A Survey on Load Balancing Technique for Resource Scheduling In Cloud

Heena Kalariya, Jignesh Vania
Dept of Computer Science & Engineering,
L.J. Institute of Engineering & Technology, Ahmedabad, India

***Abstract*-"Cloud computing" is a term, which consist virtualization, distributed computing, software & web services. A cloud consists of several elements such as clients, data centers and distributed server. It provide a on-demand services, scalability, high performance, flexibility etc. cloud is based on powerful datacenter that handle large number of users, so it required load balancing for managing a client request. Using a load balancing cloud all over performance is increase. For a load balancing so many algorithms based on static & dynamic types. All are use based on requirement. Static algorithms most useful for same type request come every day. If run time resource required based on need than dynamic algorithms used. According to the literature review various author have research based on task scheduling, response time etc. In the model based method to predicate and calculate the resource requirement of each virtual machine. In this algorithms only consider CPU & memory two parameter. We research a predicted load balancing technique for resource scheduling in cloud consider CPU, memory, disk i/o, VPC & region. so we will meet a more accurate result.**

***Index Terms*- cloud computing; load forecasting; task scheduling; virtual machine.**

I. INTODUCTION

Cloud Computing is the most recent emerging paradigm promising to turn the vision of "computing utilities" into reality, Cloud computing facilitates flexible and efficient resource management via virtualization at anytime and from anywhere, so that users can get the demanded IT resources. Basically, there are three types of services in cloud computing: Iaas, paas,saas. As the number of users on cloud increases, the existing resources decreases automatically which leads to the problem of delay between the users and the cloud service providers. Thus, the load balancing needed. where infrastructure or theactual hardware is provisioned to customers who are responsible to install operating systems and necessary softwares. IaaS cloud is usually provided to users in the form of Virtual Machines (VMs), such as Amazon EC2 and VMware vCloud. In an IaaS cloud, users can apply VMs on-demand to deploy and run their applications also provide services to their clients. Overview if cloud computing is covered in Fig.1

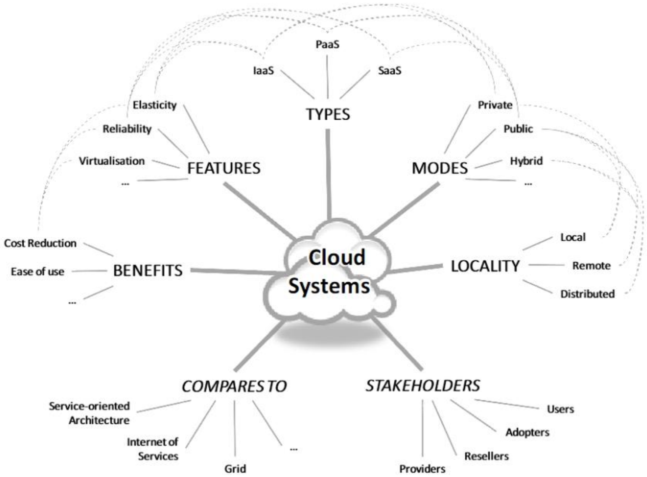

Fig. 1. View on the main aspects forming a cloud computing [4]

In IaaS, the physical resources can be split into a number of logical units called Virtual Machine (VM). User make request for resources by making an image of configuration requirement. These images get mapped on VM which is present on the server at provider side. Load balancing (LB) is done on consumer as well as on provider side. On provider side, load balancing is the problem of allocating VMs to servers at runtime. VM need to be reassigned so that servers do not get overload as demand changes. Just like VM load is distributed across servers, application load of client can be balanced across VMs on Consumer side. This paper addressed the load balancing issue on consumer side. All VM load balancing methods are designed to determine which VM is selected for the execution of next task units. So Resource management plays vital role in the performance of the entire cloud system and the level of user satisfaction provided by the cloud system. To provide the all most user satisfaction, the datacenter ought to make proper resource management so as to insist the system with minimum SLA (Service Level Agreements) violation

Load balancing is one of the key terms that affect the system performance dependent on the amount of work allotted to the system for a specific time period. In general, it can be described as anything that distributing communication and computation evenly among resources, or a system that divides many client requests into several servers. So there is need to manage the work and resources accordingly. The discussion on such various objectives and some LB algorithm which can be used to achieve those objectives are studied in this paper. There are several LB algorithms for the improvement and optimization of cloud performance parameters such as,

1) Throughput- The total number of tasks that have completed execution is called throughput. A high throughput is needed for better performance of the system.

