International Journal of Advanced Research in Engineering Technology & Science

ISSN: 2349-2819

Email: editor@ijarets.org Volume-8, Issue-12 December- 2021 www.ijarets.org

OPTIMIZING RESOURCE ALLOCATION IN CLOUD COMPUTING BASED ON ENERGY EFFICIENT LOAD BALANCER TECHNIQUE

Safia Khanam
Research Scholar
Deptt. of Computer Science and Engg.
SJJTU, Jhunjhunu.

Dr. Prasadu Peddi
Supervisor
Deptt. of Computer Science and Engg.
SJJTU, Jhunjhunu.

ABSTRACT:

In cloud computing, a review of previous cloud-related strategies is attempted to improve cloud performance. However, several existing approaches have certain hard issues, and these limits are harming cloud performance. Cloud computing is an important technology for on-demand services, in which cloud-based service providers provide resources, software, and information at a certain time based on the user's needs. Cloud vendors dynamically balance service loads, and additional CPUs, memory, and resources are necessary to handle a larger number of user requests. Clients' business requirements are the basis for this service. The load balancing technique aids in the fulfillment of two important requirements: it first ensures that cloud resources are available and, second, it improves cloud performance.

KEYWORDS: Cloud Performance, Cloud Computing, Cloud Vendors, Cloud Resources.

1.1 INTRODUCTION: Early in the business world, scheduling was a critical tool for managing and improving many activities, as well as the distribution of workload, while carrying out a method or process. Scheduling has long been utilized for resource allocation, such as signing internal/external machinery, hardware/software assets, as well as developing procedures, purchasing supplies, and designing. It used to be the responsibility of the administration to distribute the task/workload among the staff.

A computer's resources include a display, CPU (Central Processing Unit), networking devices, primary and auxiliary storage devices, printer-scanner, trackpad, and other components. To handle these resources, an operating system requires a scheduler that can define resource pre-arrangement if a certain situation or operation requires it.

Cloud computing systems are on the verge of being economically successful, as they can provide a vast array of services and resources to their users. Furthermore, intelligently built suggestion systems have greatly aided in allowing customers to determine whether a particular service is required for them. Scheduling is one of the primary ways for assigning user-defined requests to resources allocated in a specific time in current era of cutting-edge technology. Requests can take the shape of virtual computations, with elements like process and thread running on hardware resources like extension cards, network links, and CPUs. Because a cloud contains an endless number of resources, scheduling methodologies are critical for extracting maximum value from those resources by effectively employing them. To properly execute the requests, many resources should be intelligently automated. When evaluating the purchase of automation, an algorithm is a critical component that is responsible for coordinating job execution across multiple resources while maintaining data security.

The cloud computing integration of different heterogeneous technologies aims to take advantage of the deployment of multiple services. Platform as a Service gives a development environment to the user without taking care of the hardware. Software as a Service provides on-demand leverage of software offered online, whereas Software as a Service provides on-demand leverage of software provided online. Infrastructure as a Service (IaaS) provides a simple, expandable, elastic, and adaptable infrastructure for the deployment of a variety of services. IaaS can be improved in terms of server utilization by following best practices for scheduling. Customers do not have to worry about service execution using various resources because IaaS providers affirmatively provide the ability to deploy surplus services inexpensively.

A CSP (Cloud Service Provider) such as iCloud, Amazon Web Services, IBM (International Business Machines) Corporation cloud, and others provide services dynamically through a network of virtualized and scalable resources that interact with one another. Cloud services are defined as computing clusters in which data transmission takes place across multiple data centers. CSP services can be accessible through a variety of different liveware. Furthermore, these resources must be appropriately managed so that they can be used to their full potential while meeting minimum requirements. An efficient and effective scheduling system must be implemented to appropriately manage the demands.

1.2 OVERVIEW OF LOAD BALANCING IN CLOUD COMPUTING

Because there are so many workloads in a cloud computing system, load balancing is a difficult problem. The main goal of the load balancing idea is to balance the burden among all nodes by reducing execution time, communication delays, and increasing resource usage and throughput. The procedure is depicted in detail in Figure 1.1 below. The most important issue to note in cloud data processing is workload equalization and allocation to the current nodes. Load balancing is a practice of moving burden on a shared basis that increases the system's sustainability over time.

