International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 7 4442 - 4447

Load Balancing and Resource Allocation Model for SaaS Applications with Time and Cost constraints forcloud-computing

Kalyani Ghuge[1]
1PG Fellow, Department of Computer Engg
G.H. Raisoni College of Engineering & Management
Pune, India
kalyani.ghuge@gmail.com

Prof. Minaxi Doorwar[2]
[2]Assistant Professor Department of Computer Engg
G.H. Raisoni College of Engineering & Management
Pune, India
minaxi.rawat@gmail.com

Abstract—Instead of Traditional Software, nowadays we are using Cloud Computing. It enables the on-going revenue for software providers..Advancement of Cloud Computing due to use of well established research in Web Services, networks, utility computing and virtualization has resulted in many advantages in cost, flexibility and availability for service users. These advantages has further increased the demand for Cloud Services, increasing both the Cloud's customer base and the scale of Cloud installations. This has resulted in many technical issues in Service Oriented Architectures and Internet of Services (IoS) type applications such as high availability and scalability, fault tolerance. Central to these issues is the establishment of effective load balancing techniques. In this paper focus on the load balancing and resources provisioning approaches.Here, using the linear programming approach for dynamically allocates the resources with balancing the load.Mainly focus on the time and cost constraints.

Keywords- *Cloud Computing,LoadBalancing,Resource provisioning,Service Oriented Architecture.*

*****

I. INTRODUCTION

The Word Cloud Computing is buzzing everywhere among organization, enterprises, independent software vendors (ISV), end users etc. Cloud computing is nothing but distributed computing over the internet where user can access their data from the database in the cloud. Cloud computing is different from traditional grid computing it is more dynamic, flexible and scalable offered by independent organizations where deployment and maintenance of the services & data is managed by the organizations themselves.[1][2] Cloud computing varies from one cloud provider to another, as some cloud providers provide storage over network with small monthly rentals for end users, whereas some other providers offer applications for software companies which helps in reducing costs in deployment or installations of applications[3].Cloud computing signifiesmain changes in how to accumulation information and run application. The Software sales model available in the market and this model require customers.[4] Customers need to purchase software and manage the deployment themselves. If the customer requires software for specific period of time then also need to pay full amount for the license copy of the software.

Here, the Cloud computing concepts come into attention. Cloud is based onpay as per use model. [5][6] The customers only pay for the period of time how much they used up software. Not necessary to buy the license copy of the software. This is the main benefit of the cloud computing which is mostly used by the industry, colleges and many users. Cloud computing is useful and scalable service which provides efficient services to the cloud users. Numbers of the loads are present in cloud computing environment.So, maintaining the constancy of processing is a very complex problem. Here, load balancing gets much consideration. Resources can also be shared by more than one customer. One cloud can connect to another cloud anywhere and anytime.. The command and control of the cloud is handled by cloud management. Still there is need of progress in this area. Research and development is being made to make the cloud self-managed.[7]The Load balancing and resource provisioning is focused in this paper and aimed at the private cloud which has number of the nodes distributing among computing resources. This model divides the load into the number of partitions. Categorize the clusters of processors with their cost. The Linear Programming is used for dynamically allocating the tasks to the processors with minimizing cost and time. Resources consider the CPU, memory, bandwidth. Dynamically allocating the jobs to each processor shows the utilization of the resources after completion of the number of task on each node. Analyze the cost and time of each job.

II. RELATED WORK

VectorDot proposed a novel load balancing algorithm called VectorDot. It handles the hierarchical complexityof the data-center and multidimensionality of resource loads across servers, network switches, and storage in an agile data center that has integrated server and storage virtu-alization technologies. VectorDot uses dot product to distin-guish