2) Associated Overhead- The amount of overhead that is produced by the execution of the LB algorithm. Minimum overhead is imagining for successful implementation of the algorithm.

3) Fault tolerant- It is the ability to perform correctly and uniformly even in conditions of failure at any arbitrary node in the system.

4) Migration time- The time taken in migration or transfer of a task from one machine to any other machine in the system. This time should be minimum for improving the performance of the system.

5) Response time- It is the minimum time that a distributed system executing a specific load balancing algorithm takes to respond.

6) Resource Utilization- It is the degree to which the resources of the system are utilized. A good load balancing algorithm provides maximum resource utilization.

7) Scalability- It determines the ability of the system to accomplish load balancing algorithm with a restricted number of processors or machines.

8) Performance- It represents the effectiveness of the system after performing load balancing. If all the above parameters are satisfied optimally then it will highly improve the performance of the system.

Load forecasting technology is of great importance in many parts of workload balancing and resource scheduling to guarantee the service quality. Based on the load forecasting techniques, the load-balancing approach reallocates resource as load increasing, releases resources as scheduling VMs as load is imbalance and workload decreasing. Therefore, load forecasting technology could more effectively improves the service quality and resource Utilization in IaaS cloud.

II. RELATED WORK

For resource management in cloud data center, previous work has focused on the problem of replacing and placing virtual machines in servers, in order to optimize resource management for different measure, including power, performance and cost. There are some products and research projects on virtualized resource management, such as VMware DRS, Open QRM. They dynamically allocate the CPU, I/O and memory resources to partitions virtual machines according to customer’s requirement. Propose a resource scheduling scheme to enhance resource utilization and guarantee Qos of service. They Quality of service guarantee under the proposition of improving resource utilization cloud environment. Some work use the similar method used in the software minimized the number of physical machines, to manage the virtual machines using dynamic modification technique based on off line analysis of application performance as a function of machine utilization. The load forecasting techniques are widely used in the management of cloud resources such as resource scheduling and load balancing. In a model-predictive controller is proposed to minimize the total power consumption of the servers in an enclosure subject to a given set of Quality of service constraints. Neural network prediction estimates the dynamic incoming load, and serves as input to a green scheduling algorithm for turning servers on and off. A green scheduling algorithm combine with neural network has also been proposed for optimizing resources in cloud. These decisions work to minimize the number of currently running servers. We should choose a suitable load forecasting method for the VMs in IaaS cloud environment. The virtualization technology is the basis of the IaaS cloud. The existence of virual machine will impact the resource allocation because of the overhead for CPU, I/O and memory virtualization. Some research found the complex relationship between physical machine and virtual machine, but they mainly focus on the migration of the virtual machine such as the work in. About virtual machine performance presented, modeling the VM performance modeling challenges and highlighted that visible and invisible resource interference cause a significant performance degradation and need to be considered when modeling VM performance. In their models, they considered not only CPU, I/O and Memory, but also shared cache, virtual machine monitor overhead and memory bandwidth. In this paper, a model for describing the relationship between workload of VM and the resource of physical host will be created.

Cloud Computing is a kind of distributed computing where massively scalable IT-related capabilities are provided to multiple external customers "as a service" using internet technologies. The cloud providers have to achieve a large, virtualization computing infrastructure; and general-purpose of infrastructure for different customers and services to provide the multiple application services. Furthermore, the ZEUS Company develops software that can let the cloud provider cost-effectively and easily offer every customer a dedicated application delivery solution. The ZXTM software is much more than a shared load balancing service and it offers a low-cost starting point in hardware development, with a smooth and cost-effective upgrade path to scale as your service grows.
The ZEUS provided network framework can be utilized to develop new cloud computing methods, and is utilized in the current work. In this network composition that can support the network topology of cloud computing used in our study.

