JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Unification of AutoScaling Servers with Elastic Load Balancer in Cloud Computing : A Review

Viswaprasad Kasetti[1],Desineedi Jyosmitha Durga [2],Dasari Naga Deepika[3] ,Bandla Sanjay [4],Marni Tanuja[5],M.Parthiban[6]

Department Of Computer Science &Engineering ,Sasi Institute Of Technology & Engineering

***Abstract* -Although we have Auto Scaling Algorithms and Load Balancing Algorithms still they are facing numerous challenges, but their main aim are to fast response to the user requests and balancing the resources but many of the algorithms considering only few of the Quality of services parameters like allocation time, response time .. etc and remaining parameters are ignoring, we are over coming that problem by providing high availability of the resources to the requests and low latency time by comparing all the collected algorithms which are mentioned in this paper like Round robin algorithm, shadow routing algorithm, State-Based load balancing algorithm and finding out the drawbacks and over coming to that drawbacks**

Keywords*— *Autoscaling, Load balancing, Round Robin ,High Availability ,Low Latency ,Cloud Computing

I.INTRODUCTION

For storing the data we usually use the books or records or some files but these are up to the certain limit only, we can't store a huge and lots of data simply bigdata we can't store them, so to store such type of data we should need some thing i.e Cloud. In cloud we can store bigdata and simple and even small data can also be stored in the clouds their is no limits to store the data up to some limit so up to now we solved the problem of storing but what about the retrieving the data from the cloud ,we use some queries like SQL or some techniques to get the data but the major problem is if n no of people are accessing the cloud definitely their will be the slow response to the requests to avoid this problem their should increase the resources and decrease the resources based up on the demand this is nothing but scaling but it should done automatically so it termed as "Auto Scaling" there were the algorithms like shadow routing ,Artificial Neural Networks ,Hybrid Autoscaling these are all the algorithms which auto scales the resources based up on the demand but will happen when we are continuously getting data from the same server or resource ,the resource or server will get effect so we should balance the requests to the servers here we are balancing the load infact distributing the requests from the user equally to the servers which is termed as "load balancing" for this Round Robin ,Data Type file formatting ,State-Based load balancing and some mor algorithms are used for balancing the load in the servers but all these works should have the high efficiency ,high accuracy and low latency but collected algorithms are some lack of availability and resource allocation so we are going to study the points where these algorithms are lacking off. We study all that points and provides high availability ,Low latency to the requests of the user

II. Literature Overview

This paper[1] examined the auto-scaling issue with cloud-hosted application hosting. Their main goal is to reduce the number of Physical Machines used to host Virtual Machines. To do this, they devised the shadow technique It uses a specially designed created a dynamically usage of virtual queueing system deliver the best outcome that controls VM auto-scaling and VM-to-PM packing..

A cloud load balancing method based on dynamic virtual machine reconfiguration was also suggested in the paper[2] when variations in load or user request volume were observed. They developed a dynamic reconfiguration tool called the inter-cloud load balancer (ICLB) that allows the virtual resources to be scaled up or down.

In a similar vein, The Paper[3] examined load balancing in the cloud, where the main goal is to map out workloads. They also suggested metaheuristic algorithms for precise solutions.

The paper[4] also looked at an intermediary company's automated provisioning of cloud resources. This business offers a private cloud that functions as a virtualized public cloud for a single customer company. It proposed a hybrid auto-scaling method based on the fusion of proactive and reactive methodologies to expand resources in response to user demand..

This study [5] analyzed the biggest weakness of the load unbalanced, one of the major issues for cloud providers. On the Min-Min algorithm's foundation, an improved load balanced method (LBIMM) was developed in order to reduce the Make span and increase resource usage. Additionally, cloud service providers offer their clients pay-per-use access to computer resources.

The issues of workload balancing in cloud computing was investigated in this study [6]. Even with the use of Infrastructure as a Service under the cloud model, this challenge is still challenging to resolve. They recommended implementing the MOABCQ method a reinforcement learning approach that speeds up the ABC process. The recommended course of action is to increase VM throughput while improving scheduling and resource usage.

