مجلة العلوم البحتة والتطبيقية
Journal of Pure & Applied Sciences
www.Suj.sebhau.edu.ly ISSN 2521-9200
Received 23/04/2019 Revised 21/10/2019 Published online 17/11/2019

Developing a load balancing technique in cloud computing

*Lina Samir Malouf , Bassim Oumran
Department of Automatic Control Engineering and Computers, Faculty of Mechanical and Electrical Engineering, Homs Al-Baath University, Syria
*Corresponding author: Leena_m84@hotmail.com

Abstract Cloud computing is a utility to deliver services and resources to the users through high speed internet. It has a number of types and hybrid cloud is one of them. Cloud computing provides scalable services that include SaaS, PaaS and IaaS over the Internet. Because of the dynamic nature of cloud environments, the VM workload swings dynamically, resulting in unbalanced loads. Cloud Load Balancing is the process of distributing workloads and computing resources in a cloud computing environment. Enterprise load balancing lets you manage application requests or workload by allocating resources across multiple computers, networks, or servers. In this paper a load balancing technique in cloud computing is developed to distribute the work load across multiple nodes in order to insure that no single node is overloaded. The proposed technique is expected to provide maximum throughput with minimum response time.

Keywords: Cloud, Cloud computing, Load balancing.

تطوير تقنية لموازنة الحمل في الحوسبة السحابية

*لينة سمير معلوف و بسيم عمران

قسم هندسة التحكم الآلي والحواسيب– كلية الهندسة الميكانيكية والكهربائية– حمص جامعة البعث، سوريا

*للمراسلة: Leena_m84@hotmail.com

ملخص الحوسبة السحابية هي أداة لتقديم الخدمات والموارد للمستخدمين من خلال الإنترنت عالي السرعة. لديها عدد من الأنواع والسحابة الهجينة واحدة منها. تقديم الخدمات في السحابة الهجينة مهمة شاقة. أحد التحديات المرتبطة بهذا النموذج هو التوزيع المتساوي بين موارد السحابة الهجينة، والتي يتم الحكم عليها غالبًا على أنها موازنة التحميل. من خلال الاستفادة من موازنة التحميل، يمكن تحسين وقت الاستجابة الوظيفية. هذا سوف يؤدي لتحسين نتائج الأداء. الحل لمثل هذه السيناريوهات باستخدام موازن التحميل وهو مكون يحافظ على مجموعة من الخدمات (عادة نفس الخدمات ولكن يعمل على أجهزة مختلفة) ويتعرض لتلقي الرسائل بدلاً من الخدمة الحقيقية. لا يعرف العميل عن هذا الإعداد ويعتقد أنه يتصل مباشرة بالخدمة. في هذه الورقة، تم تطوير تقنية موازنة التحميل في الحوسبة السحابية لتوزيع عبء العمل عبر عقد متعددة لضمان عدم وجود زيادة في عقدة واحدة. من المتوقع أن توفر التقنية المقترحة أقصى إنتاجية مع وقت استجابة أدنى.

الكلمات المفتاحية: السحابة، الحوسبة السحابية، موازنة التحميل.

1. Introduction:

Cloud computing or the next-generation computing future offers customers access to the virtual network for applications or services. Regardless of where the client accesses the service, it is automatically routed to the available resources. Load Balancing helps prevent a server or network device from devouring requests and helping distribute work.

Load balancing is generally presented as a "service", the services in the cloud consist of 3 types:

Infrastructure as a Service (IaaS): the infrastructure layer is build on the virtualization layer by providing virtual machines as a service to users. Instead of buying servers or even hosted services, IaaS customers can create virtual machines and network them together as desired. Platform as a service (PaaS): The platform layer is based on the virtual machines of the infrastructure layer. In this layer clients do not manage their virtual machines; they just create applications within the programming API or language. There is no need to manage an operating system. Software-as-a-Service (SaaS): Software-level services consist of complete applications that do not require development. Such applications can be e-mail,

Customer Relationship Management and other office productivity applications. Bills can be invoiced monthly or by use, while the software is provided as a service provided directly to consumers, such as e-mail, free of charge.