According to the ZEUS in consequence of the properties and network framework of cloud computing structure, a three-level hierarchical topology is adopted to our investigate framework. According to the whole information of each node in a cloud computing environment, the performance of the system will be enhanced and managed. There are several methods can collect the relevant information of node that includes, the centralized polling, broadcasting and agent.

Agent is one of the technologies used extensively in recent years. It has inherent navigational autonomy and can ask to be sent to some other nodes. In other words, agent should not have to be installed on every node the agent visits, it could collect related information of each node participating in cloud computing environment, such as CPU utilization, remaining memory capability, remaining CPU, transmission rate, etc. Therefore, when agent is dispatched, it does not need any control or connection, and travel flow can be reducing in maintaining the system. However, in this study, the agent is used to gather the related information, and reduce the cost and resources wasting.

There are distinct characteristics of each scheduling algorithm. Opportunistic Load Balancing is to attempt each node keep busy, therefore does not consider the present workload of each computer. OLB assigns each task in free order to present node of useful .The advantage is quite reach and simple load balance but its shortcoming is not consider each expectation execution time of task, therefore the whole completion time (Make span) is very poor. OLB dispatches unexecuted tasks to currently available nodes at random order, regardless of the node's current workload.

Minimum Execution Time[3] assigns each job in arbitrary order to the nodes on which it is expected to be executed fastest, regardless of the current load on that node. MET tries to find good job-node pairings, but because it does not consider the current load on a node it will often cause load imbalance between the nodes and not adapt application in the heterogeneity computer system.

Minimum Completion Time (MCT)[3] assigns each job in arbitrary order to the nodes with the minimum expected completion time for the job. The completion time is simply the ETC, but this is a much more successful heuristic as both node loads and execution times are considered. Min-Min scheduling algorithm establishes the minimum completion time for every unscheduled job, and then assigns the job with the minimum completion time to the node that offers it this time. Min-min uses the same mechanism as MCT. However, because it considers the minimum completion time for all jobs at each round, it can schedule the job that will increase the overall make span the least. Therefore, it helps to balance the nodes better than MCT.

Because of OLB scheduling algorithm is very easy and simply to implement and each computer often keep busy. In this paper research, the OLB scheduling algorithm is used to assigns the job and divides the task into subtask in a three level cloud computing network. In addition, in order to provide the working load balance of each computer in the system, the Min-Min scheduling algorithm will be enhanced in this investigates on which it is expected to be efficiently reducing execution time of each node.

The motivation of the survey of existing load balancing techniques in cloud computing is to encourage the amateur researcher to contribute in developing more efficient load balancing algorithms. This will benefit interested researchers to carry out further work in this research area. The existing load balancing algorithms prevalent in cloud environment are,
A. Vector[4] Dot Environment used: Datacenters with integrated server and storage virtualization. It uses dot product to distinguish node based on the item requirement. It handles hierarchical and multidimensional resource constraints and removes overloads on server, switch and storage.

B. Carton[4] Environment used: Unifying framework for cloud control. It uses Load balancing to minimize the associated cost and uses Distributed Rate Limiting for fair allocation of resources. It is simple and Easy to implement and very low computation and communication overhead.

C. Compare and Balance[4] Environment used: Intra-Cloud. It is based on sampling process and uses adaptive live migration of virtual machines. It balances load amongst servers and reaches equilibrium fast. It assures migration of VMs from high-cost physical hosts to low cost host. It assumes that each physical host have enough memory.

D. Scheduling strategy on LB of VM resources[4] Environment used: Cloud Computing. It uses Genetic algorithm, historical data and current state of system to achieve best load balancing and to reduce dynamic migration.

E. LBVS[4] Environment used: Cloud Storage. It Uses Fair-Share Replication strategy to achieve Replica Load balancing module which in turn controls the access load balancing and uses writing balancing algorithm to control data writing load balancing.

