ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

Survey of Different Load Balancing Approach-Based Algorithms in Cloud Computing: A Comprehensive Review

Monika Verma[1], Saurabh Sharma[2]

Research Scholar, Department of Computer Technology & Applications, Gyan Ganga College of Technology, Jabalpur (M.P.), India[1]

Assistant Professor, Department of Computer Science & Engg, Gyan Ganga College of Technology, Jabalpur (M.P.), India[2]

ABSTRACT: Cloud computing may be a kind of computing technology which may be thought-about as a brand new model of computing. It can also be thought-about as a chop-chop rising new technique for providing computing as a service. In cloud computing, several cloud users demand varied services as per their daily new wants. that the perform of cloud computing is to produce all the specified services to the cloud users. however thanks to restricted resources, it's terribly difficult for cloud suppliers to produce all the users desired services. From the cloud suppliers, perception cloud resources should be assigned during a rational manner. So, it's a serious issue to fulfill cloud users satis- faction and QoS necessities. The aim of this paper is to gift a study of previous works in load leveling and QoS strategies utilized in the cloud computing setting. This paper in the main addresses key performance challenges and completely different modeling with their applications for QoS management and simulation toolkits in cloud computing.

KEYWORDS: Cloud computing QoS management Load leveling Virtual machine

I. INTRODUCTION

Cloud computing field has unfold in finding in recent years thanks to its fast user demand. it's associate rising computing technology within the field of knowledge tech- nology (IT). several cloud operators have activated within the market to produce an expensive providing, as well as Platform as a Service (PaaS), Infrastructure as a Service (IaaS), and package as a Service (SaaS) solutions. The QoS among the case of flash and additionally the likes of, storage of our digital photos is, from the customer purpose of scan, where within the cloud. we tend to don't ought to be guaranteed to obtain where, specifically, we've got associate affinity to simply would really like our flash login identification and an online affiliation. we'll see this model as evident in Web-based e-mail too.

There are some necessary characteristics of the cloud computing [1].
Self Service on demand—the cloud users will use on-demand services like server time and network storage while not human interaction with, severally, service supplier.

Location independence—customer typically has no data or management over the specific location of the provided resources however is also ready to specify location at a higher level of abstraction e.g., information center, country, or state. samples of resources embody storage, network information measure, processing, memory, and virtual machines.
Broad network access—the cloud computing capabilities are on the market over the net and accessed through normal

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

techniques that promote use by varied thick or skinny consumer platforms (e.g., laptops, mobile phones, and PDAs).
Rapid elasticity—the cloud computing capabilities will be elastically and apace helped in some cases mechanically, to quickly scale in and apace discharged to quickly scale in. The capabilities on the market typically appear to be infinite and may be bought at anytime in any amount, to the user.

Measuring service—the usage of cloud computing resource will be controlled, reported, and monitored by the cloud suppliers.
The cloud technology stack may be a versatile and straightforward thanks to retrieve and store immense information without concern regarding the package and hardware required. because the variety of customers on cloud will increase, mechanically, the existing resources decreases that creates the matter of delay between the customers and therefore the service pro- viders. The traffic over the net should be dealt.
To rise on top of this downside, plenty of load leveling techniques are meant by
researchers. However, cloud computing has significantly straightforwarded the flexibility provisioning technique though there are several challenges within the field of QoS management. QoS shows the strength of concert, convenience, associated consistency bestowed by the platform further as an application. This paper presents the thoughts of previous load leveling algorithms, the observance of work within the system, and their modeling to the management of QoS in cloud computing setting [2].
The rest of the paper is fastidiously meant as follows: In Sect. 2, we tend to describe the analysis challenges of cloud computing setting. In Sect. 3, we tend to review pre- vious cloud computing work centered on load leveling, QoS management. Section four presents overall discussion regarding the QoS add cloud computing. Finally, Sect. five concludes the paper.

II. RESEARCH CHALLENGES IN CLOUD COMPUTING

There are several research challenges which indicate the need of further improve- ment. The main challenges are as follows:

I. Security and Privacy: Security and privacy is the biggest issue in cloud. It occurs because of movement of networks data and application, loss of control on data, various natures of resources, and several security policies.
II. Performance: The performance is also a big issue in cloud computing. It deliberates the capability of the cloud organization. The outcome may be poor due to not have appropriate assets viz. limited bandwidth, memory, diminutive CPU speed, etc.
III. Efficient Load balancing: By this method, workload has been distributed equally across all the nodes in cloud environment. Load balancing is used to reach good consumer contentment and examine the ratio of resources, and ensure that no any particular node is overloaded, therefore refinement of the whole performance of the cloud.
IV. Resource Management and Scheduling: it can be considered at several
levels viz. software, hardware, virtualization level with performance, privacy, security, and other attributes being dependent on the resources and man- agement. It includes the management of disk space, memory, CPU's, cores, VM images, threads, I/O devices etc.
V. Require a constant and Fast Internet speed: With the help of cloud sys- tem, business gets the capability to save money on software and hardware but still requires spending additional on the bandwidth. This is not possible to fully exploit the services of cloud without high-speed communication channels.
VI. Data center Energy Consumption: With the help of a survey done by Amazon, the cost consumption of its data centers is 53%, and the total cost is used by the servers for a 3-year amortization period while cooling and energy requirements use 42% of the total budget including both the cooling requirements (23%) and direct power consumption (*19%) for amortization period of fifteen years.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

VII. Scale and Quality of Service Management: Although loud computing has significantly eased the competency based method, a lot of challenges of quality of service management. Quality of service means the levels of con- cert, availability, and reliability on hand by the platform and a use or infrastructure that hosts it. Quality of service is elementary for cloud con- sumers and to be expecting from the providers to provide the declared features.