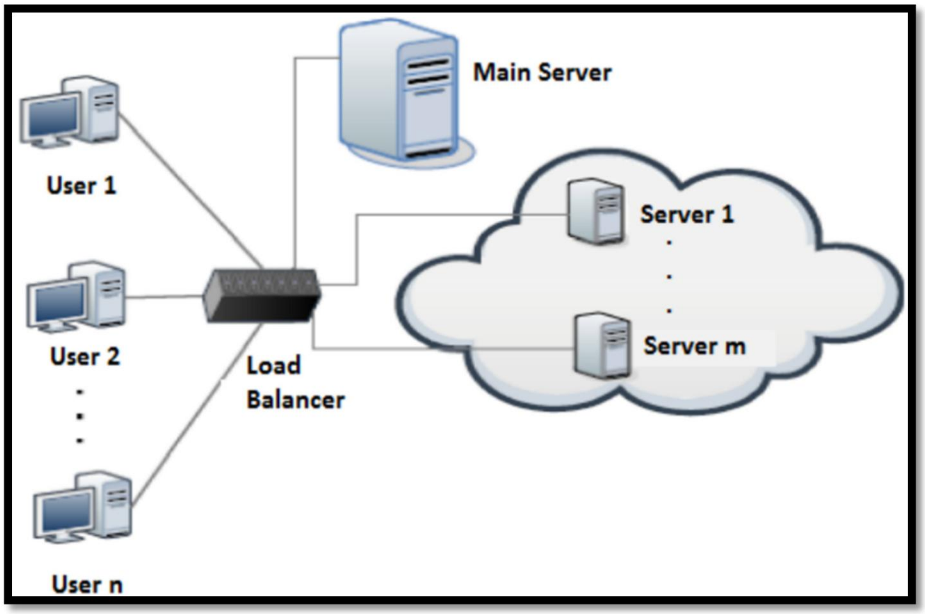

FIGURE.1.1 LOAD BALANCING

2.1 REVIEW OF LITERATURE

Jena, Tamanna, and J. R. Mohanty [2020] used a Genetic Algorithm to demonstrate client-aware job scheduling and resource allocation in a multi-cloud environment (GA). Because the concept of movement is very subjective, mapping future occupation demand to available VMs is a non-polynomial finish problem. Client requirements and use limits change frequently. They developed a GA-based

Customer-Conscious Resource Allocation and Task Scheduling in multi-distributed computing to overcome any barriers between constantly changing client necessities and accessible basis for administrations. The purpose is to assign tasks to VMs in the multi-cloud alliance to reduce make span time and increase customer loyalty. The results showed that the proposed computation beat previous calculations in terms of execution. The simulated multi-cloud setup has a lot of scalability. The simulated scenario does not include data localization costs, latency arbitration, energy consumption, or the running costs of a multi-cloud system.

Abdul Hameed et al. [2020] conducted a review of studies based on taxonomy of strategies for efficiently assigning cloud resources while consuming the least amount of energy. The allocation of diverse virtualized ICT mode of resources in the cloud computing paradigm is a complex problem created by the inclusion of heterogeneous types of applications such as content delivery nodes-based networks, web apps, and Map Reduce, among others. Workloads with more disputed needs for ICT capacity resource allocation, such as network bandwidth, response time, processor speed, and so on, are developed and investigated. With a varied point of achievement, many recent studies have addressed challenges in the enhancement of energy quality and the allocation of cloud resources applications. In any event, there are no distributed processes related to this phenomenon that we are aware of, as it evaluates the exploration issue and provides a scientific categorization based on existing methodologies. As a result, the primary goal of this research is to identify open challenges associated with vitality proficient asset assignment. Furthermore, the currently available approaches in this method deal with the classification of an effective research measurement system. The merits and limitations of present procedures are thoroughly analyzed considering methodology, particularly regarding measuring scientific categorization: asset adaptation approach, target work, chunk strategy, assignment operation, and interoperability.

