance and energy savings.

In cloud computing for energy saving, a three dimensional virtual resource scheduling method (TVRSM) was introduced by Zhu et.al [61]. For the cloud data center they

build the resource and dynamic power model of the PM in their work. There are three stages of virtual resource scheduling process as follows; virtual resource optimization, virtual resource scheduling and virtual resource allocation. For different objective of each stage they design three different algorithms respectively. The TVRSM can effectively reduce the energy consumption of the cloud data center when compared with various traditional algorithms.
For the dynamic consolidation of VMs in cloud data centers, Khoshkholghi et.al [62] has exhibited several novel algorithms. Their objective is to reduce energy consumption and improve the computing resources utilization under SLA constraints regarding bandwidth, RAM and CPU. By conducting extensive simulation the efficiency of their algorithm is validated. While providing a high level of commitment their algorithm significantly reduces energy consumption. When compared to the benchmark algorithms, the energy consumption can reduce by up to 28% and SLA can improved up to 87% based on their algorithms.

Boominathan Perumal et.al [63] has proposed method to apply firefly colony and fuzzy firefly colony optimization algorithms to solve the problems of server consolidation and multi objective virtual machine placement problem. The firefly approach which rely on ant colony optimization algorithm where fireflies are collaborating learning agents. This firefly colony algorithm works fast as pair-wise assessment of flies. Fireflies are simple insects living in groups. The firefly flashes act as a signal system to attract (communication) other files. The firefly algorithm is based on these flashing patterns and the behavior of fireflies. The randomization capacity of the basic FA is gradually decreased as it reaches the optima; the performance of the FCO is improved. The hybridization of greedy meta-heuristic algorithm and the exploitation and exploration process of firefly algorithm. The proposed work is distributed, autocatalytic and constructive greedy meta-heuristic. The decision-making is based on the probabilistic choice which is biased by the concentration of the phosphorescent glowing in its path.

When Host machine is overloaded so the selection of virtual machine for migration is needed where Monil et.al [64] explored a fuzzy logic approach for handling uncertain, imprecise, or un-modeled data in solving control and intelligent decision making problems. They innovated a new VM selection method which provides a balanced and best result in quality of service. The previous methods can be judged and rules of inference can be made to generate a result. It considers all the options and depending on those a fuzzy value will be generated based on the predetermined rules of inferences. The fuzzy system for VM selection is designed and developed three metrics; the membership functions and inference rules based on the real cloud workload data. The novel fuzzy VM selection method implemented in cloudsim toolkit and compares the performance with the existing methods where results of simulation balanced performance by trading off power and QoS parameters.

Mohammad Alaul Haque Monil [65] has proposed VM selection with migration control algorithm which is based on fuzzy logic and heuristic based virtual machine consolidation approach to achieve energy Qos balance. Also mean, median and standard deviation based over load detection algorithms is designed in this work. In addition work also explores migration control in fuzzy VM selection method that will improve the performance. The algorithm portrays the hosts are created and then VM data is taken as input. Based on the real life data of VM, the cloudlets are created then VMs are assigned to host and cloudlet is assigned to VM. Status is checked for every scheduled interval. For very scheduled interval, under load detection algorithm is executed and less utilized hosts are put into sleeping mode by transferring all VM to other active VM. Then overload detection is executed and overload hosts are identified. Proposed work is simulated in Cloudsim and checked the performance on real world work load traces of Planet lab VMs where method is more energy efficient.

Ant Colony Optimization (ACO) is a meta-heuristic inspired by the observation of real ant colonies and based upon their collective foraging behavior. The work presented by Yongqiang Gao et. al [66] multi-objective ant colony algorithm for VM Placement with objective to obtain non-dominated solutions which parallel minimize total resource utilization and energy consumption. Multi-objective evolutionary algorithms use population based method which are random or stochastic optimization methods to get pareto optimal solutions. The VM placement is a permutation of VM assignment. The algorithm works in two phases. (i) Initialization phase where all parameters are initialized (ii) iterative part each ant receives all VM requests and physical machines has been assigned with VM. In order to achieve this use of a pseudo-random-proportional rule, which explain the need of an ant to choose a particular VM as the next one to pack into its current host. The rule is about the current pheromone concentration on the movement and a heuristic which helps the ants towards selecting the most promising VMs. The proposed work is better than famous algorithms like bin packing and max-min ant systems.