F. Honeybee Foraging Behavior[4] Environment used: Large scale Cloud Systems. It achieves global load balancing through local server action.

G. Biased Random Sampling[4] Environment used: Large scale Cloud systems. It achieves load balancing across all system nodes using random sampling of the system domain.

H. Active Clustering[4] Environment used: Large scale Cloud systems. It optimizes job assignment by connecting similar services by local re-wiring.

I. ACCLB[4] Environment used: Open Cloud Computing Federation. It uses small-world and scale-free characteristics of complex network to achieve better load balancing.

J. OLB + LBMM[4] Environment used: Three-level Cloud Computing Network. It uses OLB (Opportunistic Load Balancing) to keep each node busy and uses LBMM (Load Balance Min-Min) to achieve the minimum execution time of each task.

K. Server-based LB[4] Environment used: Distributed web servers. It uses a protocol to limit redirection rates to avoid remote servers overloading and uses a middleware to support this protocol. It uses a heuristic to tolerate abrupt load changes.

L. Join-Idle-Queue[4] Environment used: Cloud data centers. It first assigns idle processors to dispatchers for the availability of the idle processors at each dispatcher and then assigns jobs to processors to reduce average queue length of jobs at each processor..

III. PROBLEM STATMENT

"Model based algorithm consist only two parameter CPU & memory. In further work we shall used more parameter Region, Disk i/o, VPC (virtual private communication)."

IV. PROPOSED WORK

Algorithm: A predicted load balancing technique

Step 1: Check threshold value of VM's.
Step 2: If all one a level then stop request on that VM and direct request to other.
Step 3: Parallel check response time of every VM's on a particular time.
Step 4: Based on the response values replace dynamic queue from low to high response value.
Step 5: Update queue every given time.
Step 6: Allocate next request as per the queue value.

V. CONCLUSION AND FUTURE WORK

According to the literature review various author have research based on task scheduling, response time etc. In the model based method to predicate and calculate the resource requirement of each virtual machine. In this algorithms only consider CPU & memory two parameter. We research a predicted load balancing technique for resource scheduling in cloud consider CPU, memory, disk i/o, VPC & region. so we will meet a more accurate result. In future, we can prediction based work.

REFERENCES

[1] Zhenzhong Zhang1, Limin Xiao1*,Yuan Tao1 & Ji Tian2,Shouxin Wang2,Hua Liu2, "A Model Based Load-Balancing Method in IaaS Cloud", 0190-3918/13 ,2013 IEEE

[2] Shridhar G.Damanal and G. Ram Mahana Red dy, "Optimal Load Balancing in Cloud computing By Efficient Utilization of Virtual Machines", 978-1-4799-3635-9/14,2014 IEEE

[3] Shu-Ching Wang, Kuo-Qin Yan *(Corresponding author), Wen-Pin Liao and Shun-Sheng Wang, " s a Load Balancing in a Three-level Cloud Computing Network", 978-1-4244-5540-9/10/2010 IEEE

[4] Mr. M. Ajit & Ms. G. Vidya, "VM Level Load Balancing in Cloud Environment", IEEE – 31661, 4th ICCCNT 2013

[5] N. S. Raghava* and Deepti Singh, "Comparative Study on Load Balancing Techniques in Cloud Computing", OPEN JOURNAL Of MOBILE COMPUTING AND CLOUD COMPUTING Volume 1, Number 1, August 2014

[6] Klaithem Al Nuaimi, Nader Mohamed, Mariam Al Nuaimi and Jameela Al-Jaroodi, "A Survey

of Load Balancing in Cloud computing: Challenges and Algorithms", 978-0-7695- 4943-9/12, 2012 IEEE

[7] Shridhar G. Domanal and G. Ram Mohana Reddy, "Load Balancing in Cloud Computing Using Modified Throttled Algorithm" IEEE, International conference. CCEM 2013

[8] Jitendra Bhatia , Tirth Patel , Harshal Trivedi , Vishrut Majmudar , "HTV Dynamic Load Balancing Algorithm for Virtual Machine Instances in Cloud",978-0-7695-4931-6/12,2012 IEEE