Optimizing the Cloud Performance By Dynamic Workload Allocation

Aishwarya T M [1], Ajay Karan A [2], Ashwini M [3], Dr S.Ambareesh[4]
UG students[1, 2, 3], Associate Professor [4]
Department of Computer science and Engineering
Vemana Institute Of Technology, Bengaluru-34.
tm.aishwaryaraj@gmail.com[1], ajaykaranom@gmail.com[2], ashwini.3696@gmail.com[3] , ambihce@gmail.com[4]

Abstract— Cloud computing technology that uses the internet and focal remote servers to keep up information and applications. Resources allocation is to dispense the resources based on infrastructure as a service. Cloud computing offers dynamic provisioning and thus can distribute machines to store data. Need of resources are essentially expanding step by step. There is still absence of apparatuses that empower designers to analyse different resource allocation techniques in IaaS with respect to the two servers and client workloads. Cloud computing technology takes into consideration considerably more proficient registering by centralization storage, memory use and CPU clock cycle. A workload allocation algorithm, named max-min-cloud, is conceived to optimize the execution of the cloud service.

Index Terms—Cloud computing, dynamic provisioning, optimize.

I INTRODUCTION

The Cloud computing rises as another registering worldview which means to give dependable, customized and QoS (Quality of Service)[1] ensured processing dynamic situations for end-clients. The organizations which give distributed computing administration could oversee and keep up the activity of these server farms. The clients can get to the put away information whenever by utilizing Application Programming Interface (API) gave by cloud suppliers through any terminal gear associated with the web. Because of the adaptable idea of distributed computing, we can rapidly get to more assets from cloud providers. Cloud computing is proficient and adaptable yet keeping up the dependability of handling such a significant number of employments in the distributed computing condition is an exceptionally complex issue with load balancing[8] getting much consideration for specialists.

Load balancing schemes depending upon whether the system dynamic are vital can be either static or dynamic. Static schemes don't utilize the system information and are less intricate while dynamic schemes will bring additional expenses for the system however can change as the system status changes. A dynamic scheme is utilized here for its adaptability

The cloud computing clients get great quality administrations from their specialist organizations with a reasonable cost. The quality and cost of the services depend on their source designation process in the particular service environment. The provider should appoint the asset to the clients in an optimal way.

The load balancing [8] model is aimed for the cloud which has different centres with appropriated computing resources in an extensive variety of geographic zones. Thus, the model divides the cloud into several cloud partitions. At the point when condition is tremendous and complex, these partitions simplify the load balancing. The cloud has a fundamental controller that picks the appropriate segments for arriving jobs while the balancer for each cloud partition picks the best load balancing methodology. Optimal allocation infers that the amount of request that can be raised, and the power consumed by a request is constrained.

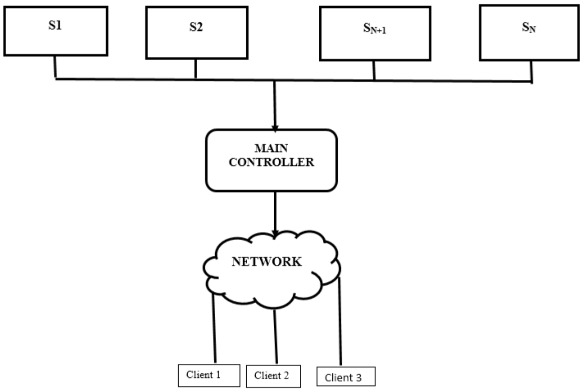


Fig. 1. System Architecture

Resource allocation [9] is the process of granting the resources to the clients according to their need. There are various algorithms which are being used for resource allocation in cloud computing. These algorithms help in scheduling virtual machines on the server at various data centres. In S1 to Sn client application (Exclusive Buy) is running in all servers. Main controller is a responsible to get the request and sent the request to corresponding servers as shown in Fig. 1.

