CHAPTER 4

ADAPTIVE LOAD BALANCING APPROACH IN CLOUD INFRASTRUCTURE FOR ENERGY CONSUMPTION

4.1. INTRODUCTION

Cloud computing infrastructure of gained a wide knowledge due to the benefits of high scalable services with massive computation power and storage capability. Even the resources are being provided as service to the clients by the cloud environment. The service provided by the cloud environment definitely ensures Service Level Agreement (SLA). The major issue in cloud infrastructure convinces the user needs regarding resource requests with appropriate energy utilization. The needs of the subscribers are grown to an extent, like expecting a wide energetic platform for load balancing, even if the resources are shared. Further, cloud computing scenario needs to optimally balance the load at the middle of the servers in order to avoid hotspot and to improve its resource utility.

In order to perform energy conservation in cloud infrastructures, the exploit of the chronological traffic data from data centers is used efficiently with a service request prediction model. Collaborative provable data possession scheme has adopted Homomorphic verifiable responses and hash index hierarchy to offer better resource service. However, the drawback is that the match index structure is not matched properly with the clustering model. The different levels of power tariffs and requests made to the servers affect the decisions like where to serve the cluster needs. SLA Laws on privacy comprises an impact that decides whether the loads are moved in or out of a cluster. On the other hand, SLA affects the overall energy consumption.

In this chapter, the proposed work aims to address the high energy utilization problem as well as imbalance load issues, with the development of Adaptive Load Balancing Approach (ALB). ALB Approach balances the load from every cluster group by minimizing the bandwidth and energy consumption. With repetitive query messaging, ALB gathers information about the recent load of the other group. Moreover, repetitive query messaging forecasts the average energy and bandwidth consumption of each group after information updating. The ALB Approach not only balances the energy consumption, but also enhances the utilization of resources with minimal bandwidth usage. An extensive level of experimental studies is conducted to illustrate the efficiency and effectiveness of the proposed ALB Approach. An experimental evaluation is accepted to estimate the performance of ALB Approach with the Virtual Machine (VM) energy-efficient cloud data centers. the performance metric for evaluation of ALB Approach is measured in terms of energy consumption, bandwidth utilization rate, performance tradeoff, and the response time to the service request.

4.2. CLOUD COMPUTING FOR EFFICIENT BANDWIDTH UTILIZATION AND ENERGY CONSUMPTION

Cloud computing is changing the trends of recent updates and has greatly revised the system the people parse information. On the other hand, the Cloud service provider offers various platforms, permitting enormous level of diversities of terminal devices responsive by individuals to function. The next generation of user devices provides not only constant readiness for operation, but also stable information consumption. In such a situation, computing, information storage and communication become more applicable.

Cloud computing is a successful way to afford mechanisms that include suitable and secure infrastructure, with reduced cost of operations. Cloud computing depends on the data centers as their main backend of users, including computing infrastructure. Cooperative provable data possession scheme adopts the technique of homomorphic verifiable responses and hash index hierarchy. Yan Zhu., and Shanbiao Wang., (2012) [72] reveal that the homomorphic verifiable responses and hash index hierarchy are still a challenging problem in scheduling with the length, irrelevant to the size of the data blocks. Even the energy consumption and load balancing play a vital role in enhancing the successful data transmission.

4.2.1. Load Balancing with Energy Consumption in Cloud

One of the major reasons behind energy inefficiency in data centers is the inactive power, which is unused when the servers operate at low utilization.

Even at a very low load, such as 10% CPU utilization, the power consumed is over 50% of its peak power. At the same time, if the memory, network, or any such resource is given restricted access, the inactive power unused in the other resources will be high. In the cloud computing approach, multiple data center applications are depended on a regular set of servers. This provides consolidation for the application workloads on a smaller number of servers. Lower range of server is managed properly, as the different loads vary in their different resource utilization paths and also they differ in their chronological details. Consolidation, thus permits amortizing the inactive power usage more efficiently. Nevertheless, effective consolidation is not as insignificant as stuffing the maximum workload in the negligible number of servers, keeping each resource like CPU, disk, network, etc on every server at 100% utilization.

Moreover, according to Ching-Hsien Hsu, et al., (2011) [9] execution of consolidation to consume energy utilization while offering sufficient amount is problematic. At first, theconsolidation methods need to suspiciously choose which workloads should be fused on a common physical server. The workload resource usage, performance, and energy usages are not supplementary. Knowing the scenario of their work is thus compulsory to choose which loads can be stuffed together. Then arises the best concert and energy point. This is due to the consolidation cause performance degradation, that results in the execution time so as to enhance the energy consumption from the reduced inactive energy. In addition, the best point varies with acknowledgeable degradation in concert and application mix. Evaluating the best point and extracting the workload variations, thus becomes significant for energy efficient consolidation. Yan Zhu., et al., (2012) [72] point out that, Homomorphic verifiable responses and hash index hierarchy are still challenging problems in resource scheduling. Cooperative provable data possession scheme adopts the technique of homomorphic verifiable responses and hash index hierarchy.