nodes based on the item requirements and helps in removing overloads on servers, switches and storage nodes.[8] CARTON proposed a mechanism CARTON for cloud control that unifies the use of LB and DRL. LB (Load Balancing) is used to equally distribute the jobs to different servers so that the associated costs can be minimized and DRL (Distributed Rate Limiting) is used to make sure that the resources are distributed in a way to keep a fair resource allocation. DRL also adapts to server capacities for the dynamic workloads so that performance levels at all servers are equal. With very low computation and communication overhead, this algorithm is simple and easy to implement. Compare and Balance addressed the problem of intra-cloud load balancing amongst physical hosts by adaptive live migration of virtual machines. A load balancing model is designed and implemented to reduce virtual ma-chines' migration time by shared storage, to balance load amongst servers according to their processor or IO usage, etc. and to keep virtual machines' zero-downtime in the process.[9][10] A distributed load balancing algorithm COMPARE AND BAL-ANCE is also proposed that is based on sampling and reaches equilibrium very fast. This algorithm assures that the migration of VMs is always from high-cost physical hosts to low-cost host but assumes that each physical host has enough memory which is a weak assumption. To find the reliability of the system which handle the load consider the some factors such as Throughput used to calculate number of task whose execution been completed in unit time. In given scale of time throughput should be high to improve the performance of the system.Response Time defined as amount of time taken in distributed cloud environment to riposte with a load balancing methodology. The response time should be minimized for effective system performance.[9][10] Resource Allocation algorithms for SaaS providers minimize infrastructure cost and SLA violations. Various algorithms are designed to check the work of SaaS providers. Dynamic change of customers, mapping customer requests to infrastructure level parameters and handling heterogeneity of Virtual machines are some of the tasks managed by the SaaS providers. Also customers Quality of Service parameters such as response time, and infrastructure level parameters such as service initiation time are also considered. By taking into consideration, the predefined SLA clauses and submitting their QoS parameters, the customer's requests for the enterprise software services from a SaaS provider are considered. The requirements and the usage of the hosted software services can be managed are accordingly changed by the customers. The SaaS provider can use their own infrastructure or outsourced resources from public IaaS providers. The main objective of SaaS providers is to work in such a way that its profit is maximized while the customers' requirements are also assured. The platform layer of a SaaS provider uses mapping and scheduling mechanisms to interpret and analyze the customer's QoS parameters, and allocates respectively.SaaS providers lease enterprise software as hosted services to customers. To increase their reputation in the marketplace, they are only interested for maximizing profit and ensuring QoS for customers.[10][11]

In linear scheduling strategy the resource allocation is taken into thought usually the parameters like CPU utilization, memory utilization and throughput etc. The cloud environment has got to take into consideration of these things for every of its clients and could offer maximum service to all of them. It suggests that when we are taking the scheduling of resources and tasks in an individual basis it imposes giant waiting time and response time. So as to beat this drawback a new approach specifically Linear Scheduling for Tasks and Resources (LSTR) is introduced. Here scheduling algorithms mainly target on the distribution of the resources among the requestors which is able to maximize the chosen QoS parameters. The QoS parameter selected in this approach is the cost function. The scheduling algorithm is designed based on the tasks and the available virtual machines together and named LSTR scheduling strategy. This is often designed so as to maximize the resource utilization. [11]

The scheduling algorithm is meted out based on the prediction that the initial response to the request is formed solely when assembling the resource for a finite amount of time (say 1 day or 1 hr. like that) but not allocating the resource as they arrive. However dynamic allocation could be carried out by the scheduler dynamically on request for a few extra resources. This is often achieved by the continuous evaluation of the threshold value in the system. The authors states that this approach suitable when we consider the "shortest job first (SJF)" instead of the "first come first serve (FCFS)" way of scheduling. The algorithm sorts the requests by excluding the arrival times. It solely considers the "threshold" of the request for the scheduling purpose. In Pre-Copy Approach for scheduling talks regarding the live migration of the virtual machines. Clark et al. Suggest that migration of the operating system instances across distinct physical hosts is a great tool for the administrator of data centers and clusters. It in addition offers a separation between hardware and software and provides fault management, low level system maintenance and load balancing. Clark et al. came out with the idea of "pre-copy approach". In this approach pages of memory are repeatedly copied from the source machine to the destination host and additionally there is an undeniable fact that all these things are done without ever stopping the execution of the system.[12] Page level protection hardware is employed to make sure that a consistent snapshot is transferred. For controlling the traffic of different running

services a rate-adaptive algorithm is used. And during the final phase it pauses the virtual machine and copies any leftover pages to the destination and afterwards resumes the execution there. The factors touching the total migration are link bandwidth, migration overhead and page dirtied rate. In the Match Making and Scheduling tells that the "Match making" is the first step and "scheduling" is second within the resource allocation in cloud environment. [13]Matchmaking is that the method of allocating jobs associated with user requests to resources designated from the obtainable resource pool. Scheduling refers to determining the order in which jobs mapped to a selected resource are to be executed. It additionally tells that there are some uncertainties that are associated with such type of "match making" and scheduling

III. SYSTEM MODEL

Figure1. Show the flow of the system. It is made up of of the central server, number of the processors and cloud users. Users are first fill the sign in form and mentioned their approximate cost and get the login on cloud service provider. After that user are ready to submit their tasks on the central server. Several users can login into cloud service at the same time and submit their jobs to cloud central server.Here consider the private cloud environment. Private cloud is constructed on standard cloud computing model, which delivers software as service to cloud clients. Private cloud contains several nodes (processors),number ofusers and the cloud server. Load balancing strategy and resource provisioning is based on the k-means clustering algorithm and linear programming.