For cloud data access management, **Pratik P. Pandya and Hitesh A. Bheda [2020]** devised a dynamic resource allocation technique. The Resource Allocation Strategy (RAS) is linked to coordinating cloud supplier exercises for using and assigning rare assets inside the cloud application's point of confinement to address difficulties. To accomplish improved data transfer, it is necessary to follow the norms and procedures of the cloud environment. Because of its actual applicability in server applications, Dynamic Resource Allocation (DRA) is a particularly prominent research subject in cloud computing. Cloud experimental outcomes are varied and dynamic in nature. Different parameters effect VM identification, such as QOS, time consumption, cost, carbon footprint, and so on. Details of the allocation mechanism, VM affinity, and thoughts behind the high performance over non-affinity groups are identified in the proposed system, as well as some suggestions for additional techniques to improve performance. Some fundamental strategies for VM allocation and grouping were demonstrated in this literature.

Daji Ergu et al. [2019] established a hierarchical systems analytical process for scheduling task or work allocation and resource assumption in the cloud environment. Because there are several elective PCs with varying constraints, resource allocation in the cloud computing environment is a complicated problem. The goal of this study is to develop a model for performing focused asset distribution in a cloud computing environment. The Analytic Hierarchy Process and the pairwise examination network method provide the available assets and customer preferences for the resource assignment project. The rating of projects can be used to assign figuring assets. In addition, when competing weights in distinct tasks are assigned, an initiated inclination lattice is employed to separate the conflicting components and improve the consistency proportion. The findings suggest that it is important to continue measuring inconsistent data, and that improving the consistency ratio and task weight can be utilized to dynamically assign computer resources in a cloud computing environment.

In enormous DCs, **J. Shi et al. [2019]** presented the multi-resource allocation technique. This resource allocation strategy was successful in distributing asset administration in the cloud DC in a timely and efficient manner. To improve asset portion productivity, allotment of multi-dimensional resource requests to servers should be enhanced so that asset utilization on servers may be advanced. As a result, the difficulty of resource allocation increases as the measure of DC increases. As a result, current techniques are incapable of achieving speedy and efficient asset distribution for large-scale DC. A pattern-based multi-resource allocation method is offered as a solution to this problem. After classifying the resource requests, this technique generates the pattern information. A multi resource allocation describes the different types of requests that can be issued to the servers.

3.1 OBJECTIVES OF THE STUDY

The load balancing approach is used to achieve energy efficiency. Load balancing functions as a layer between the VMs and the client, allowing the task in demand to be executed as quickly as feasible. The scheduled load balancing between the available VMs clearly saves money and energy.

4.1 PROPOSED METHODOLOGY

The proposed energy efficient load balancing technique of dynamic defragmentation model for cloud storage provides high reliability, resistivity, and superior cloud storage decisions to each client. In small-scale distributed systems, the Min-Min and Max-Min algorithms are frequently used.

When the number of little tasks in a meta-assignment exceeds the number of significant activities, the Max-Min schedules the tasks, with the framework's make span largely determined by the number of tiny tasks running concurrently with the large ones. A modified Max-Min algorithm is developed to solve this constraint.

4.2 CLOUD LOAD BALANCING FOR AUTOMATED DYNAMIC DEFRAGMENTATION

Certain cloud capacity in the form of a virtual DC is included in the suggested paradigm (VDC). The storage servers are grouped into clusters, which are made up of individual servers. When a cloud user seeks to share or upload a large volume of data, the model momentarily moves data into fragmented servers rather than storing it in the accessible cloud data server. The quantity of accessible space on the cloud server for resource allocation is determined using an adaptive method. Based on prior allocation requests, the model determines whether defragmentation is required. If this is the case, a trigger for automated defragmentation is in the works. Once the defragmentation is complete, the data is organized and transported back to the cloud server from the fragmented cluster. The design has a number of benefits, including the ease with which VMs can be moved between servers, and the fact that the whole allocation of VMs is found on a server-side system with all of its physical quality. When a collection of VM systems is dispersed across multiple clusters, however, a central method becomes impracticable, because dynamic per-VM information must be made available under a cloud-level database shared by hundreds or even thousands of clusters.