Avinash Mehta., et al., (2011) elaborate that a significant amount of energy is conserved by shifting the VM running an underutilized equipment. A significant amount of energy is conserved as shown shifting theby migrating virtual Machines (VM) running on underutilized equipment to additional machines and hibernating such underutilized machines. VM aims to design such a policy for energy-efficient cloud data centers. It makes use of the chronological traffic data from data centers and uses a service request prediction model. Yan Zhu., et al., (2012) [72] emphasize that, Cooperative Provable Data Possession (CPDP) Scheme is based on homomorphism demonstrable response and hash index hierarchy, but faces certain limits to satisfy the user service requests. Fan Zhanga., et al.,(2013) [16] explain that the scheduling scheme for multi-objective (MOS) is specifically designed and based on the ordinal optimization method for clouds. However, MOS uses different memory and disk requirements, increasing the workload while performing multi-tasking.

Based on the aforementioned techniques and methods, in this work, an Adaptive Load Balancing Approach, that makes the data center to consume less energy by shutting it down or scaling down its performance, is presented. The ALB Algorithm is designed, based on the clustering where a subset of cluster heads is elected to maintain a certain level of balance within their respective clusters while minimizing the overall communication cost in the cloud environment. The primary goal of ALB Approach is to minimize the total response time of the tasks by distributing the workload. The second goal of ALB Approach is to extend the performance tradeoff results reached by the distributing load system in the cloud environment.

An adaptive approach of a cloud computing system optimizes the energy consumption of the data center equipment, while providing the load balancing of traffic flowing surrounded by the data center. An effective distribution of load system improves the Quality of Service (QoS) of running cloud applications by reducing communication related delays and congestion related information losses in the cloud environment.

4.3. CLOUD INFRASTRUCTURE WITH ADAPTIVE LOAD BALANCING APPROACH

Adaptive Load Balancing (ALB) Approach performs clustering where a set of servers are grouped. ALB Approach works with two main phases, namely repetitive query messaging and clustering server group. Repetitive query messaging is performed with the aid of clusters to satisfy the computational demand of the scheduled load using the mathematical instruction, followed in adaptive load balancing. The concept behind ALB Approach is to find the most suitable group to share the load in order to avoid imbalances and which causes an engendered overhead. In addition ALB Approach with energy resourceful scheduler optimizes the energy consumption of data in the cloud computing. Moreover, the ALB approach equilibrates the communication flows, produced by the users with a minimum amount of bandwidth consumption.

ALBS scheduler examines both the load on the links and the occupancy of outgoing queues as the difficulty lies in the communication overflow at cloud infrastructure. Further, ALB Approach allocates the users the required resources in such a manner that it offers most of the obtainable bandwidth. Further, it penalizes the resources whenever the load exceeds the available transmission capacity when the queue grows in size. The design of the queuing analysis aids in ALB avoiding the tradeoff between the congestion and information losses. The overall architecture diagram of the Energy-efficient cloud computing environment with high bandwidth using Adaptive Load Balancing is shown in Fig. 4.1.

