International Journal of Electronics Communication and Computer Engineering
Volume 6, Issue 2, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

A Survey on Load Balancing Techniques in Cloud Computing

Shubhakankshi Goutam
M.Tech. Scholar, ITM University,
Gwalior, M.P. India
Email: shubhakankshi@gmail.com

Arun Kumar Yadav
Associate Prof., ITM University,
Gwalior, M.P., India
Email: arun26977@rediffmail.com

Priya Shrivastava
Assistant Prof., ITM University,
Gwalior, M.P., India
Email: priyashk32@gmail.com

***Abstract* – Cloud computing is a computing paradigm that delivers on demand IT services to consumers. Cloud computing provides development of large scale, on–demand, flexible computing infrastructure. Confidentiality, availability, privacy and performance are major concerning areas of cloud computing. As an increase demand of resources is based on performance factor, Load balancing is an important technique for virtualization of resources over the internet. Load balancing techniques can be use to schedule resources to maximize the utilization in cloud computing environment. Many researchers are currently working over load balancing techniques to improve QoS in cloud computing Environments. This paper presents a survey of existing load balancing techniques; identify the limitations and various domains for future research.**

***Keywords* – Cloud Computing, Throughput, Reliability, Load Balancing Techniques, P2P System, Virtualization.**

I. INTRODUCTION

Load balancing is an approach to reassign the loads from overloaded nodes to underutilize nodes. It is generally dynamic in nature because of traffic flow and need of server, node is depending over the user request. When a node is getting over loaded through user requests then we need a load balancing technique to reassign the loads. The main points to be considered are estimation of load, comparison of load, stability of different system, performance of system, interaction between the nodes, nature of work to be transferred, selection of nodes. This load considered can be in terms of CPU load, amount of memory used, delay or Network load. There are two types of algorithms used in load balancing techniques 1) static load balancing, 2) dynamic load balancing algorithm. In static load balancing algorithm, fixed no. of steps and prior knowledge is used for load balancing and it cannot depend over the current status of the network whereas in dynamic algorithm load changes according to the current system status of the network. Generally, dynamic algorithm works better then static algorithm.

Numbers of load balancing techniques are available and can be compared or characterized on following parameters:

- Maximum throughput- Maximize number of finished user requests in a define time unit.
- Completion time- The Maximum time unit required to complete a job.
- Communication cost- The overall cost of transmissions and receiving of the data bits.
- Min network delay- Total delay that occurred by the intermediate network devices.
- Performance factor- Describes the overall performance of algorithm to complete the user requests in a predefine time limit.
- Resource utilization- utilize a resource in such a way that it never get free.
- Execution time- Maximum time required to execute a user request
- Scalability- future scope to extend the network resource.

Using this following parameter, we are discussing the various load balancing techniques over architecture to show the process of load balancing among the various resource nodes or cloud nodes in section IV.

II. OVERVIEW OF CLOUD COMPUTING

Cloud computing is a model that provides online, on demand computing resources (storage, data, network, servers, software, applications) from an available resource pool to completely maximize the effectiveness of services that is required by the customers . Cloud computing provide data storage and different service models according to need of users. Cloud computing is a cost effective way to use various services on pay per use basis. Cloud computing based on virtualization, management and availability concepts where multiple users can able to access the single server without purchasing any license for different applications. So we can conclude that cloud computing is SOA (service oriented architecture).

Cloud computing model includes:

Reliability: Data storage should be reliable it will give results to user in one click so we have to maintain data recovery all the time by using backup copies or redundant data, we can say that reliability is related to disaster recovery.

Security: Security is related to data security of customer from unauthorized access over a public link.

Elasticity: Elasticity is related to scalability, a cloud network should be scalable to complete user demand and allow more customers to connect to the cloud services.

Cost: Cloud services model is a cost effective mechanism to use the services and applications provided by cloud providers for an organization or customers.