The Pseudo code for Max-Min Algorithm

```
for i=1 to M
    for J=1 to N
        $C_{ij} = E_{ij} + R_j$   // $C_{ij}$ is the completion time of the, $E_{ij}$ is the task execution
```

```
time, $R$j is the ready time of the task i on VM j.
end for
end for
do until all the unscheduled tasks are exhausted
for each unscheduled task
find max. completion time (T) and VMs that obtains it
end for
find the task $tp$ with T
assign task $tp$ to the VM that give the T
delete task $tp$from pull of unscheduled tasks
update the initial time of VM that gives the T
end for
```

5.1 EXPERIMENTAL RESULT AND DISCUSSION

The software model is built on the widely used Java platform. For replicating the cloud defragmentation model, all the modules have been programmed to the highest level possible. To replicate the proposed model, a Servlet-based application is created. The typical login handle provided by the cloud service provider is used to authenticate a cloud user. When a user requests that a chunk of data be stored, the data size is examined. The cluster of servers then examines and predicts whether the cloud storage space is heavily fragmented with files and requires defragmentation based on the data size, and if so, a defragmentation schedule is activated.

5.2 PERFORMANCE ANALYSIS

The performance of the suggested energy-efficient load balancing technique in a cloud storage model for automated defragmentation is measured using a variety of metrics including processor speed, throughput, instruction volume, and data volume. Assume we have four tasks T1, T2, T3, and T4 in meta-tasks and the scheduling manager has two resources R1 and R2 as a problem set to show the suggested technique.

TABLE.1.1: MAX-MIN AND MODIFIED MAX-MIN TECHNIQUES IN ACTION

Problem sample	Resources	Processing Speed (MIPS)		Throughput (MBPS)	
		Max-Min	Modified Max-Min	Max-Min	Modified Max-Min
P1	R1	29	50	87	100
	R2	73	100	300	500
P2	R1	128	150	279	300
	R2	283	300	130	150
P3	R1	272	300	282	300
	R2	17	30	130	150

The performance of the existing max-min algorithm and the suggested Modified max-min method is shown in Table. The problem sets p1, p2, and p3 are comprised of R1, R2, and R3 resources.

TABLE.1.2: INSTRUCTIONAL PERFORMANCE AND DATA VOLUME

Problem sample	Task	Instruction Vol. (MI)		Data Vol. (MB)	
		Max-Min	Modified Max-Min	Max-Min	Modified Max-Min
P1	T1	112	128	39	44
	T2	52	69	53	62
	T3	201	218	83	94
	T4	11	21	43	59
P2	T1	212	256	73	88
	T2	21	35	25	31
	T3	298	327	89	96
	T4	201	210	432	590
P3	T1	11	20	63	88
	T2	321	350	27	31
	T3	192	207	83	100
	T4	17	21	37	50

The volume of data and instructions necessary for the task is represented in Table.4.2. The sample problem sets are labeled P1, P2, and P3 in this example. Each problem set has four tasks, denoted by the letters T1, T2, T3, and T4. The modified Max-Min method produces better outcomes than the Max-Min method. The node-to-node communication rate is really high. Modified Max-Min transfers a significant amount of instruction and data in a short length of time.

TABLE.1.3: MAX MIN TECHNIQUE AND MODIFIED MAX MIN METHOD COMPLETION TIMES.

Task/Resource	R1		R2		R3	
	Max-Min	Modified Max-Min	Max-Min	Modified Max-Min	Max-Min	Modified Max-Min
T1	152	102	250	236	178	123
T2	198	150	175	143	278	232
T3	98	58	210	188	256	193
T4	252	232	198	140	290	231

The task completion time of the max-min and modified max-min with various resources is shown in the table. The value of (T) from the Modified Max-Min technique is superior to the classic Max-Min technique in terms of outputs.

The proposed defragmentation model aids cloud providers in administering and systematically distributing cloud users' data and storing it on a distributed cloud host in an efficient manner. An effective scheduled load balancing strategy has been developed among the previously reviewed strategies to save energy and money when the operation is being carried out and satisfying demands. Load balancing functions as a layer between the VMs and the client, allowing the task in demand to be executed as quickly as feasible. The scheduled load balancing between the available VMs clearly saves money and energy.