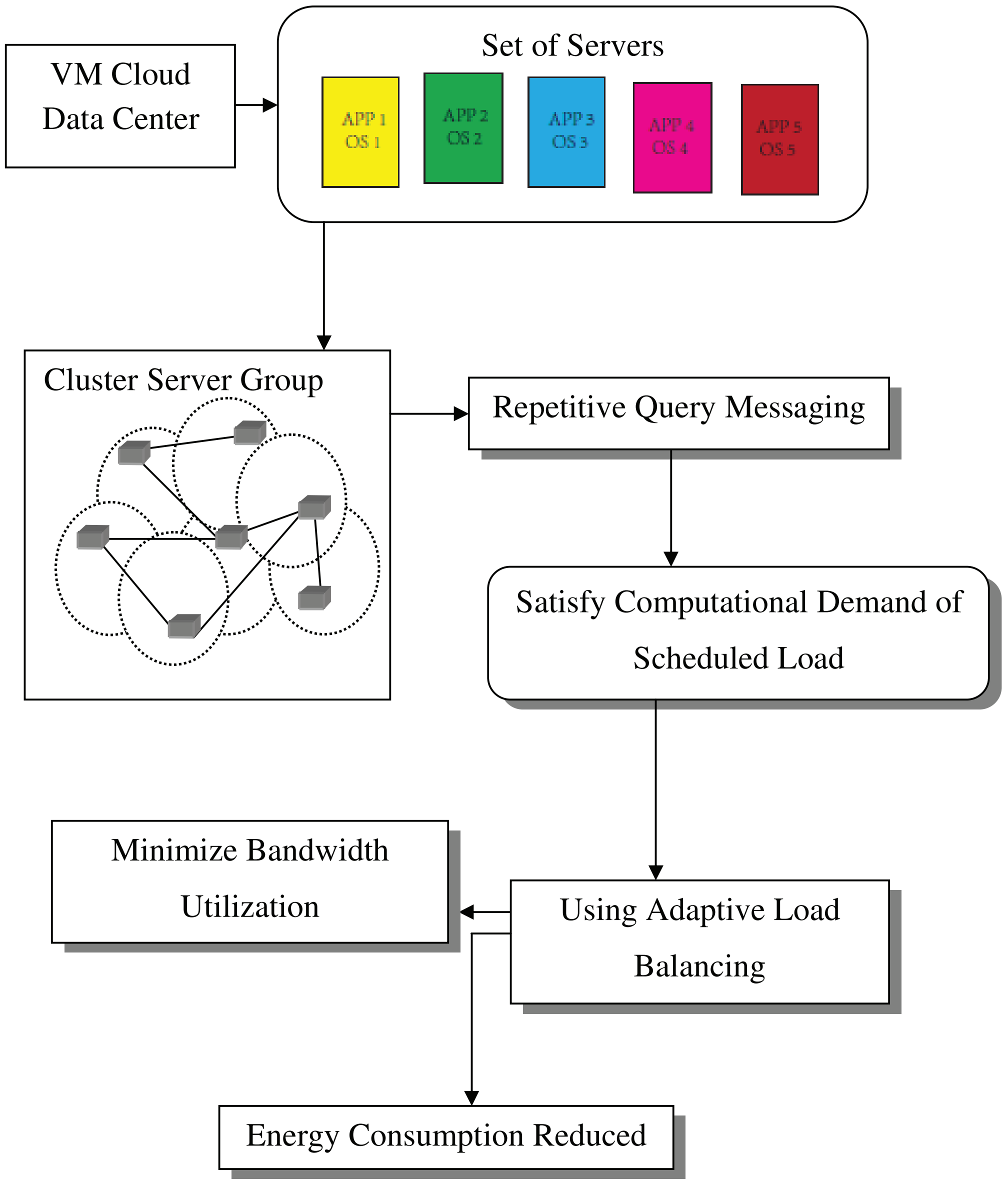

Fig 4.1: Architecture of Energy-efficient Cloud Computing Environment with High Bandwidth Using ALB Approach

ALB Approach defined with the VM cloud data center follows a few steps executed for every received cloud computing data center. A group of servers set connected to the VM data center networks with the highest available bandwidth, provided with at least one of the servers in the set a accommodate the computational demands of the scheduled users. The available bandwidth is defined as the unused capacity of the link connecting the clustered server to the rest of the VM data center network. Within the clustered servers, a computing server with the smallest available computing capacity is chosen to satisfy the computational demands of the scheduled task.

Indeed, the ALB Algorithm balances the loads with repetitive query messaging with little communication between the clouds, as the clouds are powered with servers. The ALB Algorithm relies on the cluster head each time when imbalances occur with regard to a load threshold. The performance of ALB finally achieves the minimal bandwidth utilization and reduces the energy consumption through evaluation. The process of ALB is initiated using the mathematical instruction which is briefed below.

4.3.1. Design Phase of Adaptive Load Balancing

In the VM cloud data center, the server is arranged in the form of a frame ‘F’ and arranged in the form of component ‘C’. Consequently, the frames form a set of components and select the group of servers with the largest available bandwidth. Moreover, ALB initially finds a component such that

$$B(c_i) = max_{\forall c \in C}(B(c)) \quad \dots (1)$$

Where, B is the available bandwidth of a component c_i computed on a per-server basis. For a component c_i∈ C, the available bandwidth computed is given as,

$$B(c_i) = \frac{Tc_{i-}\lambda c_i}{(Sum\ of\ c_i)} \quad \ldots (2)$$

Where, Tc_i is the transmission capacity of a component c_i, calculated as the sum of the maximum transmission speeds of all links connecting a component 'c' to the cloud infrastructure, λc_i is a currently effective transmission rate, and $Sum\ of\ c_i$ is the total number of servers hosted in the component. Equation (2) provides an instantaneous measure of the available bandwidth. However, as most of the transmissions use full link capacity for a short response time, the available capacity is evaluated as an average over the time interval 'TI'

$$Bc_i(TI) = \frac{1}{TI}\int_0^{0+TI}(\frac{Tc_i-\lambda c_i(t)}{(Sum\ of\ c_i)})dt = \frac{1}{Sum\ of\ c_i}(Tc_i - \frac{1}{TI}\int_0^{0+TI}\lambda c_i(t)dt \quad \ldots (3)$$

Similar to the case of components, a frame is identified with most of the available bandwidth, ALB find a frame f_iЄF such that

$$Bf(f_i) = \max_{\forall f \in F}(Bf(f)) \quad \ldots (4)$$

Bf is the available bandwidth of a frame f_i computed on a per server basis cloud infrastructure. For a component $f_i \in$ F the available bandwidth is computed as

$$Bf_i(t) = \frac{1}{TI}\int_{0}^{0+TI} \frac{Tf_i - \lambda f_i(t)dt}{Sum\ of\ f_i}$$

$$= \frac{1}{Sum\ of\ f_i}(Tf_i - \frac{1}{TI}\int_{0}^{0+TI} \lambda f_i(t)dt \qquad \ldots (5)$$

Where, Tf_i is the transmission capacity of a frame 'i', calculated as a sum of the maximum transmission speeds of all links connecting a frame 'i' of cloud group, λf_i is a currently effective transmission rate of the user, and $Sum\ of\ f_i$ is the number of servers, hosted in the frame of a cloud infrastructure.

4.3.2. Clustering Method in Load Balancing

Clustering Method is mainly performed to balance the load in cloud nodes for efficient energy consumption. Clustering is done once the components and frame are identified. Basically, the component and frame are extracted with most of the available bandwidth in ALB Approach. The main task of clustering is to elect a Cluster Head (CH). CH is elected for its relatively high energy capacity.

Energy is the critical resource in ad-hoc networks. Energy consumption of nodes in cloud environment, during communication does not balance linearly with the workload. A node utilizes most of the energy even when the transmission is idle or lightly loaded. Therefore, energy consumption plays a vital role in the successful data communication of resource management in large data centers. ALB Approach widely focuses on balancing the load on a subset of servers and smartly controls the rest of the servers to one of the feasible sleep states which in turn consumes much energy. ALB Approach intends to minimize the energy utilization as the whole at clusters of the servers while taking into account the systems latency and throughput.