Cloud allocation of resources to users on demand creates a load balancing problem. If the workload is not distributed properly, some nodes will be loaded into the cloud heavily and some nodes will be loaded. Similarly, if the resources provided by the cloud are not allocated efficiently, they delay service delivery to users [3]. An imbalance may cause system throttling to load. In order to make use of resources and any delay in service delivery, resources should be allocated in an effective manner [3].

The system nodes can be grouped logically into the group and the load balancing task is distributed among the groups. Each individual group will distribute the load to the nodes that belong to that group. This can be arranged in a hierarchical format. For the cloud environment, different load balancing methods were applied to provide an efficient load distribution among available devices. Such as Round Robin load balancing, throttle load balancing, Min-Min load balancing, Min-Max load balancing, honey-based load balancing, ant colony, etc. To balance effective loads, the single load balancing algorithm is insufficient. Hence there are algorithm requirements that combine two or more load balancing algorithms.

The load balancer is an important component in the cloud system that provides a high level time saving, and effective consumption of resources before load allocation between different cloud nodes. Beside its rule in resolving the virtual machines exploit problem. Load balancing provides a solution to the communication overhead problem and focus on increasing productivity, enhancing resources usage and reduce response time. It is a prerequisite for maximizing cloud performance and take advantage of resources strongly. Load Balancing can make use of cloud resources that have been strengthened through the allocation of resources they have Perform presets [2]. Effective load balancing: The critical concept of cloud computing helps determine which device within a given servers pool is best and able to process an incoming data packet to optimize the resource benefit. Thus the load will be distributed among sources on the cloud, so that each resource does the same The amount of work in any given time is allocated by a Load balancer. The load balancer controls different demand Customize for different servers. The load balancer uses different Algorithm to organize the server it has to deal with the demand. [1] we will propose load balancing algorithms Achieve high productivity and reduce response time.

2. Literature review:

There are a number of load-balancing algorithms that work on achieving their mission on different layers of clouds and different level of complications. To improve load balancing, researchers aim to develop more complex load balancing algorithms. But while processing some constants may also increase, like load handling and execution time. Most Balancing algorithms can be classified based on Spatial distribution of the contract (topology) and the environment.

Load balancing is one of the major important problems of heterogeneous computer networks. To work around this problem, many central approaches have been suggested in the literature and had proved its role in raising scalability tribulations. A comparative analysis of various load balancing algorithms was provided(Honey bee, random sampling biased, active groups). Their analysis highlighted that the honey bee algorithm has maximum productivity with increased system diversity compared to the other two algorithms.

The honey bee algorithm is stimulated by the behavior of biological bees that move in search of their food. Similarly, in load balancing, there are virtual servers that provide virtual services. Each server requires services calculates throughput and publishes it on the billboard. Servers interested in applying also have their own throughput account and compare it with colony throughput. If the case is related to the high-throughput server the interest of the colony serves the current virtual server returns to scout behavior, ie randomly selected another server [12].

in honeybee foraging algorithm, throughput does not increase with the increase in system size. Biased random sampling and active clustering do not work well as the system diversity increases. The main problem is that throughput is not increased with an increase in system size. When the distant population of service types is required then this algorithm is best suited.

A scheduling mechanism has been proposed based on the genetic algorithm for load balancing between virtual machines. This mechanism determines the virtual machine with shortest practicable job to be chosen first and improves the higher level cost of migration. However, due to the large number of virtual machines and frequent service requests in the data center, there is an opportunity to schedule an inefficient service. Shortest job had a benefit that the waiting time for the processes is a smaller amount that makes it a robust approach. The genetic algorithm approach computes the impact in advance, that it will have on the system after the new VM resource is deployed in the system, by utilizing historical data and current state of the system. It then picks up the solution, which will have the least effect on the system. By doing this it ensures the better load balancing and reduces the number of dynamic VM migrations[13],[14].