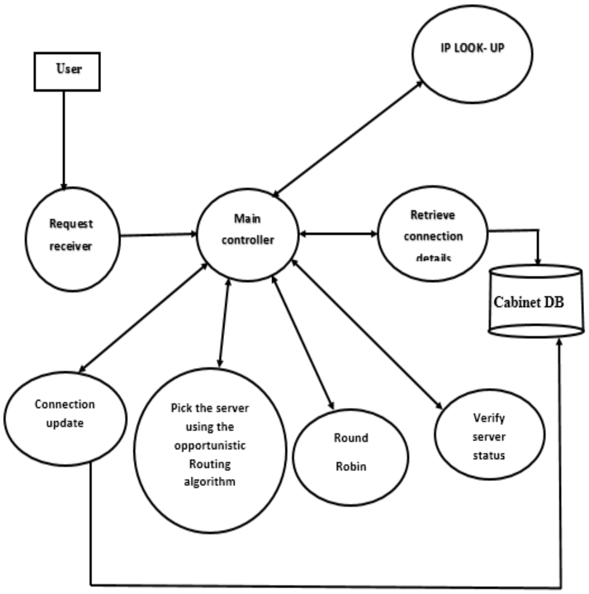


Fig. 2. Data flow diagram

The main controller initially assigns jobs to the reasonable cloud partition and after that communicating with the balancers in each partition to revive the status data. The main controller manages data for each partition, smaller data sets will prompt the higher processing rates. The balancers in each partition assemble the status data from each node and after that pick the correct technique to convey the occupations. Primary controller get the server request from customers, and it get present details from Database and confirm the server status and pick the server utilizing opportunistic routing algorithm on change to shortlisted server as appeared in the Fig. 2.

II RELATED WORK

In [1] "**Service Performance and analysis in cloud computing",** Cloud computing is another cost-effective figuring worldview in which data and PC power can be gotten to execution has turned out to be basic for benefit applications in distributed computing. For the business accomplishment of the new figuring worldview, the capacity to convey Quality of Services (QoS) ensured administrations is vital.

In particular, with an end goal to convey QoS ensured benefits in such a figuring situation and finding the relationship among the maximal number of customers, the insignificant administration assets and the most abnormal amount of administrations are required. The obtained results give the rules of PC benefit execution in distributed computing that would be incredibly helpful in the plan of this new registering worldview.

Advantages: The method provides an efficient and accurate solution for the calculation of probability and cumulative distributions of a customer's response time. It will be useful in the services performance prediction of cloud computing.

In [2] **"Performance evaluation of cloud service considering fault recovery",** Cloud computing is a promising processing worldview which permits circulation of services from a pool of resources. The services are required by the customers through on-request by means of pay and utilize strategy. The best usage of resources and greatest benefits with planning is the principle objective of the cloud specialist co-ops. The significant issue in cloud computing is booking of administrations with enhanced worldwide throughput and occupation planning. Since, distributed computing is an administration based one, the execution assessment is vital criteria to be managed. Need based Queuing model assesses the administrations rented by the cloud specialist organizations. The general administration time, reaction time for arriving demands and the holding up demands are put away in the line. Lining model is developed with markovian entry rate, general administration rate and 'm' number of servers, need line train and a cushion of size 'r'. The advantage of the proposed diagnostic model is within the time traverse, the cloud service provider schedules the services to result in maximum profit.

III PROPOSED WORK

To handle customer requests, different servers are maintained. To serve a customer request, a set of virtual machines will be utilized to execute the request concurrently. To speed up the completion of a customer request, Memory Usage and Central Processing Unit cycles of server is considered. Optimal workload allocation is achieved, if one server is busy, resource is allocated to other server based on memory and Central Processing Unit utilization of the servers.

3.1 MAIN CONTROLLER

The main controller appoints jobs that arrives. Exactly when the load status of a cloud partition is ideal or normal, this partitioning can be capable locally. In the cloud partition if the cloud segment load status isn't typical or sit still, this action should be exchanged to another segment. The segment load balancer by then picks how to dole out the occupations to the hubs. Server stack status is isolated into three sorts. Fig. 3.