After generating the cloud environment load balancing is started.When a task reaches at central server, tasks are breaks into number of chunks. After partition, resolve how to allocate tasks to each of the nodes (processors). Linear programming decides which partitions are assigned to which processors.

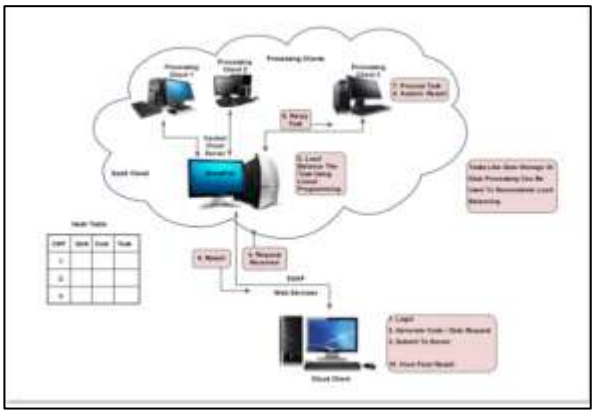

Figure 1: System Flow

The whole process is shown in figure 2.The load balancing solution is designed by the cluster of the processors on the basis of cost and the linear programming.

In the system model cloud users submit their tasks to the server for the processing. At the same time all the nodes extract their own properties such as CPU, memory, speed, cores etc. And submit it to the cloud server. The cloud server applies the K-means algorithm finds the centroid and assigns cost to the particular processor according to their properties or the configurations. Here, mainly three clusters are formed low,medium and high. Processors clusters are formed on the basis of the cost once the cost is assigned to the processors they check for the availability of the jobs on the central server. Once task is finding they start to process the job and after completion of the tasks assign it to sever. The serverassigns processed task to the user with the user requirement and the satisfaction. The main heart of the model is tasks are assigned to processors with the user satisfactions such as minimum cost and time. Use the linear programming to find the minimum cost and time of the processors which process the jobs of the users. Linear programming is used for the dynamically assign jobs to the cluster nodes with minimum time and the cost.

A. Job assignment Strategy

The processors wrinkle their properties and submit it to cloud central server. Server responds with cost and three clusters. The task reaches at central server, whichbecomesseparated into a number of pieces and reliant upon the users' prerequisite the tasks areallotted to a specific processor. The central focus in this article is that the tasks should allot the processors with minimum cost and processing time. The idea of linear programming is used for allocating tasksto nodes of the clusters.

Linear programming contributes the effects of great resource provisioning with minimum cost and processing time, which provides optimal solution towards the load balancing and resource provisioning. LP considers the constraints such as cost per partition of the computers which are categorized into low, medium and high; and also the minimum time to process the petitions. Other constraints are time per partitions in CPU cycles for low, medium and high performance of the computer. Solve the equations. [14] [15] Then by using the results we decide which job should be assigned to which processors with minimum cost and time. These are the parallel processes where jobs are dynamically assigned to each node by using the LP logic.

The processor cost log is generated at server side at each time when new processor adds in the cloud environment. The log refreshes after a fixed period of time

 ISSN: 2321-8169
4442 - 4447

T. The cluster log is created along with their category and is maintained accordingly. These tables are used by the LP solver which calculates the status of each processor with minimum cost and time. Tasks are allotted to processors based on the present load balancing approach. This approach is changed when linear programming constraints changes.

B. Mathematical Model

Let S be System such that,

{P, S, Se, SDb, J, F, f, O }

S is a system that divides into the subsystem Submitters, Processors, Functions, and Output.

P is an infinite set of the processors.

$P = \{P_1, P_2 \ldots , P_n\} \in P$

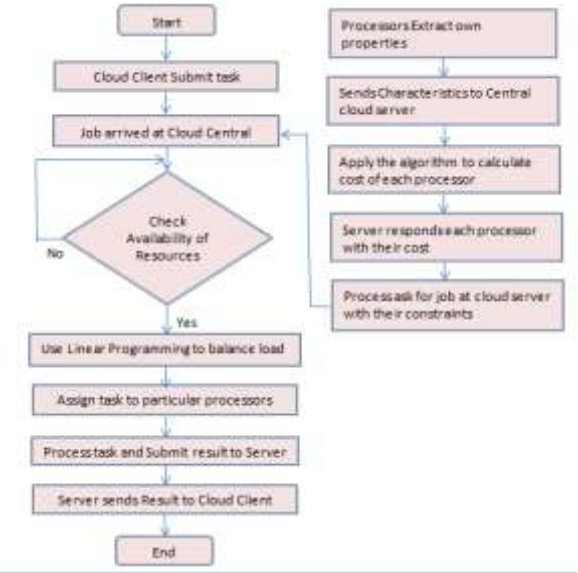

Figure 2: Job assignment Strategy

S is an infinite set of the submitter.

