International Journal of Technology Research and Management
ISSN (Online): 2348-9006
Vol 2 Issue 5 May 2015

An Introduction to the Modern Cloud Load Balancing Algorithms

Diptanshu Pandya[1], Dr. M.K. Rawat[2] and Jitendra Dangra[3]
LNCT Indore CSE Department, RGPV Bhopal, Ujjain, Madhya Pradesh, India[1]
LNCT Indore CSE Department, RGPV Bhopal, Ujjain, Madhya Pradesh, India[2]
LNCT Indore CSE Department, RGPV Bhopal, Indore, Madhya Pradesh, India[3]
diptanshupandya@gmail.com[1], drmkrawat@gmail.com[2], jitendra.dangra@gmail.com[3]

Abstract: *In modern days cloud computing is one of the greatest platform which provides storage of data in very lower cost and available for all time over the internet. But the cloud computing has more critical issue like security, load balancing and fault tolerance ability. In this paper we are focusing on Load Balancing approach. The Load balancing is the process of distributing load over the different nodes which provides good resource utilization when nodes are overloaded with job. Load balancing is required to handle the load when one node is overloaded. When the node is overloaded at that time load is distributed over the other ideal nodes. Many load balancing algorithms are available for load balancing like Static load balancing and Dynamic load balancing.*

1. Introduction

In Cloud Computing [1] scalable resources are provisioned dynamically as a service over Internet in order to assure lots of monetary benifits to be scattered among its adopters. Different layers are outlined based on the kind of services provided by the Cloud. Moving from bottom to top, bottom layer contains basic hardware resources like Memory, Storage Servers. Hence it is denoted as Infrastructure-as-a-Service (IaaS). The distinguished examples of IaaS are Amazon easy Storage Service (S3) and Amazon Elastic Compute Cloud (EC2). The layer above IaaS is Platform-as-a-Service (PaaS) which mainly supports deployment and dynamic scaling of Python and Java based applications. One such an example of PaaS is Google App Engine. On top of PaaS, a layer that offers its customers with the capability to use their applications referred to as Software-as-a-Service (SaaS). SaaS supports accessing user's applications through a browser without the knowledge of Hardware or Software to be installed.

The basic services of Cloud have been considered as the following. Platform as a Service (PaaS): PaaS is an deployment and development platform for applications provided as a service to developers over the Web. Third party renders develop and deploy software or applications to the end users through internet and servers. The cost and complexity of development and deployment of applications can be reduced to a great extent by developers by using this service. Thus the developers can reduce the cost of buying and reduce the complexity of managing the required Infrastructure. It provides all of the services required to build and deliver the web services to support the complete life cycle of web applications entirely from the Internet. This platform consists of infrastructure software, databases, middleware, and development tools. Infrastructure as a Service (IaaS): is a delivery model associated with Hardware and Software as a service. Hardware such as Storage, server and network along with supporting software such as operating system, virtualization technology and file system. It is an 3 evolution of traditional hosting to allow users to provide resources on demand and without require any long term commitment. Different from PaaS services, the IaaS provider does very little management of data other than to keep the data center operational. Deployment and managing of the software services

must be done by the end users just as the way they would in their own data center. Software as a service (SaaS): SaaS allows access to programs to large number of users all the way through browser. For a user, this can save some cost on software and servers. For Service provider's, they only need to maintain one program, this can also save space and cost. Naturally, a SaaS provider gives access to applications to multiple clients and users over web by hosting and managing the given application in their or leased data centers. SaaS providers also runs their applications on platforms and infrastructure provided by other cloud providers.

