International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

A Review on Various Approaches of Load Balancing In Cloud Computing

Charanjeet Singh[1], Amandeep Kaur[2]

[1]Student, Desh Bhagat University

[2]Assistant Professor , Desh Bhagat University

Abstract: ***Cloud computing refers to the delivery of computing resources over the Internet. Instead of keeping data on your own hard drive or updating applications for your needs, you use a service over the Internet, at another location, to store your information or use its applications. In this paper briefly defined the cloud computing and various approaches used to cloud computing.***

Keywords: load balancing, cloud computing. Directed Acyclic Graph, Round Robin and Shortest Job First

1. Introduction

1.1 Cloud Computing

Cloud computing is the long dreamed vision of computing as a utility, where data owners can remotely store their data in the cloud to enjoy on-demand high-quality applications and services from a shared pool of configurable computing resources. Cloud is a new business model wrapped around new technologies such as server virtualization that take advantage of economies of scale and multi-tenancy to reduce the cost of using information technology resources. It also brings new and challenging security threats to the outsourced data. Since cloud service providers (CSP) are separate administrative entities, data outsourcing actually relinquishes the owner's ultimate control over the fate of their data.

Frameworks provide mechanisms for:

- self-healing
- self monitoring
- resource registration and discovery
- service level agreement definition

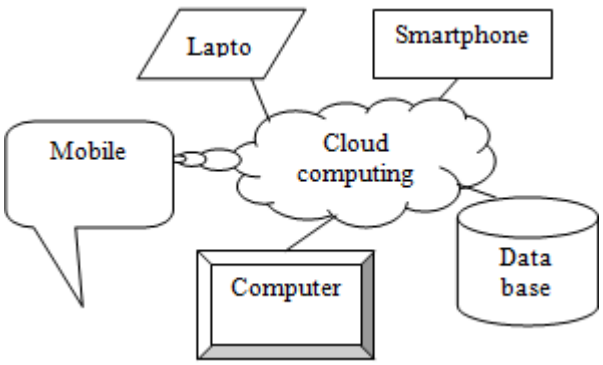

Figure 1.1: Cloud Computing

1.2 Benefits of Cloud Computing

The following are some of the possible benefits for those who offer cloud computing-based services and applications:

1.2.1Cost Savings — Companies can reduce their capital expenditures and use operational expenditures for increasing their computing capabilities. This is a lower barrier to entry and also requires fewer in-house IT resources to provide system support.

1.2.2Scalability/Flexibility — Companies can start with a small deployment and grow to a large deployment fairly rapidly, and then scale back if necessary. Also, the flexibility of cloud computing allows companies to use extra resources at peak times, enabling them to satisfy consumer demands.

1.2.3Reliability — Services using multiple redundant sites can support business continuity and disaster recovery.

1.2.4Maintenance — Cloud service providers do the system maintenance, and access is through APIs that do not require application installations onto PCs, thus further reducing maintenance requirements.

1.2.5Mobile Accessible — Mobile workers have increased productivity due to systems accessible in an infrastructure available from anywhere.

1.3 Service models of Cloud Computing:

1.3.1 Software-as-a-Service

This was the earliest cloud service and the first to enjoy widespread adoption. In a nutshell, SaaS is the online delivery of software functionality and capability without the need for locally running software. Rather, SaaS runs on a Web browser.

1.3.2 Platform-as-a-Service

Broadly speaking a Platform-as-a-Service (PaaS) is a cloud-based application development environment. Using a PaaS, companies can produce new applications more quickly and with a greater degree of flexibility than with older development platforms tied directly to hardware resources. Running application development on a PaaS has a number of key benefits. Programmers and development managers especially appreciate that the cloud provider handles all the care and maintenance of the underlying operating system(s), servers, storage, and application containers. PaaS environments can be extremely useful when development teams are widespread geographically or when partner companies or divisions share development efforts.

1.3.3Infrastructure as a Service (IaaS)
Infrastructure Providers manage a large set of computing resources, such as storing and processing capacity. Through Virtualization, they are able to split, assign and dynamically re-size these resources to build ad-hoc systems as demanded by customers. They deploy the software stacks that run their services.