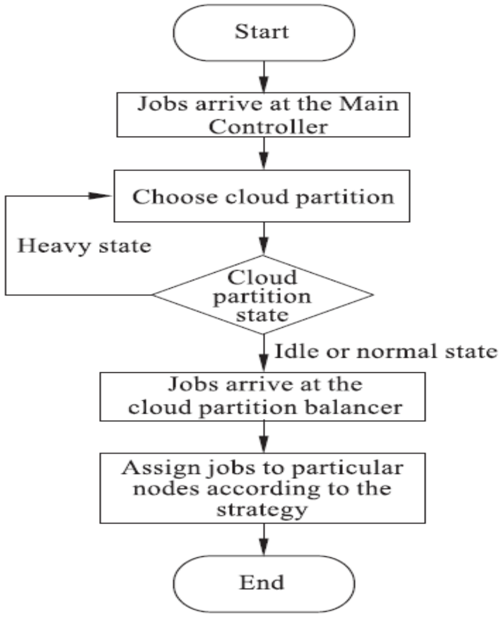

Start
Jobs arrive at the Main Controller
Choose cloud partition
Heavy state
Cloud partition state
Idle or normal state
Jobs arrive at the cloud partition balancer
Assign jobs to particular nodes according to the strategy
End

Fig.3. Flowchart for job arriving at main controller

If one cloud server is over-burden and it again getting new customer requests for while diverse servers are in Idle or Normal state as showed up in the Fig. 3

3.2 METHODOLOGY

The load balancing arrangement is finished by the fundamental controller and the balancers. The load balancing methodology depends on the cloud partitioning ideas. Subsequent to making the cloud partitioning, the load balancing then begins: when a job arrives at the cloud portioning. The system and the main controller chooses which cloud segment ought to get the activity. The partition load balancer at that point chooses how to concede the employments to alternate hubs.

At the point when the load status of a cloud partitioning is typical, this partitioning can be proficient locally. On the off chance that the cloud parcel status isn't typical, this activity ought to be exchanged to another segment. Server load status is separated into three kinds.

- Idle: If it is in idle status, this job should be shifted to another partition by using Round Robin algorithm.
- Normal: If it is normal, this job should be shifted to another partition by using Opportunity Routing algorithm.
- Overload: If it is overload, this job should be shifted to another partition. That partition selected using above two algorithms.

ROUND ROBIN ALGORITHM

```
Step 1: Let Server A is an Overloaded
Step 2: Let s[n] is an array consists of server which
        are in Idle state.
Step 3: Let c=1.
Step 4: If new connection Came for Server A
             then
                   Send the connection to s[c]
               After that make c=c+1;
         If c==n
         then
         c=1
              Else
                   Wait;
Step 5: Send the connection to s[c]
        After that make c=c+1;
        If c==n
   then
         c=1
Step 6: go to step 3.
```

OPPORTUNISTIC ROUTING ALGORITHM

```
Step 1: Let Server A is an Overloaded.
Step 2: Let w[n] is an array consists of server which
        are in Normal state. n is the total number of
        server.
Step 3: If new connection Came for Server A
        then
                Calculate distance of server A and w[n]'s.
           Select minimum distance server
           Send the connection to w[n]
             Else
                 Wait;
Step 4: go to step 3.
```

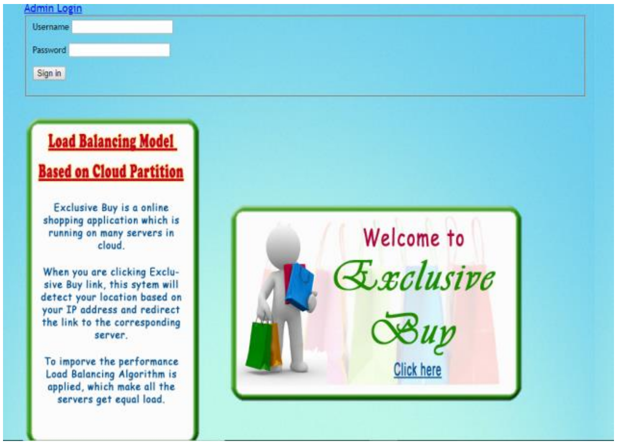
Distance Formula:

$$d=\sqrt{\Delta x^2+\Delta y^2}=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$$

$x_1 \& x_2$=Latitude and longitude of Server A.

$y_1 \& y_2$= Latitude and longitude of w[n].