In ALB Approach, the CH consumes less energy than a normal server as it has other functionalities to perform. Here, the other functionalities denote coordination between its members, cluster maintenance and load balancing. Moreover, to avoid frequent changes in electing CH in VM cloud data center, the information is updated. Frequent information about the states of nodes widely helps in the process of CH selection in a significant way. The significant updated information exchanges are updated followed by cluster head evolution. The task involved in clustering method for load balancing is depicted in fig. 4.2.

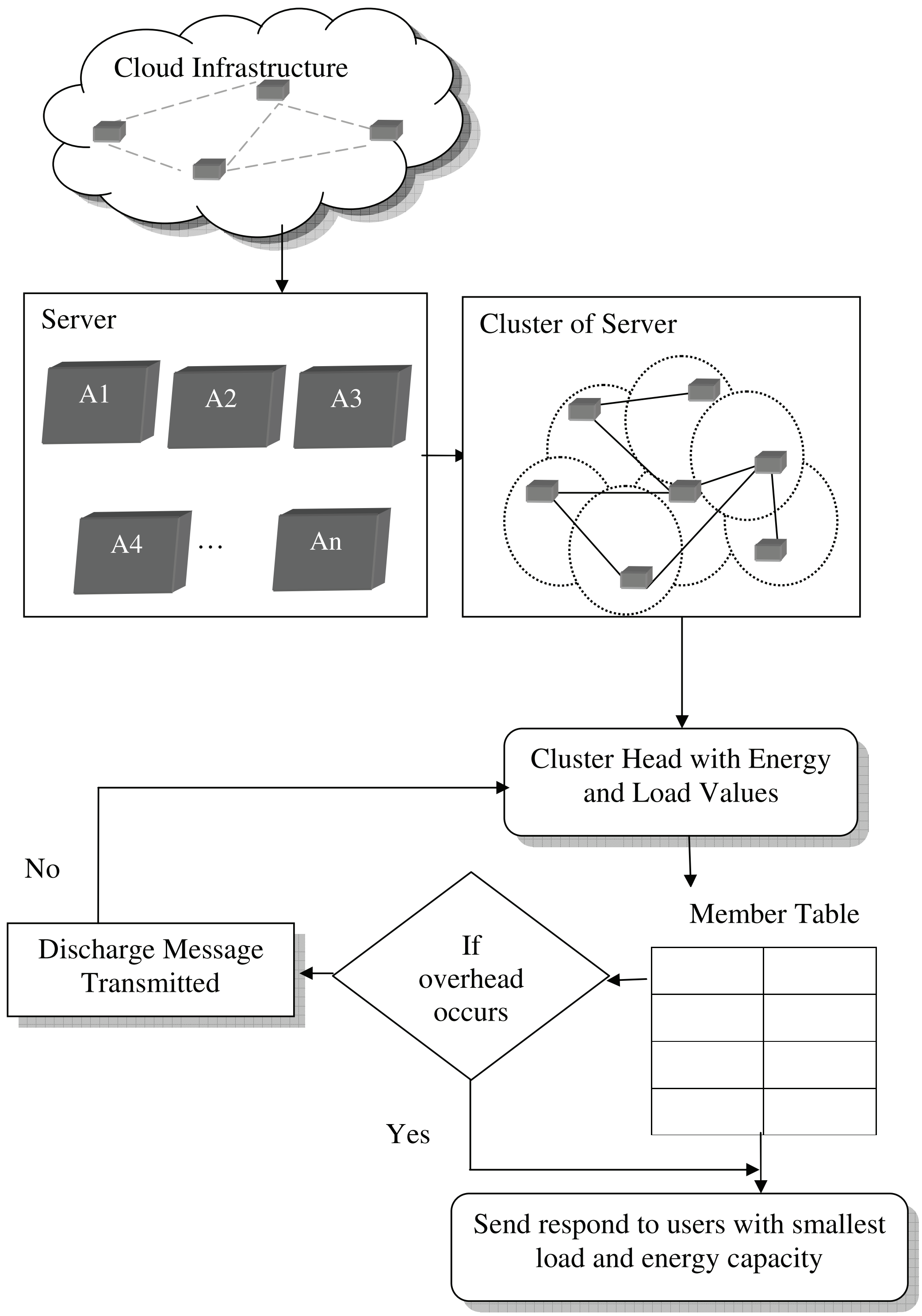

Fig 4.2 Conceptual View of Load Balancing Using Clustering

As mentioned through the shape in Fig 4.2, the pivotal role of a CH is to maintain the load balancing in each cluster. The CH periodically gathers the information about each data center, such as energy and load values. In addition, the respective values are stored in the member table. Whenever a server attains the overload from the users, a discharge message is transmitted to its CH. The secondary server consults its member table and chooses the CH which has the minimum load and the energy capacity followed by a response, sent to the concerned users. Whenever a new server joins a cluster group, the member table is updated accordingly. The load balancing algorithm in ALB is explained as below:

Load Balancing Algorithm in ALB:

Begin

Initialize

Input: Time Interval 'TI', Tf_i is the transmission capacity of a frame 'i',
$Sum\ of\ f_i$ is the number of servers hosted in VM data center

Output: Minimal bandwidth utilization and Energy Consumption

Allocate VM cloud Data Center

  While TI> Tf_i do

        Condition=False

  For 'n' servers do generate cluster head from server group

        Compute member table with energy and load values

        If (overhead occurs)