This research [7] examined how cloud computing still has a lot of load balancing difficulties despite substantial infrastructure improvements. They put forth SVM and a modified version of Cat Swarm Optimization are combined in the load balancing method known as Data Files Type

Formatting (DFTF). All performance metrics showed an improvement, including throughput (7%), response time (8.2%), migration time (13%), energy use (8.5%), optimization time (9.7%), overhead time (6.2%)

This research [8] described the autoscaling and load-balancing capabilities of a single, decentralized architecture and discussed ways to reduce operational overhead while enhancing response time performance. The creation of an analytical system model demonstrates that there is a good chance that the suggested approach will result in asymptotic zero-wait time.

This paper[9] explained how the load balancing algorithm takes deadlines into account while allocating work efficiently in light of the constrained resources and virtual machines. The proposed method aims must consider the Quality of Service (QoS) job characteristics, the priority of VMs, and resource allocation in order to optimize resources and enhance load balancing.

This paper [10] discusses the effectiveness of a transparent auto-scaling algorithm option for Kubernetes that scales-in/out containers using absolute usage measures, as well as the issue of selecting the best performance metric to initiate actions designed to ensure QoS constraints.

This paper [11] discussed about the prominent use of cloud in IT services to start a business or to utilize the resources without any capital investment and it also discussed about the Horizontal Cloud Scalability and Vertical cloud scalability

This research [12] explored how as input data size and complexity increase, system overhead increases. They suggested using an artificial neural network (ANN) with linear regression to accurately anticipate resource needs.

This paper [13] discussed about the load balancing in this way most of the physical hosts in data centers for the requests of users are overloaded which makes whole could imbalance and algorithms that we are using have high complexity so this paper proposed a heuristic algorithm which achieves the over all load balancing and improving the efficiency

In the cloud, when some servers are overcrowded and some sources are underloaded, load balancing is crucial, as was highlighted in this paper[14]. There are currently a large number of algorithms that may be used to tackle these issues, but this algorithm was developed utilizing a new paradigm for optimization search called osmotic computing, which effectively balances the load in the cloud.

This paper[15] discussed about the elasticity concept like this although elasticity should need in the cloud it gradually increase or decrease the loads in web, consider the area of HPC , by using the previous knowledge and it also undergoes modification ,this paper proposed PaaS-level elasticity model for HPC which gives high performance without user intervation

This Paper [16] discussed about the Auto Scaling of resources , there are huge no of auto scaling of resources are their in cloud but the no of time of auto scaling should also be in control and they proposed Cloud resource auto-scaling to minimize the number of auto scaling systems.

In this paper [17], the authors explored the significance of resource consumption and offered a framework that would reduce the need for Service level Agreements while increasing resource utilization by using less active servers

In order to reduce resource waste, the authors of this work [18] presented a resource allocation system based on the concepts of coalition formation.

The topic of cloud load balancing, which is performed by consolidating servers, was discussed in this paper [19].

To maximize network throughput while dynamically balancing workloads, this paper[20] discussed the ovel dynamic load-balanced scheduling (DLBS) method.

III. Methodologies and Approaches

The Methodologies used for Auto Scaling and Load Balancing are shadow routing to reduce the Physical Machines packing into the Virtual Machines ,inter- cloud elasticity used to evaluate the frame work for effective resource utilization, ,Meta heuristic load balancing, To shorten the Make span, use the Min-Min algorithm. algorithm., Artificial Bee Colony algorithm for resource Utilization, Data Files Type Formatting for partitioned data as output.