Load Balancer: This mechanism contains algorithms for mapping virtual machines onto physical machines in a cloud computing environment, for identifying the idle virtual machines and for migrating virtual machines to other physical nodes. Whenever a user submits an application workload into cloud system, one can create a new virtual machine. Now the mapping algorithm of Load balancer will generate a virtual machine placement scheme, assign necessary resources to it and deploy the virtual machine on to the identified physical resource. Unmanaged and forgotten virtual machines can consume Data center resources and cause energy waste. Another algorithm of Load balancer will identify idle virtual machines and shut them off. In the process of optimally placing the virtual machine onto the destination, we need to relocate the existing virtual machines. For doing this operation, virtual machine migration algorithm of load balancer is invoked.

2. Cloud Load Balancing Algorithms

Round Robin Round robin performs the basic type of load balancing and functions simply by providing the list of IP address of cloudlet. It allocates first IP address to the first requester then second IP address to the second requester for a fixed interval of time known as time slice. If the request is unable to finish within the given slice time, it will have to wait for the next cycle to get it turn for execution. This will continue till submitted tasks are not completed.

Active Monitoring Load Balancer This load balancer find outs the active VM and also to event out the active task at any point of time.

Throttled Load balancer This load balancing technique ensures that only a per-defined number of Internet cloudlets are allocated to a single VM at any point of time. If more groups are presents in the data center than the number of available VM than some of the requests have to be queued until the next VM is available.

Load Balancing Algorithms:

A number of load balancing algorithms existing which are distributing the load among the data center. Each of them has their own functionality. Some of the major load balancing algorithms has been discussed as follows:

A method using Estimated Time to Compute (ETC) matrices was presented to model heterogeneous systems; it is found that the Minimize Completion Time (MCT) scheduling algorithm [2] attempts to minimize the total computational time, required for any job performed for the best out of a set of well-known scheduling algorithms.

The work done by A. Singh et al. [3] proposed a novel load balancing algorithm called VectorDot. This algorithm handles the hierarchical complexity of 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 virtualization technologies.

The work done by Stanojevic et al. [4] proposed a mechanism CARTON for cloud control that unifies the use of LB and DRL. The 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.

Author Y. Zhao et al. [5] addressed the problem of intra-cloud load balancing amongst physical hosts by adaptive live migration of virtual machines. The load balancing model is designed and implemented to reduce virtual machines migration time by shared storage to balance load amongst servers according to their processor or IO usage.

Work done by V. Nae et al. [6] presented an event driven load balancing algorithm for real-time Massively Multiplayer Online Games (MMOG). The

algorithm after receiving capacity events as input, also analysis its components in context of the resources and the global state of the game session, then generating the game session load balancing actions.

The J. Hu et al. [7] proposed a scheduling strategy on load balancing of VM resources that uses historical data and current state of the system. Proposed strategy achieves the best load balancing and reduced dynamic migration by using a genetic algorithm.

The A. Bhadani et al. [8] proposed a Central Load Balancing Policy for Virtual Machines (CLBVM) that balances the load evenly in a distributed virtual machine/cloud computing environment.

The LBVS H. Liu et al. [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. The Storage virtualization is achieved using an architecture that is three-layered and load balancing is achieved using two load balancing modules. It helps in improving the efficiency.

The Y. Fang et al.[10] discussed a two-level task scheduling mechanism based on load balancing to meet dynamic requirements of users and obtain high resource utilization. Algorithm achieves load balancing by first mapping tasks to virtual machines and then virtual machines to host resources thereby improving the task response time, and resource utilization also overall performance of the cloud computing environment.

Author M. Randles et al. [11] investigated a decentralized honey bee based load balancing technique that is a nature inspired algorithm for self-organization. Algorithm achieves global load balancing through local server actions. Performance of the system is enhanced with increased system diversity but throughput is not increased with an increase in system size. This is best suited for the conditions where the diverse population of service types is required.

The work done by M. Randles et al. [11] investigated a distributed and scalable load balancing approach that uses random sampling of the system domain to achieve self-organization thus balancing the load across all nodes of the system.

Author M. Randles et al. [11] investigated a self-aggregation load balancing technique that is a self-aggregation algorithm to optimize job assignments by connecting similar services using local re-wiring. Overall performance of the system is enhanced with high resources thereby in-creasing the throughput by using these resources effectively.