IV EXPERIMENTAL RESULTS



Admin Login

Username

Password

Sign in

Load Balancing Model Based on Cloud Partition

Exclusive Buy is a online shopping application which is running on many servers in cloud.

When you are clicking Exclusive Buy link, this sytem will detect your location based on your IP address and redirect the link to the corresponding server.

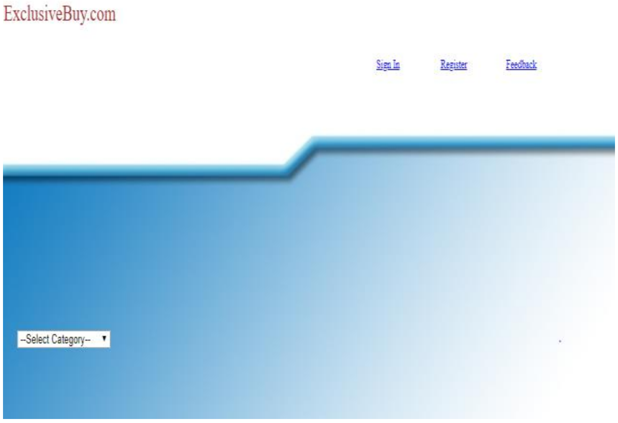
To imporve the performance Load Balancing Algorithm is applied, which make all the servers get equal load.

Welcome to Exclusive Buy

Click here

Fig. 4. Exclusive buy client application login page.

Exclusive Buy is an online shopping application which is running on many servers in cloud. When user clicks the link, the system will detect your location based on your IP address and redirect the link to the corresponding server. The admin should sign in using user name and password as shown in the Fig. 4.



ExclusiveBuy.com

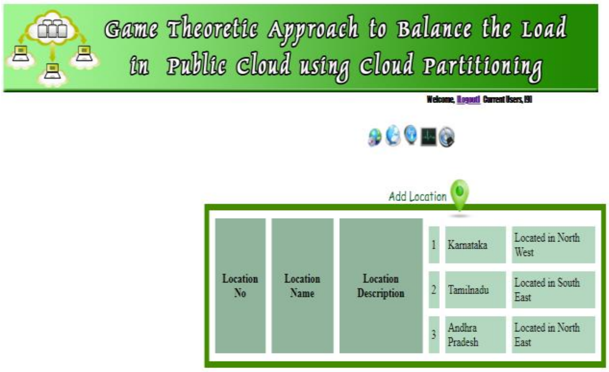
Sign In Register Feedback

--Select Category--

Fig. 5. The home page of Exclusive Buy

The home page of Exclusive buy is shown in the above Fig. 5, where the client can register there details, the client can sign in using the user name and password, can book the products which is available in the application and the client can also feedback.

The Fig. 6, describes the number of servers that can run the client application on it, and shows the location number, location name and location description. Further we can modify the details present in it.



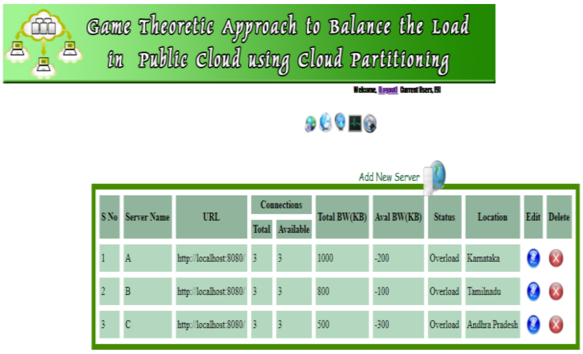
Game Theoretic Approach to Balance the Load in Public Cloud using Cloud Partitioning

Welcome, [Logout] Current Users, [9]

Add Location

Location No	Location Name	Location Description
1	Karnataka	Located in North West
2	Tamilnadu	Located in South East
3	Andhra Pradesh	Located in North East

Fig. 6. Server configuration.