1.4 Various Cloud Models:

1.4.1 Private cloud
The cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises.

1.4.2 Community cloud
The cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises.

1.4.3 Public cloud
The cloud infrastructure is provisioned for open use by the general public. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the cloud provider.

1.4.4 Hybrid cloud
The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load balancing between clouds).

1.5 Load Balancing

Load balancing is a computer networking method to distribute workload across multiple computers or a computer cluster, network links, central processing units, disk drives, or other resources, to achieve optimal resource utilization, maximize throughput, minimize response time, and avoid overload. Using multiple components with load balancing, instead of a single component, may increase reliability through redundancy. The load balancing service is usually provided by dedicated software or hardware, such as a multilayer switch or a Domain Name System server.

1.5.1 Load balancing in Cloud computing
Load Balancing is a method to distribute workload across one or more servers, network interfaces, hard drives, or other computing resources. Typical data centre implementations rely on large, powerful (and expensive) computing hardware and network infrastructure, which are subject to the usual risks associated with any physical device, including hardware failure, power and/or network interruptions, and resource limitations in times of high demand. Load balancing in cloud differs from classical thinking on load-balancing architecture and implementation by using commodity servers to perform the load balancing. This provides for new opportunities and economies-of-scale, as well as presenting its own unique set of challenges. Load balancing is used to make sure that none of existing resources are idle while others are being utilized. To balance load distribution, migrate the load from the source nodes (which have surplus workload) to the comparatively lightly loaded destination nodes. When apply load balancing during runtime, it is called dynamic load balancing — this can be realized both in a direct or iterative manner according to the execution node selection:

- In the iterative methods, the final destination node is determined through several iteration steps
- In the direct methods, the final destination node is selected in one step.

And another kind of Load Balancing method can be used i.e. the Randomized Hydrodynamic Load Balancing method, a hybrid method that takes advantage of both direct and iterative methods.

1.5.2 Goals of Load balancing:
1) To improve the performance substantially.
2) To have a backup plan in case the system fails even partially.
3) To maintain the system stability.
4) To accommodate future modification in the system.

1.5.3 Types of Load balancing algorithms:
Depending on who initiated the process, load balancing algorithms can be of three categories as

- **Sender Initiated**: If the load balancing algorithm is initialized by the sender.
- **Receiver Initiated**: If the load balancing algorithm is initiated by the receiver.
- **Symmetric**: It is the combination of both sender initiated and receiver initiated.
- Depending on the current state of the system, load balancing algorithms can be divided into 2 categories:
- **Static:** It does not depend on the current state of the system. Prior knowledge of the system is needed.
- **Dynamic**: Decisions on load balancing are based on current state of the system. No prior knowledge is needed. So it is better than static approach. Here we will discuss on various dynamic load balancing algorithms for the clouds of different sizes.

1.6 Types of Cloud Computing

Today cloud is very common thing in peoples but not everyone means the same thing when they do. The general idea behind the cloud is – that applications or other business functions exist somewhere away from the business itself – there are much iteration that companies look to in order to actually use the technology. Cloud computing offers a variety of ways for businesses to increase their IT capacity or functionality without having to add infrastructure, personnel, and software. Seven different types of cloud computing are:

1.6.1 Web-based cloud services. These services let you exploit certain web service functionality rather than using fully developed applications.

1.6.2 SaaS (Software as a Service). This is the idea of providing a given application to multiple tenants. SaaS solutions are common in sales, HR, and ERP.

1.6.3 Platform as a Service. This is saas variant. You run your own applications but you do it on the cloud provider's infrastructure.

1.6.4 Utility cloud services. These are virtual storage and server options that organizations can access on demand even allowing the creation of a virtual data center.

1.6.5 Managed services. In this scenario, a cloud provider utilizes an application rather than end-users. So, for example, this might include anti-spam services, or even application monitoring services.