In order to reduce the problem of energy-performance trade-off and power consumption Seyed Ebrahim Dashti et. al [67] proposed a modified particle swam optimization to assign the live migrated virtual machines in the overloaded host also under loaded nodes are consolidated and powered off which save the power consumption. The main work is workflow scheduling in Iaas which is energy efficient and keep the balance between the resource consumption and quality of service to achieve more benefit in our private data network which provides better SLA. The scheduling also uses heuristic method in fitness function. Dynamic provisioning and allocation strategies are needed to regulate the internal settings of the cloud to address oscillatory peaks and off peaks of the workload. Simulation of proposed approach was checked in CloudSim. Simulation results of proposed approach save 14 % more energy and the number of migrations and simulation time.

Shangguang Wang et. al [68] proposed an approach for minimizing the energy consumption is based on particle swarm optimization (PSO) which is search algorithm that is based on swarm intelligence for heterogeneous servers of datacenters. Traditional PSO is not used for discrete optimization. So need to redefine the parameters in this work. It has many commonalities between evolutionary techniques like genetic algorithms but PSO is easy to implement with fewer parameters. Here in this work the parameters and operators of PSO are redefined. After that work adopted an energy-aware local fitness first strategy to update particle position to improve the problem solving efficiency. Also a novel two-dimensional particle encoding scheme is developed. The work simulated which improved the placement of VM and reduce the energy consumption which outperform than other work.

VI. CONCLUSIONS

In cloud computing datacenter’s resources consumes a large amount of energy. In order to reduce the power consumption and heat generation for green computing VM consolidation in data center plays a crucial role for resource usage optimization where overloaded hosts migrate the VMs to proper host servers also the under-loaded host to be put in idle mode to reduce the power consumption. In our survey paper many power management strategies, load balancing and dynamic VM consolidation algorithms like genetic algorithms, heuristic algorithms and fuzzy logic algorithms are described and compared proposed by researchers. The fuzzy logic and heuristic algorithms performance in load balancing, VM consolidation is better than the traditional static algorithms for VMs overload detection, VM selection and VM placement in cloud data center’s servers.

REFERENCES

[1] Vivek Raich, Pradeep Sharma, Shivlal Mewada and Makhan Kumbhka “Performance Improvement of Software as a Service and Platform as a Service in Cloud Computing Solution”, International Journal of Scientific Research in Computer Science and Engineering,Vol.1, Issue-6, pp.13-16, Dec 2013.

[2] Karthik Kumar and Yung-Hsiang Lu, “Cloud Computing For Mobile Users: Can Offloading Computation Save Energy”, Published by the IEEE Computer Society, 2010.

[3] Nabil Sultan, “Cloud computing for education: A new dawn”, International Journal of Information Management, vol. 30, pp. 109–116, 2010.

[4] Fatma A. Omara, Sherif M. Khattab and Radhya Sahal, “Optimum Resource Allocation of Database in Cloud Computing”, Egyptian Informatics Journal, vol. 15, pp.1–12, 2014.

[5] George Pallis, “Cloud Computing The New Frontier of Internet Computing”, Published by the IEEE Computer Society, 2010.

[6]Saurabh Kumar Garg, Steve Versteeg and Rajkumar Buyya, “A framework for ranking of cloud computing services”, Future Generation Computer Systems, vol. 29, pp. 1012–1023, 2013.

[7]S. Wang, A. Zhou, C. Hsu, X. Xiao and F. Yang, "Provision of Data-Intensive Services Through Energy- and QoS-Aware Virtual Machine Placement in National Cloud Data Centers", IEEE Transactions on Emerging Topics in Computing, vol. 4, no. 2, pp. 290-300, 2016.

[8]M. Abdullahi, M. Ngadi and S. Abdulhamid, "Symbiotic Organism Search optimization based task scheduling in cloud computing environment", Future Generation Computer Systems, vol. 56, pp. 640-650, 2016.

[9] Ajay Jangra and Renu Bala “Spectrum of Cloud Computing Architecture: Adoption and Avoidance Issues”, International Journal of Computing and Business Research, Vol. 2, Issue 2, May 2011.

[10] S. Energy “Report to congress on Server and data center energy efficiency public law 109-431”, public Law Vol. 109, p.431, 2007.

[11] Faiza Fakhar, Barkha Javed, Raihan ur Rasool, Owais Malik and Khurram Zulfiqar, “Software level green computing for large scale systems”, Journal of Cloud Computing: Advances, Systems and Applications, vol. 1, no. 4, 2012.

[12] Robert Basmadjian, Hermann De Meer, Ricardo Lent and Giovanni Giuliani, “Cloud computing and its interest in saving energy: the use case of a private cloud”, Journal of Cloud Computing: Advances, Systems and Applications, vol. 1, no. 5, 2012.