A. Deployment models of cloud computing:

In cloud computing there are three deployment models, according to need of organization or customers they can choose suitable model.

a) Private cloud:

A private cloud is dedicated to a single organization or a single enterprise means cloud services are dedicated to

specific list of users. Customers can connect to cloud services using the dedicated links. Using the private cloud deployment model good security can be achieved and there are maximum benefits in bandwidth utilization but cloud providers have to manage access controls.

b) Public cloud:

A public cloud services is shared or available for any public customer want to use them as pay per use basis. It is captured by third party CSP (cloud service provider) such as Amazon. Advantages associated with this deployment model are that it is cost effective and high availability of resources. Here services can be access using the public network links.

c) Community cloud:

A community cloud is collection of organizations, industries or customers that have a common set of requirements. Using this type of clouds agencies can save cost. This cloud enables organizations to securely combine and share resources and data servers.

d) Hybrid cloud:

A hybrid cloud is a combination of two or more clouds (public, private or community). This type of deployment model is incorporative in nature. Hybrid model needs standard planning and technology to maintain and needs complex security levels by controlling parameters over the data and application.

B. Service models of cloud computing:

In cloud computing there are three service models. SaaS, PaaS, and IaaS.

a) IaaS (Infrastructure as a service):

IaaS services provide computing resources such as server, storage, firewall, memory, virtual space and load balancers on the demand of customers over the public network. User can purchase these services online. IaaS can be used by IT enterprises to create cost effective IT solution or build their own IT platform by using service license by a cloud provider. These services allow the use of costly hardwires on the rent of user demand. Customers can not have rights to change the settings of underlying cloud infrastructure but they can manage networking components, servers and operating system.

Amazon web services (AWS), Microsoft Azure, Google compute engine (GCE) and Joynet are IaaS service providers.

b) PaaS (Platform as a service):

PaaS service provides programming execution and operating system applications development platform that execute the customers programs without any extra charges or installation overhead of corresponding language software. There is no need to manage hardware complexity to customers. For users it is self service portal for developing software in any language. Here user has no control over underlying cloud infrastructure but they can manage the language libraries and deployed applications.

Apprenda, Microsoft azure, Google app engine are some PaaS service providers.

c) SaaS (Software as a service):

In SaaS services cloud providers install and operate the application programs and provide an accessibility to cloud user to use that software on pay per use basis. In this model of services it gives an infrastructure, platform and collection of application programs. SaaS model reduce the IT operation costs by providing high price application software in low cost. SaaS application software runs directly over web browser so there is no need to any downloads and installation.

Google apps, Cisco, WebEx, Workday, SQL azure are some SaaS application providers.

C. Entities used in a cloud computing model:

a) Cloud provider: cloud provider provides services to user on the basis of pay per use. Cloud providers have their own service models user can choose any service model according to their need. Services offer by cloud providers are available on cheap cost over the network connection.

b) Cloud user: cloud users are client that can buy services on the rent. Cloud clients can be a single client, group of clients or it can be a large scale organization.

c) Cloud Manager: cloud manager provides management tool for a cloud. Overall working and assignments of resources are controlled by cloud manager. Cloud manager have all the control over the cloud resources and scheduling mechanism.

d) Cloud Broker: cloud broker provides a bridge between multiple cloud service providers to maximize the resource availability.

e) Cloud operating system: it's a cloud layer which acts as a middleware between client and cloud resources. It provides a user friendly environment.

III. USE OF VIRTUALIZATION IN CLOUD COMPUTING

Virtualization provides illusion to user where a single hardware can serve multiple user requests. Virtualization is an abstraction layer between the client and cloud resources. Using the concept of virtualization we can create a virtual world where a broad range of resource is available. Virtualization is software that separates physical infrastructures to create various dedicated resources. The main component of virtualization is virtual machine. Virtualization improves availability, performance and scalability factors in cloud computing.

Virtual machine: VM provides abstraction layer to share the computing resources for network level user requests. Isolation plays major role in VM's concept, where each machine has own environment to run process separately. Multiple processes running on single host is separated by instances of machine. Each process has their own virtual machine and it does not affect another process running on same host.

Hypervisor: Hypervisor is also called as software layer because it manages and monitors the virtualization of resources to complete the user requirements. It is an interactive layer that works between operating system and hardware resources. Hypervisor act as a host machine which manages multiple users called as guest machines. Native and host operating systems are two types of hypervisor. In native hypervisor there is no need of

International Journal of Electronics Communication and Computer Engineering
Volume 6, Issue 2, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

software abstraction because it can directly run over the hardware. A host operating system has software rules that required performing virtualization of resources in a proficient way.