In addition to JSQ (Join-the-Shortest-of-d-Queues), other load-balancing algorithms include Join-the-First-Idle-Queue (JFIQ), which reduces the monitoring burden., State-Based Load Balancing (SBLB) handle dynamic user requests and resource allocation, and for auto scaling Kubernetes auto-scaling algorithm ,Horizontal Cloud Scalability & Vertical cloud scalability to increase the capacity of existing hardware more resources ,Composite ANN for resource allocation , Osmotic Hybrid Artificial Bee and Ant Colony optimization

IV .Challenges and Gaps

In the online-VM Autoscaling paper they proposed shadow Routing Algorithm for optimal solution although optimal solution came but it lacks at public cloud environment to solve this problem a hybrid auto-scaling method suggested in paper [4] that blends being proactive and reactive approaches to optimize the resource allocation which leads to the public cloud environment

The paper [8] used load balancing policy, JIFQ(Join-In -First -Idle -Queue) for the Asymptotic Zero Wait time but it lacks at the sophisticated balancing and autoscaling to overcome this Round Robin method is proposed in paper[] that increased efficiency and accuracy

The paper [6] used Artificial Bee Colony in order to reduce the make span ,cost reduction but it lacks at dynamic resource allocation to overcome that problem the paper[7] introduced the Data Files Type Formatting it allocates any type of files and convert it and the files like audio ,video ,images ..etc which supports the dynamic resource allocation

The paper [5] used Min-Min scheduling algorithm that is user-priority-driven in order to increase the load balancing and it was not adoptive much for advanced features to overcome this The paper[9] introduced state-Based Load Balancing Algorithm it uses different methods for allocation which is adaptive in nature

Whenever a container's auto-scaling mechanism is used to manage CPU-demanding workloads, QoS constraints, the results presented should be taken into consideration as a guideline. The paper [10] used Kubernetes' auto-scaling algorithm to transparently leverage absolute usage measures rather than relative ones. To avoid such type of problems the paper[12] is completely about Neural Networks about AutoScaling for efficient resource allocation.

VI. FUTURE WORK AND DIRECTION

Paper no	Proposed Work	Algorithm Used	Parameters
1	Optimal Solution for VM Auto scaling	Shadow Routing	Physical Machine Utilization -53.4%
2	HTTP load balancing	Round Robin	Response Time(seconds)-16 Active Server(secs)-26
3	Providing the ideal solution	Metaheuristic load balancing	Resources Utilized by 75%
4	Hybrid auto-Scaling method	a blend of proactive and reactive measures	Mean Service Time-50 mins Mean Arrival Rate (in Requests)-0.0083 Load Factor- 0 to 1
5	Load Balancing ,Task Scheduling	User Priority Guided Min-Min Algorithm	ResourceUtilization-84 Make span(ms)-834 Execution Time(ms)-534
6	Reduce the time, expense, and simultaneous resource utilization	Artificial Bee Colony	Make span
7	enhancing load balance and performance as a goal	Data Files Type Formatting	Accuracy -0.974,Precision-0.963 Recall-0.951,F-Measure-0.977
8	Load balancing policy	Join in First Idle queue	Response time-106 ms Threshold-104 ms
9	Dynamically Assign tasks to idle	State-Based Load Balancing	Make span-892 Execution Time- 650 Resource Utilization-89
10	Detailed Analysis of state-of-the-art	Kubernetes	Utilization Time-85
11	Allocating resources based on user demand	Horizontal Scaling, Vertical Scaling	Cost Saving 40%
12	Optimal Resource Allocation	Artificial Neural Networks	Resource Allocation-90
13	physical host selection issue for carrying out specified activities	novel heuristic approach	Make span -830
14	overcoming the shortcomings of the earlier load balancing studies	Osmotic Hybrid artificial Bee and Ant Colony optimization	Time complexity reduced 93%
15	high performance programs without user involvement	PaaS-level elasticity model for HPC	Cost Reduced 18%, Resource Utilization increased 20%
16	Reducing the no of Auto Scaling Systems	Cloud resource auto scaling system	Threshold increased 50%
17	Efficient resource utilization ,Power Saving	Resource management frame work	Resource Utilization increased 89.3% Power Consumption decreased 25%
18	Increased request satisfaction and improved resource use	Coalition Dissolving	Resource Utilization increased 90%
19	Reduces the Energy Consumption	Load Balancing Algorithm	Response Time-90 ms Make span Time –800
20	maximizing the network throughput and balancing workload dynamically	novel dynamical load-balanced scheduling	Threshold Increased 90%