1.6.6 Service commerce. These types of cloud solutions are a mix of SaaS and managed services. They provide a hub of services through which the end-user interacts. Common implementations include expense tracking, travel ordering, or even virtual assistant services

2. Review of Literature

Amir Nahir et al [1] Author proposes a novel scheme that incurs no communication overhead between the users and the servers upon job arrivals,thus removing any scheduling overhead from the job execution'scritical path. Furthermore, our scheme is oblivious, that is, it does not use any state information.Our approach is based on creating, in addition to the regular job requests that are assigned to randomly chosen servers, also replicas that are sent to different servers; these replicas are servedin low priority, such that they do not add any real burden on the servers. Through analysis and simulations we show that the expected system performance improves up to a factor of 2 (even under high load conditions), if job lengths are exponentially distributed, and over a factor of 5, when job lengths adhere to heavy-tailed distributions. We implemented a load balancing system based on our approach and deployed it on the Amazon Elastic Compute Cloud (EC2). Realistic load tests on that system indicate that the actual performance is as predicted.

Hong Tao et al [2] Author proposed that with the growing demand of data and the increase of the user scale, data allocation technology has become a key technology for improving scalability and flexibility in current mass storage system such as cloud storage system. This paper proposed an efficient dynamic data allocation strategy with data partitioning and loadbalancing. Based on the basic idea of consistent hashing algorithm, the strategy introduced the concept of virtualization technology and improved the load-balance with employing virtual node. Moreover, the strategy adopted a novel available-storage-capacity-aware and storage-capacity-utilization-aware method to enhance the performance of the cloud storage system. The simulation results demonstrate that the proposed data allocation strategy improves system performance in both homogeneous and heterogeneous distributed storage architectures.

Yuqi Zhang et al [3] Author described that in clouds, many applications need to distribute large data sets from the cloud's storage facility to all compute nodes as fast as possible, especially data-intensive parallel applications. Many multicast algorithms have been used for clusters and grid environments. In order to maximize available bandwidth and avoid bottleneck links, a common approach is to construct one or more spanning trees based on the network monitoring data and network topology. However, in clouds the available bandwidth changes dynamically, so delivering optimal performance becomes difficult. In this paper, we focus on Eucalyptus (an open-source cloud-computing platform) and propose a high performance multicast algorithms `steal-and-p2p' based on `non-steal' and `steal' algorithm mentioned. We evaluate our algorithm on Eucalyptus, and show that the algorithm can achieve high throughput and perform much better having each node downloading all data directly from storage facility.

Er.Amandeep Kaur1 et al [4] Author proposed that as the cloud computing is a new style of computing over internet. It has many advantages along with some crucial issues to be resolved in order to improve reliability of cloud environment. These issues are related with the load management, fault tolerance and different security issues in cloud environment. In this paper the main concern is load balancing and security issues in cloud computing. The load can be CPU load, memory capacity, completion time of each job and security issues to prevent the data from unauthorized user. From decades well known algorithms like FCFS, Priority has been seen into action to reduce the server load. But with the increase in the complexity of the server needs, they have failed to cope up with the current scenario. In our approach, we are developing a technique named Cross Breed Job scheduling technique which would be a combination of FCFS, Priority and would be monitored by RBAC (Role based access control). RBAC is a system which checks that whether the user of the system has the access to particular content or not. If the user doesn't have the access to the content, he will be denied and the server's load would be minimized.

Magade, Krishnanjali A. et al [5] The IEEE 802.11 standard does not provide any mechanism to resolve load imbalance in the network. To reduce this deficiency, various loadbalancing techniques have been designed. Loadbalancing provides a cost-effective, efficient and transparent method to expand the bandwidth of network devices and servers, increase the throughput, and enhance the data process capability, thus increasing the flexibility and availability of networks. There are different techniques based on persistent algorithm for loadbalancing in Wireless LAN. The goal of the proposed work is to use Persistence weighted round robin algorithm for loadbalancing in wireless LAN. This algorithm is able to distribute mobile stations among all APs and the signal strengths between stations and access points are also being maximized at the same time. This technique will be useful to reduce the congetion in the network, maintain the load in balance condition on network, as well as improve the bandwidth utilization.