Emulation: Emulation is a type of virtualization technique where the behavior of hardware can be translated into a software package. By using technique of emulation better flexibility can be achieved in a cloud model.

IV. Use of Load Balancing Techniques in Cloud Computing Environment

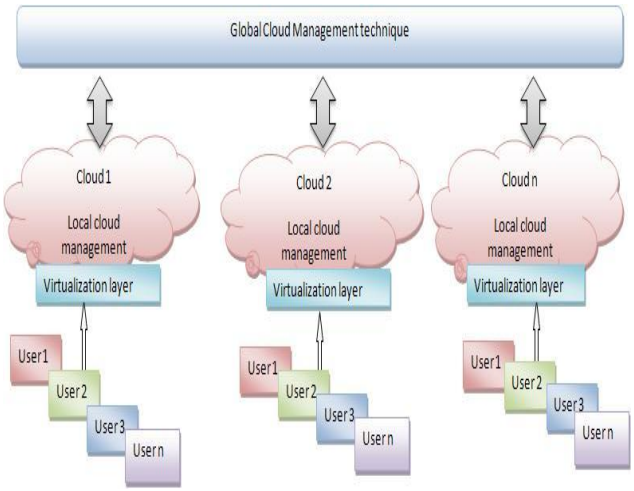

Fig.1. Techniques used for load management in cloud environment

The above diagram shows the basic idea of using the techniques of load balancing in cloud computing environment.

A. Cloud management schemes in load balancing techniques

Local load balancing: Local load balancing management scheme is used for a service cloud. We can say that there is intra domain scheduling of available resources between many user requests on a single cloud.

Global Load balancing: Global load balancing techniques can be used for scheduling of resources between more than one cloud. It is based on global traffic prediction and current status of the system where many clouds can combine in a single unit for load balancing.

B. Type of Load balancer in load balancing techniques:

Centralized Load Balancing: In centralized load balancing there is only one central decision maker for reassigning loads. This technique reduces the time required to analyze different cloud resources but creates a great overhead if central decision maker gets failed then whole cloud becomes failed.

Distributed Load Balancing: In distributed load balancing technique there are many decision maker nodes that are responsible for resource provisioning, or task scheduling decision. Multiple nodes monitor the network to make load balancing decision.

C. Load balancing on the basis of cloud environment:

a) Receiver initiated: In this receiver allocate resources by itself to a client which wants to use the cloud services.

b) Sender initiated: In this client sends a request to cloud server for assigning resources to complete the process.

c) Symmetric: In this type of load balancing it is a collection of both type of initiator. According to current status of the system, it can act like either receiver initiator or sender initiator.

d) Static environment: In static environment of cloud has a fixed design where if a task assigns to a VM then it cannot shift during the execution. Work in a static environment is easy.

e) Dynamic Environment: In dynamic environment it is more suitable for distributed cloud environment because it works on current status of the system. It can shift the process during the execution.

V. Related Works

There are many kinds of techniques have used for the load balancing approach, that are described here with merits and demerits of each techniques

A. Honey bee behavior inspired load balancing of tasks in cloud computing environment

Dhinesh babu L.D. and venkata Krishna proposed an algorithm based on HBB-LB model. In this technique of load balancing the used phenomena is inspired by the behavior of honey bees. Here scout bees forage food sources and come back to beehive and then perform a dance called as waggle dance this dance give the idea about the quality of food and distance from the beehive. Then forage bee's follows the path of scout bees to the location, when they return to beehive they also do the waggle dance to other bees for giving an idea about how much food is left. In HBB-LB technique same idea is used for load balancing, the tasks are denoted as honey bees and VM's are denoted as food sources. When a VM is find under loaded then it like foraging bee finding a new food source. Then a status has indicated like perform a waggle dance to show that how many tasks can be perform the VM and task can be chosen according to their load and priorities. This updating status will give the clear idea about which task is assign to which VM based on the availability and scalability and total load of VM.

HBB-LB proposed algorithm for minimizing the calculation of current workload, good resource utilization, high throughput, and QoS is based on the task priority.