The Algorithms which we studied provided efficient way of resource allocation and load balancing but these algorithms have some limitations like accuracy, throughput, response time, so we are going to launch the algorithms in cloud in amazon web services which provides the high availability of resources and low latency time and security mechanisms

VII. CONCLUSION

We concluding that we collected a lot of algorithms related to Auto Scaling and Load Balancing they had a more no of advantages that are responding to the user requests and balancing the requests from the users to the servers but most of the algorithms have the accuracy problems and allocating resources, efficiency. we found out all the drawbacks in the algorithms and done the work to provide High Availability of resources to the requests in cloud and Low Latency that are responding to the requests in less time and providing Security Mechanisms

REFERENCES

[1] Guo, Y. (2018). Online VM Auto-Scaling Algorithms for Application Hosting in a Cloud. *IEEE Transactions on Cloud Computing* , 889-898.

[2] Stelios Sotiriadis, N. B. (2019). Elastic Load Balancing for Dynamic Virtual Machine Reconfiguration Based on Vertical and Horizontal Scaling. *IEEE TRANSACTIONS ON SERVICES COMPUTING*, 16.

[3] Zhou, J. (2023). Comparative analysis of metaheuristic load balancingalgorithms for efficientloadbalancing in cloud computing. *SpringerOpen*, 12.

[4] Biswas, A. (2017). A Hybrid Auto-Scaling Technique For clouds Processing Applications with Service Level Agreements. *SpringerOpen*, 12.

[5] Chen, H. (2013). Load Balancing in Cloud Computing User Priority Guided Min-Min Algorithm. 12.

[6] KRUEKAEW, B. (2022). Multi-Objective Task Scheduling Optimization For LoadBalancing in Cloud ComputingEnvironment using Hybrid Artificial Bee-colony Algorithm With Reinforcement Learning. *IEEE*, 16.

[7] Junaid, M. I. (2020). Modeling an Optimized Approach for load balancing in cloud. *IEEE*, 12

[8] Desmouceaux, Y. (2021). Joint Monitorless Load- Balancing and Autoscaling for Zero-time Enivronments. *IEEE Transcations on Network and ServiceManagment*, 15

[9] Junaid, M. I. (2020). Modeling an Optimized Approach for load balancing in cloud. *IEEE*, 12.

[10] Casalicchio, E. (2019). A study performance measure for auto-scaling CPU-intensive containerized applications. *CrossMark*, 12.

[11] Kriushanth. (2013). AutoScaling in cloud. *International Journal of Advanced Research*, 12.

[12] Ekhande, A. (2018). Improvement in Auto Scaling mechanisms of cloud computing resources using composite ANN. 19.

[13] Zhao, J. (2015). A Heuristic Clustering-based Task Deployment Approach for Load Balancing Using Bayes Theorem in Cloud

Environment. *IEEE Transactions on Parallel and Distributed Systems*, 305-316

[14] Gamal, M. (2019). Osmotic Bio-inspired Load Balancing Algorithm in Cloud Computing. *IEEE*, 42735 - 42744.

[15] Righi, R. d. (2015). Automatic Resource Elasticity for High Performance Applications in the Cloud. *IEEE Transactions on Cloud Computing*, 6-19.

[16] Hasan, M. Z. (2015). Integrated and Autonomic Cloud Resource Scaling. *IEEE Network Operations and Management Symposium*.

[17] Saxena, D. (2021). OP-MLB: An Online VM Prediction-Based Multi-Objective Load Balancing Framework for Resource Management at Cloud Data Center. *IEEE Transactions on Cloud Computing*, 2804-2816.

[18] Rao, S. (2016). Resource Allocation in Cloud Computing Using the Uncertainty Principle of Game Theory. *IEEE Systems Journal* , 637-648.

[19] ALA'ANZY, M. (2019). Load Balancing and Server Consolidation in. *IEEEAcess*.

[20] ang, F. (2016). A Dynamical and Load-Balanced Flow Scheduling Approach for Big Data Centers in Clouds. *IEEE Transactions on Cloud Computing* , 915-928.