Another algorithm was used for load balancing in the general cloud using game theory. The strategy behind the game theory is that the any of the player's option that is used in the setting where the outcome of the action not only depend on that player but also on the action of the other players. The player strategy will determine the stage of the game. By using game theory in load balancing, it provides the fairness to the users and reduces the response time of the server[16].

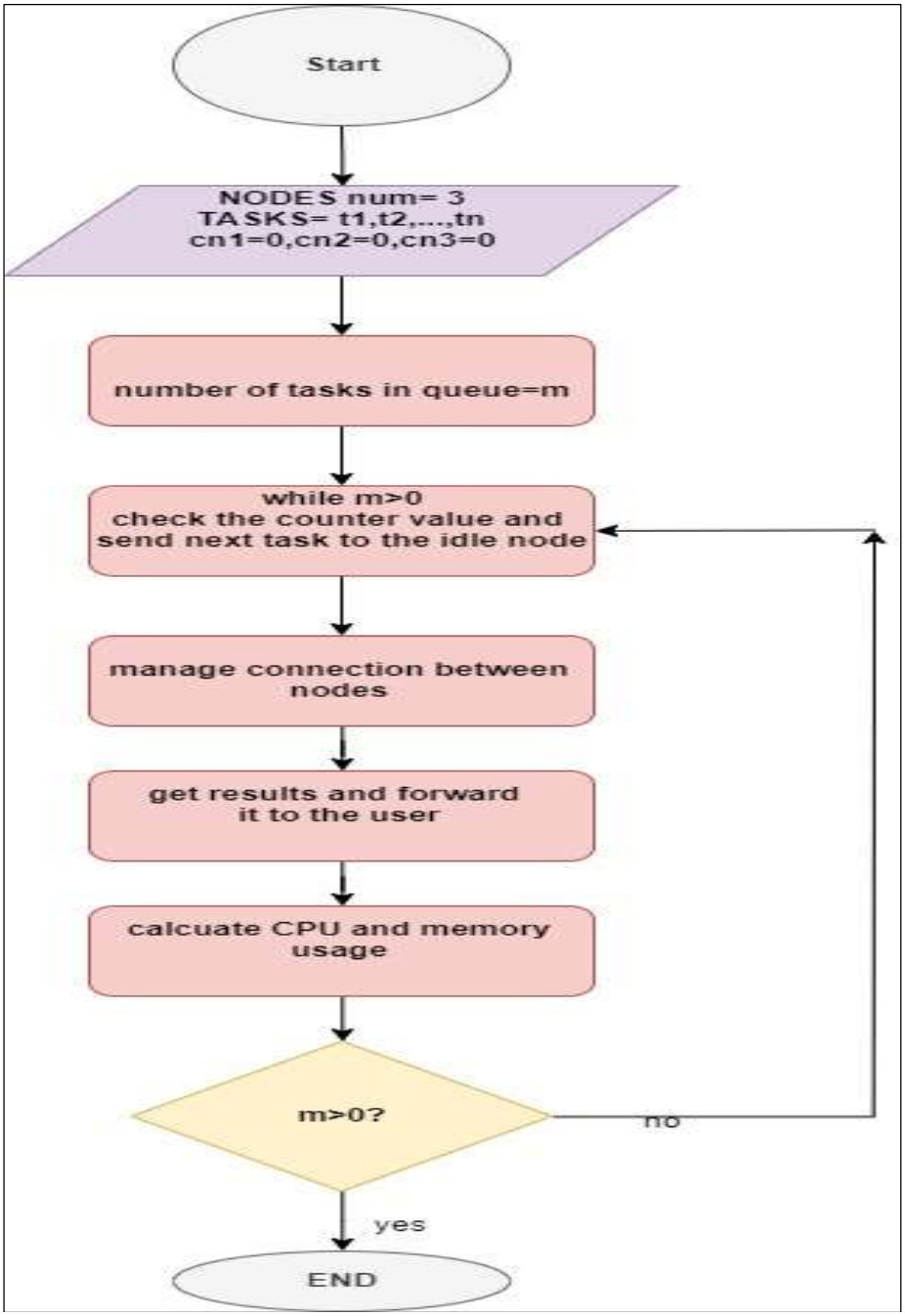
In [15] proposed two static algorithms in Cloud environment[6][7]. One is the opportunistic load balancing, in which the incoming functions received by a nodule are present minimum execution time calculated by the Service Manager. The second is the Min balance load which gets better use resources by maintaining load balance. However, both algorithms are not support dynamic environments [7].