              Discharge Message transmitted to cluster head

        Else

Member table send respond to users with least load and energy capacity
End If
End For
End While
End

The load balancing algorithm given above elaborates the steps involved in ALB approach. The CH server maintains their member tables in order to control their member loads. Periodically, in each node the member of a cluster sends a message and communicates its energy and load values to the CH which updates its member table. The thresholds for ALB Approach are defined for each server, in such a way that a server is able to perform the process with minimum energy. Each server checks the nodes load as well as the energy. The load and energy is compared with the two thresholds in a periodic manner. If one of the two thresholds is reached, the server sends a message i.e., discharge message to the elected CH. The ALB consults its member table and further it chooses the one which has the smallest load and energy capacity. If one such node is found, the CH sends a positive response to the server, indicating the address of the new server that will receive the extra load. In addition to load balancing, the server congestion is also handled by ALB Approach, which is elaborated in the next section.

4.3.3. ALB Approach in Trigger of Server Congestion

Followed by the process of clustering and query message, the next goal of ALB Approach is to maintain load balance and prevent server congestion. A supportive measure forecasted in ALB is to mainly evaluate the available bandwidth within the data center. Though, such a measure does not detain the system dynamics, like unexpected increase in the transmission rate of the cloud applications, congestion is properly handled. In order to have a more precise measure of the server congestion, ALB scales the measures of the available bandwidth with respect to the component $B\,c_i(t)$ and frame $Bf_i(t)$ with the component related to the size of the row.

$$R(t) = 1 - \frac{1}{TI}\int_{0}^{0+TI}(e^{-\frac{(r(t)-1}{R_{max}}})dt \quad \ldots (6)$$

Where, $R(t)$is an instantaneous occupancy of the row measured at the time't', R_{max} is the maximum allowed size of the row. The objective of ALB to favour the empty row with minimum occupancy and penalize the highly loaded rows is highly convinced. Moreover, ALB also establishes the speed of load, balancing with the growing congestion control. Quantitative analysis is carried out in order to justify the better performance of ALB Approach in the next section.

4.4. EXPERIMENTAL EVALUATION

The ALB Approach is measured against the Virtual Machine (VM) for energy-efficient cloud data centers. For experimental discussions, a set of parameters are taken for the evaluation and implemented using JAVA Cloud Sim Simulator. CloudSim Simulator compiles the codes through Command prompt or through CloudSim with Eclipse, Netbeans, etc. providing an easy process. The specified CloudSim Simulator has been selected as a simulation platform as it is a present simulation structure in Cloud computing environments. Cloud availability structures at transmission layer carries out the optimal analysis, based on the custom configurations supported within the CloudSim. Compared to the simulation toolkits (e.g. SimGrid, CloudSim), JAVA CloudSim provides a copy of on-demand virtualization with enabled bandwidth and submission management. The Virtual machine simulated data center comprises 8 GB of RAM and 1 TB of storage. Energy consumption by the hosts is defined according to ALB Approach.

The user present needs the provisioning of 290 assorted VMs pack power of the virtual data center. Each VM runs a web-application or any kind of application with variable workload, which is modeled to generate the utilization of the bandwidth according to the evenly distributed random variable. ALB Approach uses Statlog (Shuttle) Data Set from UCI repository. The shuttle dataset contains 9 attributes all of which are numerical. Moreover, 80% of the data belongs to class 1 so as to provide high accuracy in the validation process. The instances in the actual dataset are in time order, and this time order could most probably be related in clustering. ALB Approach is compared against the Energy conservation in cloud infrastructures with Service

Request Prediction (SRP) Model of Avinash Mehta., et al., (2011) and Homomorphic Verifiable responses Hash Index Hierarchy (HVHIH) of Yan Zhu., and Shanbiao Wang., (2012) [72] in terms of energy consumption, bandwidth utilization, performance tradeoff and response time.

4.5. RESULT ANALYSIS

In the result analysis section, ALB Approach is compared against the existing Energy conservation in cloud infrastructures with Service Request Prediction (SRP) Model and Homomorphic Verifiable responses Hash Index Hierarchy (HVHIH) in order to prove their superior performance. The evaluation table and graph given below describe ALB Approach improvements in contrast to the existing systems.

4.5.1. Measure of Energy Consumption

The average amount of system energy used for the query processing in cloud structure is termed as energy consumption. The energy consumption of ALB is measured in terms of Joules (J) on processing the resource for balancing the load. ALB Approach is compared against the existing Energy conservation in cloud infrastructures with Service Request Prediction (SRP) Model of Avinash Mehta., et al., (2011) and Homomorphic Verifiable responses Hash Index Hierarchy (HVHIH) of Yan Zhu., and Shanbiao Wang., (2012) [72].