Game Theoretic Approach to Balance the Load in Public Cloud using Cloud Partitioning

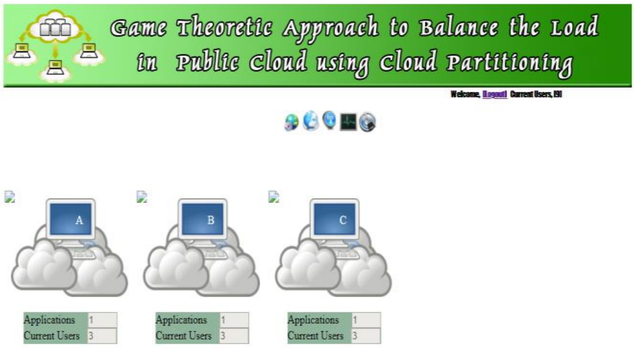
Welcome, [Logout] Current Users, [9]

Add New Server

S No	Server Name	URL	Connections		Total BW(KB)	Aval BW(KB)	Status	Location	Edit	Delete
			Total	Available						
1	A	http://localhost:8080/	3	3	1000	-200	Overload	Karnataka		
2	B	http://localhost:8080/	3	3	800	-100	Overload	Tamilnadu		
3	C	http://localhost:8080/	3	3	500	-300	Overload	Andhra Pradesh		

Fig. 7. Status of each server

The Fig. 7, shows the server name and also the status of each server whether it is idle, normal or overload. It shows the total connection of each server.



Game Theoretic Approach to Balance the Load in Public Cloud using Cloud Partitioning

Welcome, [Logout] Current Users, [9]

	A	B	C
Applications	1	1	1
Current Users	3	3	3

Fig. 8. Monitoring the servers

The fig. 8, shows the connections of each server, the number of users that are connected to the particular server and the applications that is running in the server.

V CONCLUSION

Distributed computing innovation is progressively being utilized as a part of undertakings and business markets. In cloud worldview, a compelling asset allotment technique is required for accomplishing client fulfilment and amplifying the benefit for cloud specialist co-ops. A portion of the systems talked about above for the most part centre around CPU, memory assets yet are inadequate in a few elements. Thus the paper will ideally spur future analysts to think of more intelligent and secured ideal asset distribution calculations and system to fortify the distributed computing worldview.

REFERENCES

[1] K. Xiong and H. Perros, “Service performance and analysis in cloud computing,” in Proc. IEEE World Conf. Serv., 2009, pp. 693–700.

[2] B. Yang, F. Tan, Y. Dai, and S. Guo, “Performance evaluation of cloud service considering fault recovery,” in Proc. 1st Int. Conf. Cloud Comput., 2009, pp.571–576.

[3] H. Khazaei, J. Mi_si_c, and V. B. Mi_si_c, “Performance analysis of cloud computing centers using M/G/m/m+r queueing systems,” IEEE Trans. Parallel Distrib. Syst., vol. 23, no. 5, pp. 936–943, May 2012.

[4] S. Yeo and H. Lee, “Using mathematical modeling in provisioning a heterogeneous cloud computing environment,” in Proc. 3rd IEEE Int. Conf. Netw., 2015, pp. 339–345.

[5] Z. Wang, M. M. Hayat, N. Ghani, and K. B. Shaban, “A probabilistic multi-tenant model for virtual machine allocation in cloud systems,” in Proc. 3rd IEEE Int. Conf. Cloud Netw., 2014, pp. 339–343.

[6] T. Braun, et al., “A comparison of eleven static heuristics for mapping a class of independent tasks onto heterogeneous distributed computing systems”, Parallel Distrib. Comput, vol. 61, no. 6, pp. 810–837, 2001.

[7] G. Ritchie and J. Levine, “A fast, effective local search for scheduling independent jobs in heterogeneous computing environments,” Centre Intell. Syst. Appl., School Informat., Univ. Edinburgh, Edinburgh, U.K., Tech. Rep., 2003.

[8] S. T. Maguluri, R. Srikant, and L. Ying, “Stochastic models of load balancing and scheduling in cloud computing clusters,” in Proc. IEEE INFOCOM, 2012, pp. 702–710.

[9] G. Juve and E. Deelman, “Resource provisioning options for largescale scientific workflows,” in Proc. IEEE 4th Int. Conf. e-Sci., 2008, pp. 608–613.