H. Liu[9] proposed a load balancing virtual storage strategy (LBVS) that provides a large scale net data storage model and storage as a service model based on Cloud storage. Storage virtualization is achieved using an architecture that is three-layered and load balancing is achieved using two

load balancing modules. It improves the efficiency of concurrent access by using replica balancing which reduces the response time and enhances the capacity for backup. LBVS improves the flexibility, robustness and use rate of storage resource of the system.
Nakai et. Al [11], presented a distributed server based technique for web servers. It facilitates the reduction in service response time by using a protocol that bounds the redirection of requests to the closest remote servers without overloading them. A middleware is used in this technique to implement this protocol. To endure overload, web server uses Heuristic. Heuristic scheme provide a surety to get balance of load based on the job size and also guarantee to not get repetition of same size job in single node..
It has been noted from the literature that there are some obstacles such as the constant nature of the load balance algorithms, scalability and reliability. In addition from the analysis of literature it is noted that artificial Intelligent mechanism such as genetic algorithm, honeybee algorithm, game theory and smart factors were used for load balancing in cloud computing, which highlights that researchers have found them suitable, applications so that there is a scope to employ them even more. Thus there is a strong need for effective loading mechanism of balance in cloud computing [10].

3. The proposed work.:
The proposed system demonstrates the communication scenario where the capacity of one running node/server/VM is not enough and we want to increase the performance by distributing the work load across multiple nodes.
To ensure that no single node is overloaded that we can get more utilization of the resources and improve the system performance.
On the other hand, we can detect the overloaded nodes then transfer the extra load to other nodes, so it can help to provide maximum throughput with minimum response time.
Let's assume an application as a service in cloud computing that performs some calculations algorithms.
We will define a load balancer as an intermediate layer which is responsible for transferring the end user's request to the cloud and responsible for monitoring the load of each node.
This load balancer layer is transparent to the end user and the application will communicate directly with the load balancer layer [9].
When the end user sends a request, the load balancer receives this request and to decide which node to pick from the available nodes, the load balancer multiple strategies to transfer, select, locate and collect information. Therefore, all nodes are chosen evenly The proposed system is a dynamic load balancing algorithm; it's decision depends on the work load at every node.
We have to collect information about the system status. We assume a cloud system consists of a server and three nodes, every node have a counter variable, in order to get information about every node, which make sure that the load is distributed equally to improve the whole performance, reduce the delay time and improve the cloud response time.
Every node is assigned to a counter, the counter refers to the number of tasks are handled in this node now.
The load balancer will monitor the load of every node. When the user sends a request to the cloud application, we can consider the load as the CPU load and memory usage. As the proposed work detailed in the flowchart Figure (1).

1- Request sent by end user to the web application.
2- The load balancer layer is triggered to perform the end user request.
3- It receives requests from the end user, devide the request into multiple smaller tasks (in a queue) and forwards them to the available nodes to balance the workload.
4- The load balancer get information about the availabe nodes by the conuter to select the abbrobriate node with lowest counter value to assign task to.
5- Increase the conter by one.
6- Match the task to the selected node by the transfer strategy.
7- Do computations about the CPU utilization and memory usage for every node.
8- Get results from nodes and forward answer to the end user.
9- Decrease the counter in every node by one.