Limitations related to this algorithm are that it may suffer from load imbalance between the VM's, minimizes the response time and underutilization of low priority task.

B. Dynamic Load balancing in distributed virtual environment using Heat diffusion

Yunhua deng and Rynson W.H. Lau proposed an algorithm that is based on the principle of heat diffusion, a simple concept is use for load balancing. In heat diffusion concept heat diffusion has happened from high

temperature to low temperature. These same phenomena is used for load balancing purpose in the terms of VM's the traffic flow of user request is from overloaded VM to under loaded VM. According to this algorithm the virtual environment is divided into number of cells and each cell have objects, every node in cell send load information to its neighbor node in single iteration. Heat diffusion concept use as local diffusion and global diffusion methods, where local diffusion decide the local topology or scheduling of virtual resources and user's request locally satisfy using the local decision making. And global diffusion is used at the global level of decision making for knowing which VM is assign to which user request and how to VM's overloading managed at global level.

In heat diffusion load balance environment amount of load migrates is minimum. This algorithm Efficient for multiprocessor network and network latency is minimizing for load transfer between the cells.

Limitations associated with this technique are very small connectivity for large scale graphs or cells, higher computational and communication methods are used, network delay on single path, and if there are more iteration then more time wastage.

C. Decentralized scale free network construction and load balancing in massive multiuser virtual environments

Markus Esch and Eric Tobias addressed a concept for self organized load balancing scheme. They provide an algorithm for the decentralized construction of a scale free link structure interconnecting network nodes. In this load balancing scheme a hyper verse infrastructure relies on two tier architecture. The real backbone of this type of technology is public servers. Public servers are loosely interconnected architecture p2p overlays that are used for torrent based data distribution. For self organized load balancing of server node, a world surface is divided into the small cells. Each cell managed public servers. The management of machines is based on crowded area for high crowded area there are many more powerful machines and for less crowded area there are less powerful machine. By the self organize behavior position of servers they are managed virtually. Scale free link structure algorithm is used for public servers to link them in a connected fashion, here we established a link together. So the new node can connected to an existing node with a predefine probability here global load is to be handle. So it's an adaptive self organizing method for load balancing using cell dividing method managed by global server.

In this technique here fast routing in scale free network. Main advantage of this method is sort average path length and resilience against random node failure. It also support for fault tolerance.

Limitations of this technique are that it depends on size of cell so sometimes it cannot avoid overloaded node, and there must be synchronization (global) between the distributed nodes.

D. Load balancing in dynamic structured P2P systems

Brighten godfrey proposed an algorithm for load balancing in this technique a mapping is perform, a unique identifier is associate with each of data items at each node. Peer 2 peer system used DTH (dynamic hash table) abstraction In this method two functions are associated : put (id, item), put function used to store an data item to associate identifier over a network connection, and Get(id) that requests for retrieve the data item. Load of any node can be dynamic because data items are storing and removing continuously and nodes get connected and disconnected continuously. Directory based management of load information is the most important part of this technique, a directory is collection of information of peer nodes where numbers of directories are periodically maintain and reassignment of VM is capture in the directory. Here Transfer of a virtual server to node when a node get overloaded.

Advantages of this technique are good system utilization, scalability and bandwidth improvement.

Limitations associated with this method is if there are m virtual servers per node, then routing status increases by factor of m. and reassignment of virtual server to a particular node is difficult.

E. A Fast adaptive load balancing method for parallel practical based simulations

Dongliang zhang and changjun jiang addressed to a scheme to improve the performance of distributed simulation systems. It is a universal method for adaptive load balancing technique, adaptive means system changes its status according to user requests and load is adapted according to the distance between the sub domains. Domain decomposition is based on discrete approximation method, finite element method and boundary element method. The partition of the region is done using binary tree method, each leaf represents a cell and an area is shown by the parent node. In this method the spatial region decomposes into sub domains. Each sub domain is associated to a solo processor. We group the domain hierarchy. A hybrid decomposition method is used to reduce inter communication costs. A heavy node forwards its traffic from one domain to neighbor node. In this technique load of a node is distributed according to local and global traffic notifications.