The Z. Zhang et al. [12] proposed a load balancing mechanism based on ant colony and complex network theory (ACCLB) in an open cloud computing federation. Proposed algorithm uses small-world and scale-free characteristics of a complex network to achieve better load balancing. Proposed technique overcomes heterogeneity is adaptive to dynamic environments and has good scalability hence helps in improving the performance of the system.

Author S.-C. Wang et al. [13] proposed a two-phase scheduling algorithm that combines OLB (Opportunistic Load Balancing) and LBMM (Load Balance Min-Min) scheduling algorithms to utilize better executing efficiency and maintain the load balancing of the system. This OLB scheduling algorithm keeps every node in working state to achieve the goal of load balance and LBMM scheduling algorithm is utilized to minimize the execution time of each task on the node thereby minimizing the overall completion time.

Author H. Mehta et al. [14] Proposed a new content aware load balancing policy named as work-load and client aware policy (WCAP). Proposed work uses a parameter named as USP to specify the unique and special property of the requests as well as computing nodes. The USP helps the scheduler to decide the best suitable node for processing the requests.

Author Y. Lua et al. [15] proposed a Join-Idle-Queue load balancing algorithm for dynamically scalable web services. Work provides large-scale load balancing with distributed dispatchers by, first load balancing idle processors across dispatchers for the availability of idle processors at each dispatcher and then, assigning jobs to processors to reduce average queue length at each processor

3. Conclusion

This paper has proposed a survey of load balancing methods. In cloud computing load balancing is one of the main issue. When client is requesting for service it should be available to the client. When any node is overloaded with job at that time load balancer has to set that load on another free node. Therefore load balancing is necessary in cloud computing. This paper has discussed all modern existing methods for cloud computing.

References

[1]. John Harauz, Lorti M. Kaufinan. Bruce Potter, "Data Security in the World of Cloud Computing", IEEE Security & Privacy, Co published by the IEEE Computer and Reliability Societies, July/August 2009.
[2]. National Institute of Standards and Technology- Computer Security Resource Center -www.csrc.nist.gov
[3]. Singh A., Korupolu M. and Mohapatra D., ACM/IEEE conference on Supercomputing, 2008.
[4]. Stanojevic R. and Shorten R., IEEE ICC, 1-6, 2009.
[5]. Zhao Y. and Huang W., 5th International Joint Confer-ence on INC, IMS and IDC, 170-175, 2009.
[6]. Nae V., Prodan R. and Fahringer T., 11th IEEE/ACM International Conference on Grid Computing (Grid), 9-17, 2010.
[7]. Hu J., Gu J., Sun G. and Zhao T., 3rd International Symposium on Parallel Architectures, Algorithms and Programming, 89-96, 2010.
[8]. Bhadani A. and Chaudhary S., 3rd Annual ACM Bangalore Conference, 2010.
[9]. Liu H., Liu S., Meng X., Yang C. and Zhang Y.,International Conference on Service Sciences (ICSS), 257-262,2010.
[10]. Fang Y., Wang F. and Ge J., Lecture Notes in Computer Science, 6318, 271-277, 2010.
[11]. Randles M., Lamb D. and Taleb-Bendiab A., 24th International Conference on Advanced Information Networking and Applications Workshops, 551-556, 2010.
[12]. Zhang Z. and Zhang X, 2nd International Conference on Industrial Mechatronics and Automation, 240-243, 2011.
[13]. Wang S., Yan K., Liao W. and Wang S, 3rd International Conference on Computer Science and Information Technology, 108-113, 2010.
[14]. Mehta H., Kanungo P. and Chandwani M., International Conference Workshop on Emerging Trends in Technology, 370-375, 2011.
[15]. Lua Y., Xiea Q., Kliotb G., Gellerb A., Larusb J. R. and Green-ber A,"Int. Journal on Performance evaluation", 2011.