[13] Lingjia Tang, Mary Lou Soffa and Jason Mars, “Directly Characterizing Cross Core Interference through Contention Synthesis”, ACM 2011.

[14] Gaurav Dhiman, Giacomo Marchetti and Tajana Rosing, “vGreen: A System for Energy-Efficient Management of Virtual Machines”, ACM Transactions on Design Automation of Electronic Systems, Vol. 16, No. 1, 2010.

[15] Xiangzhen Kong, ChuangLin, YixinJiang, WeiYan and XiaowenChu, “Efficient dynamic task scheduling in virtualized datacenters with fuzzy prediction”, Journal of Network and Computer Applications, 2010 (Elsevier).

[16]Andreas Merkel, Jan Stoess, Frank Bellosa, “Resource-conscious scheduling for Energy Efficiency on Multicore Processors”, EuroSys '10 Proceedings of the 5th European conference on Computer systems, pp.153-166, 2010.

[17] Sriram Govindan, Jeonghwan Choi, Arjun R. Nath, Amitayu Das, Bhuvan Urgaonkar, Member, Anand Sivasubramaniam, “Xen and Co.: Communication-Aware CPU Management in Consolidated Xen-Based Hosting Platforms”, IEEE Transactions On Computers, vol. 58, no. 8, 2009.

[18] George Kousiourisa, Tommaso Cucinottab and Theodora Varvarigoua, “The effects of scheduling, workload type and consolidation scenarios on virtual machine performance and their prediction through optimized artificial neural networks”, The Journal of Systems and Software, vol. 84, pp. 1270– 1291, 2011.

[19] Corentin Dupont, Thomas Schulze, Giovanni Giuliani, Andrey Somov and Fabien Hermenier, “An Energy Aware Framework for Virtual Machine Placement in Cloud Federated Data Centres”, ACM, 2012.

[20]S.Greenberg, E. Mills,B. Tschudi, P. Rumsey and B.Myatt, “Best Practices for Data Centers : Lessons Learned from Benchmarking 22 Data Centers”, Proceedings of the ACEEE Summer Study on Energy Efficiency in Buildings in Asilomar, CA. ACEEE, August, vol. 3, pp. 76–87256–259, Dec. 2013, 2006.

[21] C.H. Hsu, K. Slagter, S.C. Chen, and Y.C. Chung, "Optimizing Energy Consumption with Task Consolidation in Clouds", Information Sciences, 2012.

[22] D. Meisner, B. Gold, and T. Wenisch, "PowerNap: eliminating server idle power", ACM SIGPLAN Notices, vol. 44, no. 3, pp. 205–216, 2009.

[23] T. Horvath, T. Abdelzaher, K. Skadron, and X. Liu, "Dynamic voltage scaling in multitier web servers with end-to-end delay control", Computers, IEEE Transactions on, vol. 56, no. 4, pp. 444– 458, 2007.

[24] Lin Wang, Fa Zhang, Jordi Arjona Aroca and Athanasios V. Vasilakos "GreenDCN: a General Framework for Achieving Energy Efficiency in Data Center Networks", IEEE Journal on selected areas in communications, January 2014.

[25] Mingwei Xua, Yunfei Shang and Dan Lia, Xin Wang "Greening Data Center Networks with Throughput-guaranteed Power-aware Routing", ACM SIGCOMM Workshop on Green Networking 2010.

[26] J. F. Botero, X. Hesselbach, M. Duelli and D. Schlosser, A. Fischer, and H. De Meer, "Energy efficient virtual network embedding,", Communications Letter, IEEE, vol. 16, pp. 756-759, 2012.

[27] Tran Manh Nam, Nguyen Huu Thanh and Doan Anh Tuan,"Green data center using centralized power-management of network and servers" International Conference on Electronics, Information, and Communications (ICEIC), IEEE pp. 1- 4, 2016.

[28] W. Fang, X. Liang, S. Li, L. Chiaraviglio, and N. Xiong "VMPlanner: Optimizing virtual machine placement and traffic flow routing to reduce network power costs in cloud data centers", Computer Networks, 2012.

[29] A. Beloglazov, "Energy-Efficient Management of Virtual Machines in Data Centers for Cloud Computing", Ph.D. thesis, The University of Melbourne, 2013.

[30] A. Beloglazov and R. Buyya, "OpenStack Neat: A Framework for Dynamic and Energy-Efficient Consolidation of Virtual Machines in OpenStack Clouds", Concurrency and Computation: Practice and Experience (CCPE), pp. 32–36, 2014.