In this technique there is lower communication overhead and a faster convergence speed is gain in real distributed case. Fast a limitation associated with this method is that it cannot maintain topologies of cells.

F. A dynamic and adoptive load balancing strategy for parallel file system with large-scale I/O servers

Bin dong and Limin Xiao proposed a technique where it is based on distributed architecture for dynamic and adaptive load balancing. They proposed an algorithm called as a SALB (self active load balancing algorithm).In parallel file system data is collectively transferred between memory and I/O systems in a single transmission. Using this technique data is effectively transferred over a cloud structure and load balancing is needed for dynamic file migration. SALB forecast future load of server and makes its own decision by on line load prediction model. SALB works as background process without interrupting the system. For reduce the message exchange a threshold criteria is attached, this threshold value gets dynamically

adjustment according to the traffic. Main backbone of this algorithm is a decision maker for distribution of load balancing. In central decision maker, responsibility of taking decisions are assign to a central node but there are some problems such as if central decision making node gets failed then all the system gets failed and limited network bandwidth using the single central node. In group decision maker based algorithm a large system is divided into groups, each group has its own decision maker but there is only local optimization of load into each group. It does not give the global view of system load, so a distributed decision maker is used to provide scalability, availability and minimize decision delay.

Advantages of this technique are high speed processing, elasticity, based on global load management, resource utilization, works on large file system, and load migration without affecting to system processing.

A limitation associated with this algorithm is that there are some file migration side effects that degrade the system performance.

FUTURE WORK

In HBB-LB technology it can be used techniques of artificial intelligent and neural model for better performance and previous algorithm can be improved using various QoS parameters such as resource allocation method for low priority tasks in a better way.

Currently researchers are still working on coverage speed and coverage time factors in dynamic load balancing in distributed virtual environment using heat diffusion technique.

In decentralized scale free network construction and load balancing in massive multiuser virtual environments of load balancing we can define a set of rules for servers to self organize into massive environment and can work over to find an aggregation scheme of node. Find a policy how to reassign hotspots.

In dynamic and adoptive load balancing strategy for parallel file system with large-scale I/O server's technique the relationship between load balancing and data replication can be explored to improve the effectiveness of file migration.

In a Fast adoptive load balancing method for parallel practical based simulations this algorithm generally focus on convergence speed and communication overhead we can test the algorithm on the different factors such as utilization, throughput, and load management.

CONCLUSION

In this paper we are discussing the recent technologies being used and also merits and demerits of each technology. Researchers are still working to improve the performance factors and also working to come over the demerits of the various techniques. Distributed load balancing is also a broad and highlighted area for researchers to go ahead with new algorithms and improvement of existing one. Researchers can also purpose various scheduling mechanism for reassignment of resource.

REFERENCES

[1] Dhinesh Babu L.D, P. Venkata Krishna, "Honey bee behavior inspired load balancing of tasks in cloud computing environments", Applied Soft Computing 13 (2013) 2292–2303.

[2] Yunhua Deng, Rynson W.H. Lau, "Heat diffusion based dynamic load balancing for distributed virtual environments", in: Proceedings of the17th ACM Symposium on Virtual Reality Software and Technology, ACM, 2010, pp. 203–210.

[3] Markus Esch, Eric Tobias, "Decentralized scale-free network construction and load balancing in Massive Multiuser Virtual Environments", in: Collaborative Computing: Networking, Applications and Work sharing, Collaborate Com, 2010, 6th International Conference on, IEEE, 2010, pp. 1–10.

[4] B. Godfrey, K. Lakshminarayanan, S. Surana, R. Karp, I. Stoica, "Load balancing in dynamic structured P2P systems", in: INFOCOM 2004. Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 4, IEEE, 2004, pp. 2253–2262.

[5] Dongliang Zhang, Changjun Jiang,Shu Li, "A fast adaptive load balancing method for parallel particle-based simulations", Simulation Modeling Practice and Theory 17 (2009) 1032–1042.

[6] Bin Dong, Xiuqiao Li, Qimeng Wu, Limin Xiao, Li Ruan, "A dynamic and adaptive load balancing strategy for parallel file system with large-scale I/O servers", J. Parallel Distribution Computing. 72 (2012) 1254–1268.