$S = \{S_1, S_2 \ldots .., S_n\} \in S$

Se is the server. The server is many or may be one.S_{Db} is the server database which is used as MYSQL. The server database consists the how many processors manages the job, Submitters and the processors

J is an infinite number of the jobs which is submitted by the submitter

$J = \{J_1, J_2 \ldots J_n\} \in J$

f is the file which the cloud user submitted to the server for the completion of the task.

F is a set of functions' = { F_{upload}, $F_{brakjob}$, $F_{assignjob}$, $F_{processjob}$, $F_{fetchjob}$}

[1] F_{upload} = Users upload jobs on the cloud sever by using this function.
[2] $F_{brakjob}$ = Divide the load on the basis of the constraints using the linear programming.
[3] $F_{assignjob}$= Assign divided job to the processors. The processorconsiders the quality of the factors such as cost, utilization of resource etc.
[4] $F_{processjob}$= Processors process the jobs which they have and submit processed job to a server.
[5] $F_{fetchjob}$= Users fetch the processed job from the server.

Set of the output.

O = {O1, O2, ….. , On}

Prosessor process the load and submit the result after proper load distribution and task completion to the server and then server submit processed job to submitter.

C. Algorithm Steps

Algorithms Used in Proposed Solution:-

[1] Collection of nodes properties.(processor speed,RAM,etc..)
[2] Set of task submission from clients to server(Provider).
[3] Apply k-means algorithm and respond with cost to each node.
[4] Generate clusters by considering properties & categorize them in to high, medium, low cost of node.
[5] Consider execution time required for given partitions for allocated task.
[6] Find total number of partitions processed by each node using linear programming.
[7] Find:

- Min(TC) = (PartL * PartCostL) + (PartM * PartCostM) + (PartH * PartCostH)
- Min (TT) = (PartL * PartTimeL) + (PartM * PartTimeM)+(PartH * PartTimeH)

1. Combine the all partitions processed by different nodes.
   - PartH + PartM+ PartL = TotalPartx

D. Assumptions

- Processes have been split into tasks
- Computation requirement of tasks and speed of processors are known
- Cost of processing tasks on nodes is known.
- Resource requirements and available resources on a node are known
- Reassignment of tasks is possible

IV. EXPERIMENTAL RESULTS

. Figure 3. Show the resource utilization in the percentage. Here, consider the number of Clusters and each consists of different number of the inputs and the load.Clusters contain the number of nodes.

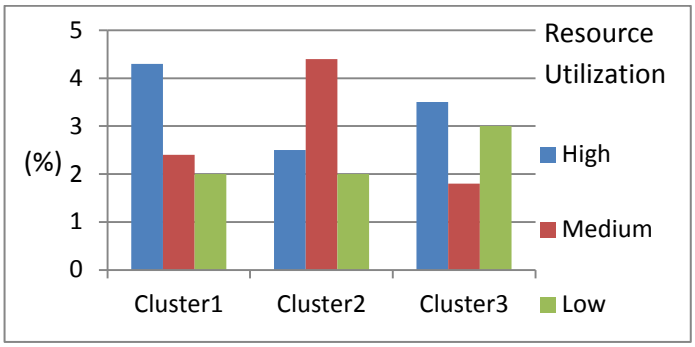

Figure 3. Resource Utilization

Figure 4. Show the size of the images in the kb and the time required for processing the image.The graph shows the required the minimum response time for processing the images.

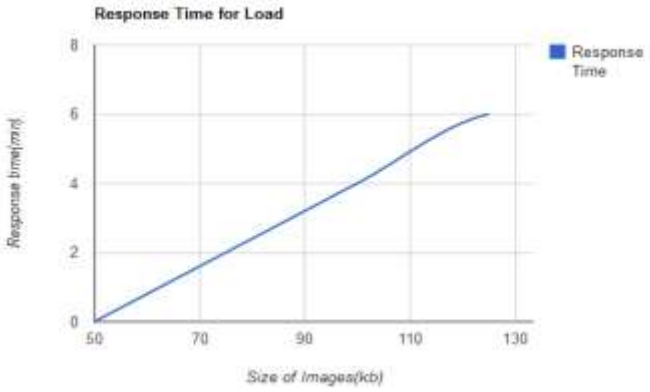

Figure 4. Processing Of Images

Figure 5. Shows the job start timing and time required to process the jobs the log of the job processing is shown below