[31] A. Beloglazov and R. Buyya, "Optimal Online Deterministic Algorithms and Adaptive Heuristics for Energy and Performance Efficient Dynamic Consolidation of Virtual Machines in Cloud Data Centers", Concurrency and Computation: Practice and Experience (CCPE), vol. 24, no. 13, pp. 1397–1420, 2012.

[32] W. Cleveland and C. Loader, "Smoothing by Local Regression: Principles and Methods", Statistical Theory and Computational Aspects of Smoothing, 1996.

[33] W. Cleveland, "Robust Locally Weighted Regression and Smoothing Scatterplots", Journal of the American Statistical Association, 1979.

[34]] A. Singh and S. Kinger, "Virtual Machine Migration Policies in Clouds", International Journal of Science Research, vol. 2, no. 5, pp. 364–367, 2013.

[35] A. Beloglazov and R. Buyya, "Adaptive Threshold-Based Approach for Energy-Efficient Consolidation of Virtual Machines in Cloud Data Centers", Proceedings of the 8th International Workshop on Middleware for Grid, Clouds and E-science, Bangalore, India, 2010.

[36] A.verma, G. Dasgupta, T. Nayak and R.Kothar "Server Workload Analysis for Power Minimization Using Consolidation", Proc. the 2009 USENIX Annual Technical Conference, San Diego, USA, 2009.

[37] S. Masoumzadeh and H. Hlavacs "Integrating VM Selection Criteria in Distributed Dynamic VM Consolidation Using Fuzzy Q-Learning", Proc. the 9th International Conference on Network and Service Management (CNSM 2013), pp. 332– 338, Oct. 2013.

[38] A. Beloglazov, J. Abawajy and R. Buyya "Energy-Aware Resource Allocation Heuristics for Efficient Management of Data Centers for Cloud Computing", Future Generation Computer Systems, vol. 28, no. 5, pp. 755–768, 2012.

[39] P. Sayeedkhan, "Virtual Machine Placement Based on Disk I/O Load in Cloud", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4), pp. 5477-5479, 2014.

[40] Y. Yu and Y. Gao, "Constraint Programming-Based Virtual Machines Placement Algorithm in Datacenter", Intelligent Information Processing, pp. 295–304, 2012.

[41] B. B. Nandi, A. Banerjee, S. C. Ghosh and N. Banerjee, "Stochastic VM Multiplexing for Datacenter Consolidation", IEEE Ninth International Conference on Services Computing, pp. 114–121, Jun. 2012.

[42] G. Wu, M. Tang, Y. Tian and W. Li "Energy-efficient virtual machine placement in data centers by genetic algorithm", Neural Information Processing, pp. 315–323, 2012.

[43] Mi, H., Wang, H., Yin, G., Zhou, Y., Shi, D. and Yuan, L "Online self-reconfiguration with performance guarantee for energy efficient large-scale cloud computing data centers", In Services Computing (SCC), IEEE International Conference on pp. 514-521, 2010

[44] Gaurav Chadha, Scott Mahlke and Satish Narayanasamy, "When Less Is MOre (LIMO): Controlled Parallelism for Improved Efficiency", ACM, 2012.

[45] Jordi Guitart, David Carrera, Vicenc Beltran, Jordi Torres and Eduard Ayguade, "Dynamic CPU provisioning for self-managed secure web applications in SMP hosting platforms", Computer Networks, vol. 52,pp. 1390–1409, 2008.

[46] Anton Beloglazov, Jemal Abawajyb and Rajkumar Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing", Future Generation Computer Systems, vol. 28, pp. 755–768, 2012.

[47] Nadjia Kara, Mbarka Soualhia, Fatna Belqasmi, Christian Azar and Roch Glitho, "Genetic-based Algorithms for Resource Management in Virtualized IVR Applications", Journal of Cloud Computing, vol. 3, no.15, 2014.

[48] F. Farahnakian, T. Pahikkala, P. Liljeberg, and J. Plosila, "Energy Aware Consolidation Algorithm Based on K-Nearest Neighbor Regression for Cloud Data Centers", IEEE/ACM 6th Int. Conf. Util. Cloud Comput., pp..

[49] Y. Wu, M. Tang, and W. Fraser, "A simulated annealing algorithm for energy efficient virtual machine placement", IEEE Int. Conf. Syst. Man, Cybern., pp. 1245–1250, Oct. 2012.

[50] Yu Huanle and Shi.Weifeng "An OpenStack-based resource optimization scheduling framework", IEEE 6th International Symposium on Computational Intelligence and Design, Oct. 2013.

[51] Moreno Marzolla and Ozalp Babaoglu "Server Consolidation in Clouds through Gossiping" IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, June 2011.