Table 4.1: Tabulation for Energy Consumption

VM Counts	Energy Consumption (J)		
	SRP Model	**HVHIH Mechanism**	**ALB Approach**
1	8.16	7.58	6.15
2	9.54	9.01	7.07
3	10.18	9.22	7.85
4	12.09	11.26	8.89
5	13.01	12.23	9.55
6	15.46	13.95	11.45
7	17.12	16.44	13.25

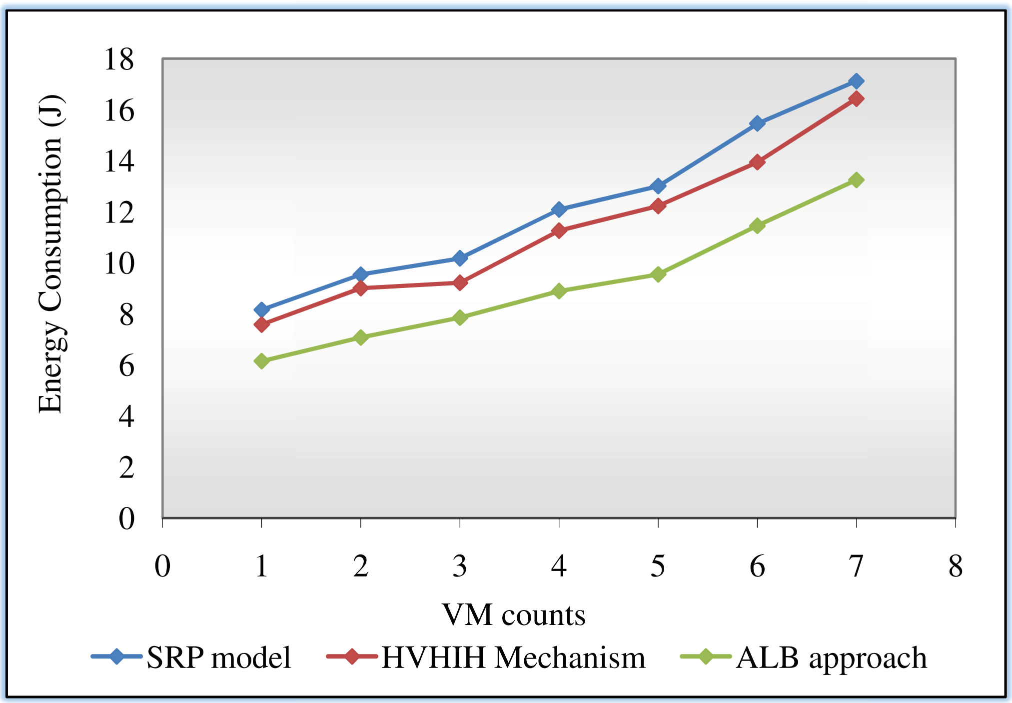

Fig. 4.3: Energy Consumption Measure

Table 4.1 and Fig 4.3 describe the energy consumption based on the virtual Machine counts (VM). The energy consumption is increased gradually as the virtual machine counts increases. As the count of the machines increases in cloud infrastructure, the energy utilization of ALB is reduced approximately to 22 – 26 % when compared with the SRP Model of Yan Zhu., (2012) [70] and approximately to 14-21% compared with the HVHIH Mechanism in Yan Zhu., and Shanbiao Wang., (2012) [72]. This is because ALB checks the load and its energy and compares them with two thresholds in the member table. The usage of the member table information for query processing in ALB minimizes the utilization leading to high energy consumption.

4.5.2. Measure of Bandwidth Utilization

Bandwidth utilization is defined as the measure of the maximum usage by processing the resources for balancing the load of each factor. The bandwidth utilization is expressed in terms of Kilo bits per second (Kbps).

Table 4.2: Tabulation for Bandwidth Utilization

No. of Tasks	Bandwidth Utilization (Kbps)		
	SRP Model	HVHIH Mechanism	ALB Approach
5	2500	2205	2005
10	2600	2310	2150
15	2800	2355	2235
20	2920	2530	2460
25	2990	2645	2515
30	3265	2890	2750
35	3620	3245	3015

Table 4.2 depicts the bandwidth utilization values of ALB Approach in comparison to the existing SRP Model and HVHIH Mechanism.