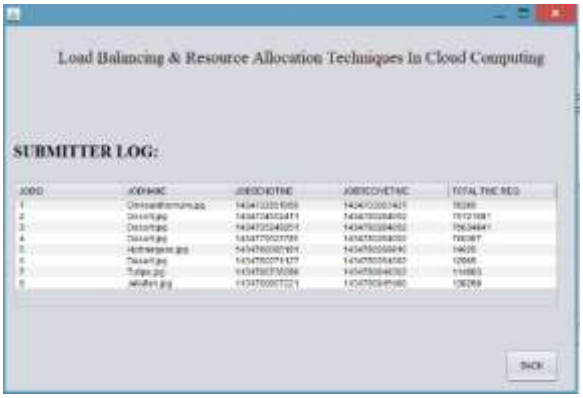

Figure 5. Time required to process jobs.

V. CONCLUSION

Load balancing algorithm based on the linear programming dynamically allocates the workload ro different processors with minimum time and the cost.Dynamically allocation of the resources is the main concern with the cloud computing which achived by algorithm.The main purpose of load balancing is to satisfy the customers requirement by distributing dynamically workload across the processors.Performance of the system increases with high utilization of the resources and minimum time .

REFERENCES

[1] Prabavathy.B, Priya.K, Chitra Babu "A Load Balancing Algorithm For Private Cloud Storage" 4th ICCCNT 2013 July 4-6, 2013, Tiruchengode, India IEEE – 31661

[2] Zhen Xiao, Senior Member, IEEE, Weijia Song, and Qi Chen "Dynamic Resource Allocation Using Virtual Machines for Cloud Computing Environment"

[3] Tushar Desai, Jignesh Prajapati "A Survey Of Various Load Balancing Techniques And Challenges In Cloud Computing" International Journal of scientific & technology research volume 2, Issue 11, November 2013.

[4] Harpreet Kaur,Maninder Singh "A Task Scheduling and Resource Allocation Algorithm for Cloud using Live Migration and Priorities" International Journal of Computer Applications (0975 – 8887) Volume 84 – No 13, December 2013.

[5] A.Meera , S.Swamynathan "Agent based Resource Monitoring system in IaaS Cloud Environment" International Conference on Computational Intelligence: Modeling Techniques andApplications (CIMTA) 2013.

[6] Martin Koehler,"An adaptive framework for utility-based optimization of scientific applications in the cloud" Koehler Journal of Cloud Computing: Advances, Systems and Applications 2014,Springer.

[7] Zheng Hu, Kaijun Wu, Jinsong Huang" An Utility-Based Job Scheduling Algorithm for Current Computing Cloud Considering ReliabilityFactor" ©2012 IEEE

[8] Linlin Wu, Saurabh Kumar Garg and Rajkumar Buyya "SLA-based Resource Allocation for Software as a Service Provider (SaaS) in CloudComputing Environments" 2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing

[9] Divya Chaudhary ,Rajender Singh Chhillar" Reverse Host Allocation Approach for Virtual Machine Cloud Computing Environment" International Journal of Computer Applications (0975 – 8887) Volume 72– No.17, June 2013.

[10] Gaochao Xu, Junjie Pang, and Xiaodong Fu "A Load Balancing Model Based on Cloud Partitioning for the Public Cloud" IEEE Transaction on cloud computing Year 2013.

[11] David W Chadwick*, Matteo Casenove ,Kristy Siu"My private cloud – granting federated access tocloud

resources" Chadwick et al. Journal of Cloud Computing: Advances, Systems and Applications 2013,Springer

[12] Ratan Mishra,Anant Jaiswal "Ant colony Optimization: A Solution of Load balancing in Cloud" International Journal of Web & Semantic Technology (IJWesT) Vol.3, No.2, April 2012

[13] Radojevic, B., & Zagar, M. (2011)." Analysis of issues with load balancing algorithms in hosted (cloud) environments". In MIPRO, 2011 Proceedings of the 34th International Convention, 416-420. IEEE.

[14] Lori M. Kaufman,Bruce Potter "Monitoring Cloud Computing by Layer, Part 1" Co-published by the IEEE computer and reliability socities IEEE MARCH/APRIL 2011.

[15] Zenon Chaczko , Venkatesh Mahadevan , Shahrzad Aslanzadeh and Christopher Mcdermid. "Availability and Load Balancing in Cloud Computing." 2011 International Conference on Computer and Software Modeling IPCSIT vol.14 (2011) © (2011) IACSIT Press, Singapore.