III. LITERATURE REVIEW

In this survey, we cover the works related to quality of service and load balancing in cloud computing. Quality of service can be increased by balancing the load on different machine, so delay will be decreased, such research work based on mod- eling of workload, quality of service management, and load-balancing algorithms in cloud computing environment. The meaning of system modeling is analyzing the concert of a cloud computing whichever at runtime or at design time. The values of quality of service such as reliability, availability, and response time are calculated using these system models. The most common aspects of load balancing of cloud computing are as first, resource pooling in which cloud service provider used the on-demand services using virtualization concepts and multi-tenancy to make readily available resources to the numerous consumers. Second, quick elasticity and flex- ibility permit the cloud system to balance up and down quickly as per desires of the cloud users and provide the capability to free the resources as soon as no longer desired. Third, scalability facilitates on-demand services and resources in the cloud computing environment. Fourth aspect is efficiency to accomplish extremely scal- able system, able to balance the loads as soon as the load increases by a huge amount and a user of cloud computing, user demands more resources online rapidly is very important. An appropriate allocation of responsibilities among the proces- sors can attain these features for the cloud systems environment. At last, Dynamic and Static Resource Allocation in this ways, the load is assigning across cloud computing system, moreover, statically or dynamically. Literature shows that the resource distribution in dynamic ways is better than the static one to retain the dynamic requirements of a cloud user.

Di et al. [3] presented workload guess and pattern analysis that is validated for long-term basis based on a Bayesian algorithm. Here, there are several workloads-based key features to find the possibility of the next features, a Bayesian classifier is used. The researchers defined several workload-based key features and used a Bayesian classifier to evaluate the next possibility of each feature. The tests, however, require resources such as thousand of machines and a large number of related contents or data which is collected from Google data center. In Caprarescu et al. [4], gave a self-organizing approach and also considered Decentralized methods. This approach is able to give proper robust solutions for resource pro- visioning, load balancing, and service deployment in the cloud infrastructure.

In order to forecast and get temporal correlations between loads of various computer clusters in the cloud, a Hidden Markov Models is used by Khan et al. [5]. Various proposals have been made by the authors such as a technique to forecast and classify workloads in cloud systems in order to supply proper cloud resources. A co-clustering algorithm has also been developed by the authors to find servers that have scheme of equal workload which is developed by analyzing and researching the correlations performance for applications on various servers.

In a similar manner, a data center simulation tool DCSIM [6] focused on the dynamic resource management of infrastructure as a service. Each host can run a large number of virtual machines and has a model or power model to give the power consumption of the entire data center.

Pattern recognition techniques are presented by Gmach et al. [7] to cloud workload and data center. Based on pattern recognition and trend analysis, the researchers proposed a workload prediction algorithm whose goal is to find a procedure to allocate servers to various workloads by using the resource pool properly. The synthetic workloads are created to reflect the later or following activities of the workload but the design and trend are analyzed first.

A best practice guide proposed in [8] by Trivedi is to build empirical models. In this paper, major matters involved are the gathering of the very useful content or data, variable-selection procedure, and the modeling technique. The benefits of various prediction approaches have also been properly described by the researchers.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

A Demand Estimation with Confidence (DEC) method has been proposed by Kalbasi et al. [9] to solve the issues of multi-collinearity in regression approaches. Proper enhancement of the estimation accuracy can be achieved through DEC. A service demand estimation from exploitation and end to end response times have proposed by Liu et al. [10]. The issue is evaluated as quadratic optimization pro- grams which are based on queuing formula, and results achieved can be backed out with experimental data. A simulator based on the event GROUDSIM [11] is used for certain applications arranged on large-scale grids and cloud. In cloud computing environment [12], different methods of resource allocation and their applications are discussed. Also, differentially adapted dynamic proportion-based network resource allocation in cloud computing was discussed. The resources are allocated in cloud computing environment via different parameters like maximum efficiency, maxi- mum energy, SLA aware, elevated throughput, highest efficiency, QoS aware, highest energy, and consumption of power. Calbasi [13], presents a method to evaluate online resource demand based on evaluation of least absolute deviations, regression techniques—least squares, and support vector regression. Casale et al. [14] presented an optimization-based inference approach and expressed as a robust linear regression problem which uses open- and close-based queuing network performance models. This approach sums up measurements (i.e., utilization of the servers and system throughput), generally retrieved from log files, to estimate service times.