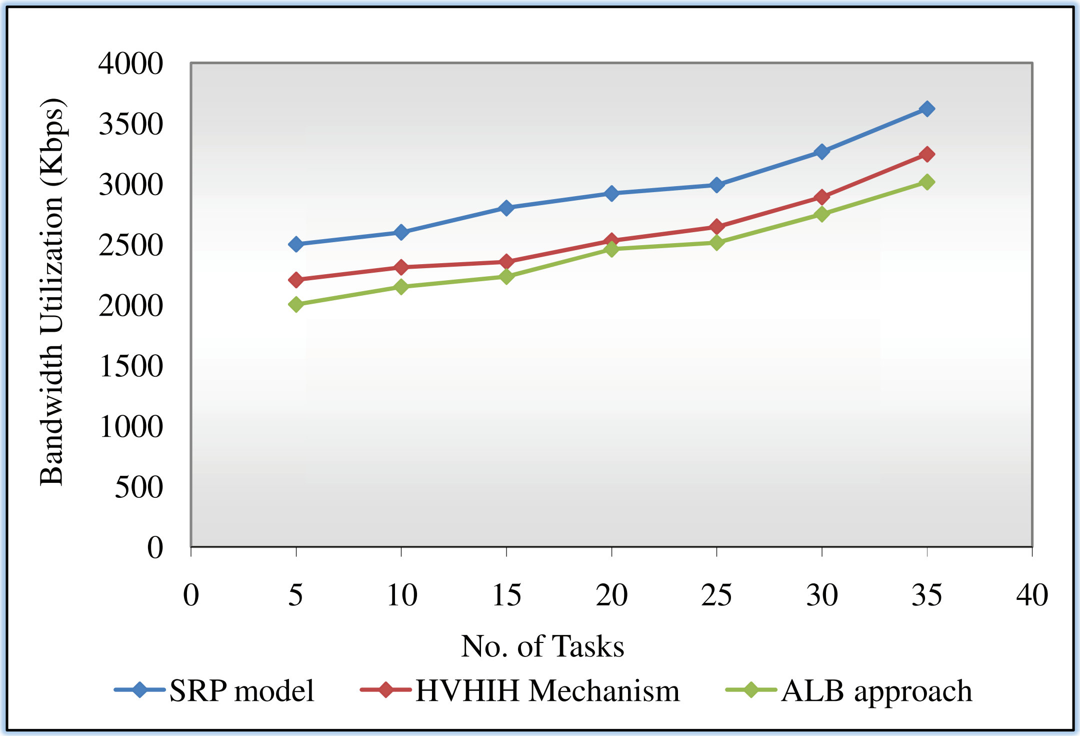

Fig 4.4: Bandwidth Utilization Measure

The bandwidth utilization of ALB Approach is compared against the SRP Model of Avinash Mehta., et al., (2011) and HVHIH Mechanism of Yan Zhu., and Shanbiao Wang., (2012) [72]. The bandwidth of a frame f_i in ALB Approach is computed in terms of per server basis which reduces the bandwidth utilization when compared with cloud infrastructure. As the task gets increased, the bandwidth utilization in the cloud environment is reduced to 15 – 20 % in ALB Approach when compared with the SRP model. The bandwidth utilized of ALB approach is also compared with the HVHIH

Mechanism, where the result for bandwidth utilization is reduced to 2 – 9 % as examined. Moreover, ALB Approach allocates VM only on satisfying the TI> Tf_i criteria, which discards the message transmission on the overhead, reducing much of the bandwidth utilization.

4.5.3. Performance Trade off

The performance tradeoff of ALB is the effective result obtained from the overall system of cloud infrastructure. The performance trade off is measured in terms of percentage (%).

Table 4.3 Tabulation of Performance Tradeoff

No. of Users	Performance Tradeoff (%)		
	SRP Model	HVHIH Mechanism	ALB Approach
20	85	74	99
40	83	70	98
60	82	72	97
80	81	78	97
100	80	71	96
120	80	73	94
140	79	68	93

Table 4.3 describes the performance tradeoff of the SRP Model, HVHIH Mechanism and ALB Approach. Users consider the evaluation, starting from 20, 40, 60 up to 140 counts. As the user count increases, the performance tradeoff is also increased using ALB Approach. The experimental results are depicted in a graph as depicted in Fig 4.5.

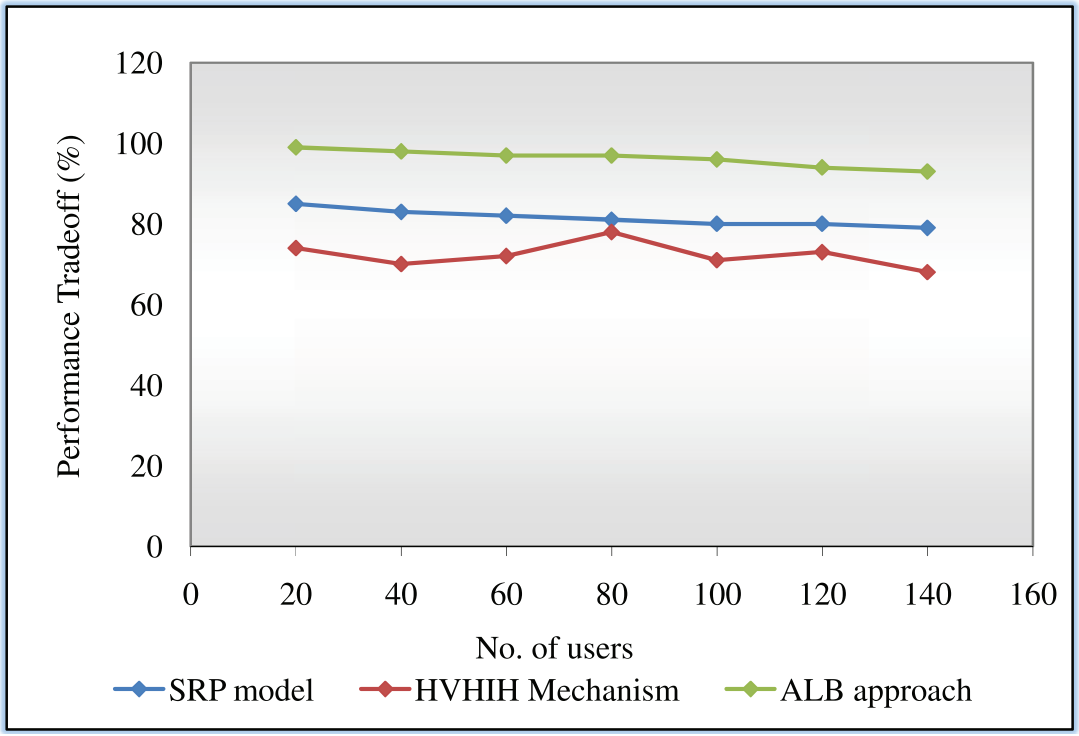

Fig 4.5 Performance Tradeoff Measure

Fig 4.5 describes the performance tradeoff of the ALB and existing system. From figure 4.5, it is evident that the performance tradeoff has been improved after using ALB. The performance of ALB Approach is approximately 14 – 16 % improved when compared with SRP Model and 20 – 30 % improved when compared with the HVHIH Mechanism. As ALB scales the measure the available bandwidth B c_i (t) and $B\ f_i$ (t) within the component with reference to the size of the row, the performance tradeoff has improved, in contrast to existing ones.