3. Approaches Used

3.1 FCFS (first come first serve)

CPU gets a lot of processes to handle. The problem is shortening the waiting time for a process to reach CPU and get processed.

- Process the requests the CPU FIRST is allocated the CPU FIRST.
- Also called FIFO.
- Non-preemptive
- Used in batch systems
- Implementation
- FIFO queues
- A new process enters the tail of the queue
- The schedule selects from the head of the queue
- Performance metric: average waiting time
- Given parameters:
- Burst time, arrival time, order

3.2 RR (Round robin)

This method is quite same as the FCFS but the difference is the in this case the processor will not process the whole job (process) at a time. Instead, it will complete an amount of job (quantum) at a turn and then will go to the next process and so on. When all job has got a turn, it will again start from the first job and work for a quantum of time/cycle on each job and proceed.

- Preemptive version of FCFS.
- Treat ready queue as circular
- Arriving jobs are placed at end
- Dispatcher selects first job in queue and runs until completion of CPU burst, or until time quantum expires
- If quantum expires, job is again placed at end.

4. Conclusion

In cloud computing various users sends request for the transmission of data for different demands. The access to different no. of user increases load on the cloud servers. Due to these cloud server does not provides best efficiency. To provide best efficiency load has to be balanced main problem in the paper is that different jobs can be divides in tasks. The job dependency checking is done on the basis of directed a cyclic graph. The dependency checking the make span has to created on the basis of shortest job first and pound robin approach. The minimization can be done on the basis of using min-min algorithm.

References

[1] Amir Nahir"Distributed Oblivious Load Balancing Using Prioritized Job Replication", ISSN 978-3-901882-48-7, IEEE, 2012.
[2] Hong Tao "A dynamic data allocation method with improved load-balancing for cloud storage system", ISSN 978-1-84919-707-6, PP 220 – 225, IEEE, 2013
[3] Yuqi Zhang "Dynamic load-balanced multicast based on the Eucalyptus open-source cloud-computing system",ISSN 978-1-61284-158-8, pp. 456 – 460, IEEE,2011.
[4] Magade, Krishnanjali A. "Techniques for load balancing in Wireless LAN's**", ISSN** 978-1-4799-3357-0, PP 1831 – 1836, IEEE, 2014.
[5] Yean-Fu Wen "Load balancing job assignment for cluster-based cloud computing", ISSN 14517061, PP 199 – 204, IEEE, 2014.
[6] De Mello, M.O.M.C "Load balancing routing for path length and overhead controlling in Wireless Mesh Networks", ISSN 14630778, PP 1-6, IEEE, 2014.
[7] R. Angel Preethima , Margret Johnson, *"Survey on Optimization Techniques for Task Scheduling in Cloud Environment"*, IJARCSSE, Volume 3, Issue 12, December 2013.
[8] Akhil Goyal, Bharti, *"A Study of Load Balancing in Cloud Computing using Soft Computing Techniques "*, International Journal of Computer Applications (0975 – 8887) Volume 92 – No.9, April 2014
[9] Navjot Kaur, Taranjit Singh Aulakh, Rajbir Singh Cheema, *"Comparison of Workflow Scheduling Algorithms in Cloud Computing"*, (IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 2, No. 10, 2011.
[10] Mayank Singh Rana , Sendhil Kumar , Jaisankar N, *"Comparison of Probabilistic Optimization Algorithms for Resource Scheduling in Cloud Computing Environment"* International Journal of Engineering and Technology (IJET)
[11] C.Kalpana,U.Karthick Kumar, R.Gogulan, *"Max-Min Particle Swarm Optimization Algorithm with Load Balancing for Distributed Task Scheduling on the Grid Environment"*, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 1, May 2012.
[12] ZHANG Yan-huaa, Feng Leia, Yang Zhia, *"Optimization of Cloud Database Route Scheduling Based on Combination of Genetic Algorithm and Ant Colony Algorithm"*, Science direct, Procedia Engineering 15 (2011), pp. 3341 – 3345.