Figure(1): The proposed work

4.Conclusion and future work:

Cloud computing is a tool to deliver services and resources to users through the internet at high speed. It has a number of species and hybrid cloud is one of them. Since part of it is special, it is considered as such more secure but hybrid cloud design is a difficult task because the complexities involved. Some benefits of hybrid drag are optimal resource utilization, risk transfer, availability, and hardware reduction cost and best quality of service. However, there are many associated challenges as well with mixed clouds as described. Some are bootable mobility, cost, security, reliability, monitoring, denial of service, load balancing. Since cloud computing is a broad area of research, one of the main topics of research is balancing dynamic loads, the following research will focus on the algorithm taking into account two basic parameters first, the load on the server and second, the current performance of the server.

The objective of load balancing is to increase customer satisfaction, increase resource utilization, increase cloud system performance, reduce response time, reduce the number of work rejects, thereby reducing energy consumption and carbon emissions.

In this paper we have proposed a distributed and scalable load balancing mechanism which is a load balancer layer transparent to the end user and the application will communicate directly with the load balancer layer for cloud computing which will provide maximum throughput with minimum response time.

In future study we will use CloudSim tool for simulation the proposed algorithm and considers Datacenter, Virtual Machine (VM), host and Cloudlet components from CloudSim for execution analysis of a few algorithms. Datacenter component is used for handling service requests. VM consist of application elements which are connected with these requests, so Datacenter's host components should allocate VM process sharing.

References:

[1]- "Fault Tolerance- Challenges, Techniques and Implementation in Cloud Computing" Anju Bala1, Inderveer Chana2.

[2]- A.N. Ivanisenko; T. A. Radivilova , "Survey of major load balancing algorithms in distributed system ", Information Technologies in Innovation Business Conference (ITIB), 2015 ,Pages: 89 - 92, DOI: 10.1109/ITIB.2015.7355061

[3]- Advanced Information Networking and Applications Workshops.

[4]- Ashalatha R; J. Agarkhed, "Dynamic load balancing methods for resource optimization in cloud computing environment ", 2015 Annual IEEE India Conference (INDICON)

[5]- G.PunethaSarmila,Dr.N.Gnanambigai, Dr.P.Dinadayalan," Survey on Fault Tolerant – Load Balancing Algorithms in Cloud Computing", IEEE Sponsored 2nd International Conference On Electronics And Communication System (ICECS 2015), Pages- 1715-1720

[6]- Jain, S., *A Survey of Load Balancing Challenges in Cloud Environment.* Proceedings of the SMART -2016, IEEE Conference ID: 39669, 2016.

[7]- Pop F, Cristea V, Bessis N, Sotiriadis S Reputation guided Genetic

[8]- Priyanka Singh, P.B., *Assorted Load Balancing Algorithms in Cloud Computing: A Survey.* International Journal of Computer Applications (0975 – 8887), 2016.

[9]- Riky subtra et.al , " Game Theoretic Approach for Load Balancing in Computational Grids", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS", vol.19.

[10]- Scheduling Algorithm for Independent Tasks in Inter-Clouds Environments.

[11]- Shridhar G.Domanal and G.Ram Mohana Reddy, " Load Balancing in Cloud Computing Using Modified Throttled Algorithm ", 2013 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)

[12]- Sidra Aslam, Munam Ali Shah, "Load Balancing Algorithms in Cloud Computing: A Survey of Modern Techniques", 2015 National Software Engineering Conference (NSEC 2015)

[13]- Surbhi Kapoor, Dr. Chetna Dabas," Cluster Based Load Balancing in Cloud Computing", 2015 Eighth International Conference on Contemporary Computing (IC3)

[14]- Wang SC, Chen CW, Yan KQ, Wang SS The Anatomy Study of Load Balancing in Cloud Computing Environment. The Eighth International Conference on Internet and Web Applications and Services 230-235.

[15]- Whitley D A generic algorithm tutorial. Statistics and Computing 4: 65-68.

[16]- Xi, S., *RT-OpenStack: CPU Resource Management for Real-Time Cloud Computing.* 2015 IEEE 8th International Conference on Cloud Computing, 2015.

[17]- Y. Fang, F. Wang, and J. Ge, "A Task Scheduling Algorithm Based on Load Balancing in Cloud Computing", Web Information Systems and Mining, Lecture Notes in Computer Science, Vol. 6318,pp.271-277.