4.5.4. Measure of Response Time

The response time for ALB Approach is the measure of the average amount of time consumed in response to the request sent from the clients or users. The response time is measured in terms of seconds (sec).

Table 4.4: Tabulation of Response Time

No. of Requests	Response Time to Service Requests (sec)		
	SRP Model	HVHIH Mechanism	ALB Approach
4	1123	1051	999
8	987	904	852
12	1521	1410	1354
16	221	202	176
20	1830	1695	1423
24	2216	1999	1847
28	2565	2340	2245

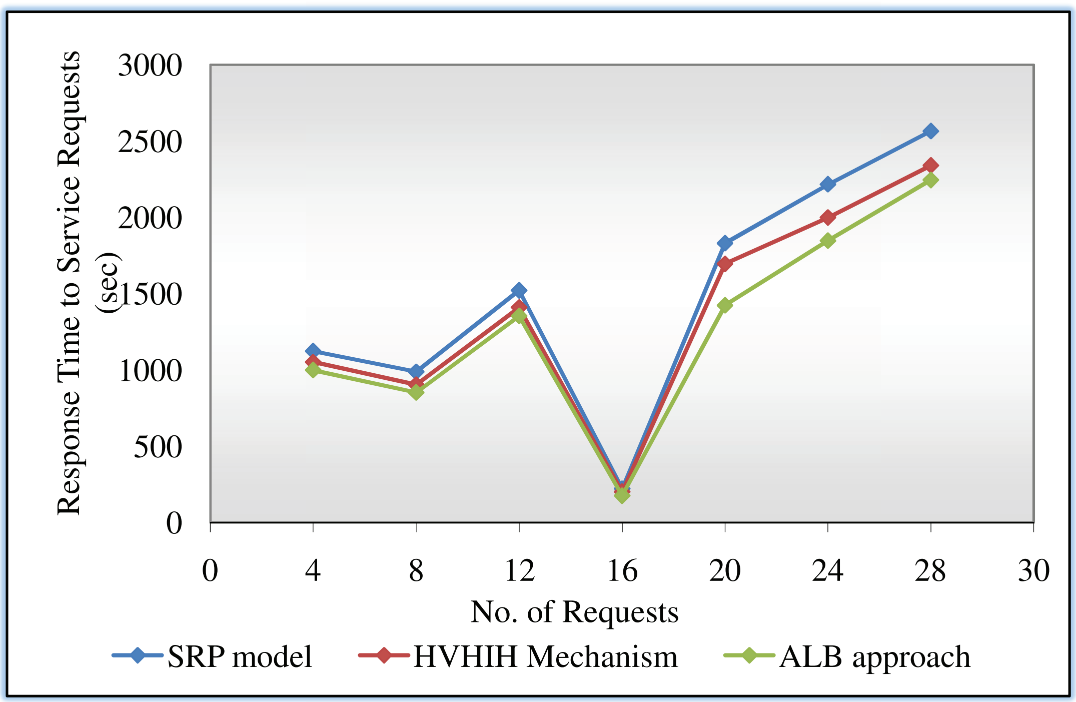

Fig 4.6 Measure of Response Time

Table 4.4 and Fig 4.6 describe the response time of the service request using the SRP Model of Avinash Mehta., et al., (2011), HVHIH Mechanism of Yan Zhu., and Shanbiao Wang., (2012) [72] and ALB Approach. The request taken for the evaluation is 4, 8, 12, up to 28 requests. ALB response time is approximately 10-22 % less in SRP Model and 3 – 16% less using HVHIH Mechanism. The response time to service the request is better than the SRP model and HVHIH mechanism as the response time using ALB of a frame 'i', is calculated through Tf_i, where the maximum response speeds of all links connecting a frame 'i' in a cloud group is minimized. Equation (3) in ALB Approach reduces the response time drastically when compared with the SRP Model.

Finally, the experimental evaluation shows that load balancing is a significant aspect to be evaluated so as to enhance the response time for the service requests and minimal bandwidth utilization based on tasks. Both the energy consumption and minimum bandwidth utilization achievement support the management of the load imbalance in the cloud infrastructure with the aid of clustering. Simulation results reveal a significant improvement of response time, bandwidth utilization and a good energy management for a great number of users.

4.6. SUMMARY

On-demand of load imbalance, modern VM cloud computing data centers has motivated a load balancing approach, termed Adaptive Load Balancing (ALB) to optimize the system energy consumption and bandwidth utilization. ALB supports adaptive load balancing in the VM data center. Transmission of information between the different servers helps to avoid the congestion hotspots and the data losses due to the overflow in cloud infrastructure. Moreover, ALB collects the information about the current load of the other groups by a repetitive query messaging. As a result, ALB improves the quality of service by reducing the delays, related to communication and data loses due to congestion. The validation results, obtained from the VM cloud data center ensure better work in load balancing aspect. Experimental result of ALB Approach is compared with the existing SRP of Avinash Mehta., et al., (2011) and HVHIH system of Yan Zhu., and Shanbiao Wang., (2012) [72] to attain minimal energy consumption and bandwidth utilization. The performance tradeoff is also improved in ALB Approach with minimal response time taken for responding the requests. Energy consumption is also achieved approximately to 14-26 %, using adaptive load balancing factor.