The existing resource scheduling approach is inefficient due to its high power consumption. The research is entirely based on the understanding of power utilization, which is directly proportional to CPU consumption from the server, and thus proposes a task consolidation technique that sequentially maps requests from the client to the machine's processor, resulting in increased CPU utility and reduced energy consumption at the same time.

6.1 CONCLUSION

Cloud computing is a large-scale distributed, parallel computing platform in which computer resources are given as a service based on a service level agreement (SLA) signed by the cloud service provider and its consumers. Amazon EC2, Microsoft Azure, and Google App Engine are the main CSPs that offer Infrastructure, Platform, and Software administrations as part of their utility processing administrations. The great part of organizations is outsourcing their IT administration to the cloud to reduce the initial capital expenditure in building up the IT framework and to reduce the weight of equipment and programming maintenance. In most cases, a cloud-based application should balance its computing load. The purpose of the Resource Allocation Algorithm is to alleviate overcrowding and improve the environment. If an application receives several requests, the resource allocation method should distribute the requests among multiple physical machines using virtual machines. The resource allocation algorithm should be aware of cloud storage space and resource use for this reason, and it should outperform other traditional algorithms.

6.2 FUTURE WORK

In a cloud environment, a task consolidation approach is employed to enhance CPU utilization while reducing energy consumption. However, security in cloud storage is lagging, and future research effort should focus on and address security considerations and provisions for virtualization, as well as the best use of cloud infrastructure.

REFERENCES

- ❖ Randles M, Lamb D, Taleb-Bendiab A (2010) A comparative study into distributed load balancing algorithms for cloud computing. In: IEEE 24th international conference on advanced information networking and applications workshops, pp 551–556
- ❖ Rodero I, Jaramillo J, Quiroz A, Parashar M, Guim F (2010) Energy-efficient application-aware online provisioning for virtualized clouds and data centers. In: International conference on green computing (GREENCOMP ’10)

- Saure D, Sheopuri A, Qu H, Jamjoom H, Zeevi A (2010) Time-of-use pricing policies for offering cloud computing as service. In: IEEE SOLI 2010, pp 300–305
- Teng F, Magoulès F (2010) Resource pricing and equilibrium allocation policy in cloud computing. In: 10th IEEE international conference on computer and information technology (CIT 2010), pp 195–202
- Efficient Resource Provisioning in Compute Clouds via VM Multiplexing by Xiaoqiao Meng, Canturk Isci, Jeffrey Kephart, Li Zhang, Eric Bouillet Dimitrios Pendarakis IBM T. J. Watson Research Center Hawthorne ICAC'10, June 7–11, 2010, Washington, DC, USA.
- Yazir YO, Matthews C, Farahbod R (2010) Dynamic resource allocation in computing clouds using distributed multiple criteria decision analysis. In: IEEE 3rd international conference on cloud computing, pp 91–98
- [Above the clouds: a Berkeley view of cloud computing by M. Armbrust, et al. Tech. Rep. CB/EECS-2009-28, EECS Department, U.C. Berkeley, Feb 2009.
- Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility by R. Buyya, C. Shin Yeo, S. Venugopal, J. Broberg, I. Brandic R. Buyya et al. / Future Generation Computer Systems 25 (2009) 599_616.
- Cloud Computing Resource Management through a Grid Middleware: A Case Study with DIET and Eucalyptus by Eddy Caron, Frédéric Desprez, David Loureiro and Adrian Muresan CLOUD '09 Proceedings of the 2009 IEEE International Conference on Cloud Computing.
- Resource provisioning for cloud computing by J. Wong, G. Iszlai, M.L. Ye Hu, Copyright © 2009 Ye Hu, Johnny Wong, Marin Litoiu and IBM Canada Ltd.
- Supporting Database Application as a Service by Zhou Yuan Data Engineering, 2009. ICDE '09. IEEE 25th International Conference on March 29, 2009-April 2, 2009.