[52]N. Kord and H. Haghighi, "An energy-efficient approach for virtual machine placement in cloud based data centers", 5th Conf. Inf. Knowl. Technol., pp. 44–49, May 2013.

[53]S. S. Masoumzadeh and H. Hlavacs, "Integrating VM selection criteria in distributed dynamic VM consolidation using Fuzzy Q-Learning", Proc. 9th Int. Conf. Netw. Serv. Manag. (CNSM 2013), pp. 332–338, Oct. 2013.

[54] F. Farahnakian, P. Liljeberg, and J. Plosila, "Energy-Efficient Virtual Machines Consolidation in Cloud Data Centers Using Reinforcement Learning", 22nd Euromicro Int. Conf. Parallel, Distrib. Network-Based Process. pp. 500–507, Feb. 2014.

[55] Yufan Ho "Server Consolidation Algorithms with Bounded Migration Cost and Performance Guarantees in Cloud Computing" in Fourth IEEE International Conference on Utility and Cloud Computing pp. 155-161, 2011.

[56] Zhe Huang, Danny H.K.”A Virtual Machine Consolidation Framework for MapReduce Enabled Computing Clouds”, ITC 2012.

[57] Z. Zhou, Z. Hu, J. Yu, J. Abawajy and M. Chowdhury, “Energy-efficient virtual machine consolidation algorithm in cloud data centers”, Journal of Central South University, vol. 24, no. 10, pp. 2331-2341, 2017.

[58]X. Li, P. Garraghan, X. Jiang, Z. Wu and J. Xu, “Holistic Virtual Machine Scheduling in Cloud Datacenters towards Minimizing Total Energy”, IEEE Transactions on Parallel and Distributed Systems, vol. 29, no. 6, pp. 1317-1331, 2018.

[59] H. Duan, C. Chen, G. Min and Y. Wu “Energy-aware scheduling of virtual machines in heterogeneous cloud computing systems”, Future Generation Computer Systems, vol. 74, pp. 142-150, 2017.

[60] F. Rossi, M. Xavier, C. De Rose, R. Calheiros and R. Buyya, “E-eco: Performance-aware energy-efficient cloud data center orchestration”, Journal of Network and Computer Applications, vol. 78, pp. 83-96, 2017.

[61]W. Zhu, Y. Zhuang and L. Zhang “A three-dimensional virtual resource scheduling method for energy saving in cloud computing”, Future Generation Computer Systems, vol. 69, pp. 66-74, 2017.

[62] M. Khoshkholghi, M. Derahman, A. Abdullah, S. Subramaniam and M. Othman, “Energy-Efficient Algorithms for Dynamic Virtual Machine Consolidation in Cloud Data Centers”, IEEE Access, vol. 5, pp. 10709-10722, 2017.

[63] Boominathan Perumal, Aramudhan Murugaiyan “A Firefly Colony and Its Fuzzy Approach for Server Consolidation and Virtual Machine Placement in Cloud Datacenters”, Hindawi Publishing Corporation Advances in Fuzzy Systems, February 2016.

[64] Monil and Mohammad Rahman “Fuzzy logic-based VM selection strategy for cloud environment”, International Journal Cloud Computing, Vol. 6, No. 2, 2017.

[65] Mohammad Alaul Haque Monil, Rashedur M. Rahman “VM consolidation approach based on heuristics, fuzzy logic, and migration control”, Journal of Cloud Computing: Advances, Systems Applications, pp.5-8, 2016.

[66] Yongqiang Gao, Haibing Guan, Zhengwei Qi, Yang Houb and Liang Liu “A multi-objective ant colony system algorithm for virtual machine placement in cloud computing”, Journal of Computer and System Sciences , pp.1230–1242 ,March 2013.

[67] Seyed Ebrahim Dashti, Amir Masoud Rahmani “Dynamic VMs placement for energy efficiency by PSO in cloud computing”, Journal of Experimental & Theoretical Artificial Intelligence, 28:1-2, pp. 97-112.

[68] Shangguang Wang, “Particle Swarm Optimization for Energy-Aware Virtual Machine Placement Optimization in Virtualized Data Centres”, published in IEEE ICPADS October 2013.

Authors Profile

Mr.Rajesh Patel pursed Bachelor of Computer Engineering from HNGU University of Gujarat, India in 2005 and Master of Computer Science and Engineering from Nirma University in year 2011. He is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Computer Engineering, GPERI, Mehsana since 2012. He is a member of CSI since 2013. He has published more than 09 research papers in reputed international journals and conferences including IEEE which are available online. His main research work focuses on Computer Network, Sensor Network and Cloud Computing. He has 13 years of teaching experience and 5 years of Research Experience.