A proposed resource allocation model [15] deals the consumer's job to an appropriate data center. Their implementation is based on an agent-based test bed using Java Agent. This is adaptively find a proper data centre based on (a) the geographical distance is man-made from network delay stuck between a user and data centers, and (b) every data centres' workload. The coordinator, monitoring agents, and users are agents in this system. The game theory-based method was proposed by author Guiyi et al. [16] to provide solution for recourse allocation in cloud computing environment. The QoS constrained based recourse allocation problem. The binary integer programming method has proposed for preliminary optimization for cloud services. This method has designed an evolutionary mech- anism on the basis of initial results by which it is able to achieve reasonable solution and final best possible. In totality, authors focused on the complicated parallel computing problem on distinct machines associated across the Internet.
A resource allocation and management algorithm of cloud were presented by Bacigalupo et al. [17]. This algorithm is based on forecasting. LQNs methods are used to evaluate the completion of an application install on the systems with strict SLA desires based on data from the past. The researchers also gave the advantage and disadvantage of key the practical use of LQNs in the cloud systems.

The relations between resource consumption and workload for cloud Web applications are studied by Desnoyers et al. [18]. Queuing scheme are useful for representation of various elements of the system, data mining, and machine learning approaches in order to assure flexibility of the model to work under various system conditions. Through the proposed method, great accuracy for forecasting usage of resources and workload is achieved.

To enhance the throughput by means of minimum response time is possible by using new load balancing techniques [19] in the network. The servers can sent and received data in minimum delays by dividing the traffic in between the servers. Lot of algorithms is presents that balance the traffic load between the servers [20]. The well known example of traffic load balancing in our day today life can be related to internet websites. Long response time and more delays experience by the users in earlier server systems without load balancing. Load balancing techniques are plays major role in enhancing quality of services in multimedia applications typically concern redundant servers which help a better allocation of the communication traffic consequently that the website ease of use is categorically advanced [19]. An economic model have presented by Ye et al. [21] based on discrete Bayesian Networks to classify end-users long-term performance considering cloud service supply of an end user. Then, by using effective diagrams, simulation, and analytical experiments, the composition QoS aware service is resolved. Susana proposed [22] a queuing theory-based technique of dynamic load balancing strategy to offer differentiated services. To achieve the differentiated services-

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

based solution, use the key attribute of intensity of concurrency in servers. For maintaining the required distance in normal service times among the service classes by using dynamic load balancing technique based on self-adaptive nature. Markov and Fault Trees models are also used by Jhawar and Piuri et al. in [23] to evaluate the availability and the reliability of the fault tolerance patterns of a cloud environment under the different deployment contexts. The authors have also proposed an approach based on the above evaluation to identify the best approach which is according to the user's requirements. The QoS in cloud computing is enhanced on proper implementing load balancing techniques on the basic of dynamically monitoring the load of the system.

IV. DISCUSSION

The research work proposed by authors on the QoS in cloud computing has greatly increased the various services provided by cloud computing service providers. QoS field aims to improve the QoS by keeping the delay to small quantity by balancing the load in cloud computing. Multimedia, audio, video conferencing, etc. are considered the main applications of this area where enhancement in the delay may be useless.
Cloud has greatly simplified the capacity provisioning method; however, it creates several issues in the management of QoS. Quality of service includes the levels of performance, availability, and reliability offered by the platform or the application or the infrastructure that hosts it. Research in workload modeling and its application in QoS management in cloud computing. Few well-known load bal- ancing techniques via which QoS can be improved in cloud computing

- Dynamic Load Balancing that results fault tolerance, low overhead, high scal- ability can be used in enhancing performance in cloud computing.
- Approaches related to load balancing based on Weighted Active Monitoring to enhance processing time and response time.
- Load balancing based on round-robin technique assigns the virtual machines in circular order; by these methods, some nodes may be under loaded/overloaded and can result in decreasing resource utilization.
- The multiple workflows in cloud computing for dynamic works deal by load balancing technique to improve quality of service.
- A lot of load balancing techniques get evaluated on the basis of different metrics of QoS like throughput, cost, resource utilization, and results show the improvement over existing others works.

This survey shows a lot of improvement over authors works with own works but there are several assumptions. So without considering assumptions taken at the time of evaluation, their improvement may not be considerable. Even though there are many existing works that show improvement, they have disadvantages also. Therefore, we say that there is no any single approach to give better solution in all conditions.

V. CONCLUSION

This paper surveyed the various research efforts that are being carried out in the workload modeling, system modeling, their different applications to QoS man- agement, and its load balancing algorithms in the cloud computing environment. We have also discussed the major challenges that are faced by cloud computing environment and designing of many load balancing algorithm and various modeling techniques for the QoS management in cloud systems. We have discussed the proposed algorithms, but the research work opens many research opportunities to us in the area of load balancing in cloud computing.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

REFERENCES

1. Mell, P., Grance, T.: The NIST definition of cloud computing, Technical report published in NIST Special Publication 800-145 at 25 Oct 2011
2. Petcu, D., Macariu, G., Panica, S., Craciun, C.: Portable cloud applications: from theory to practice. Future Gener. Comput. Syst. 29(6) 1417–1430 (2013)
3. Di, S., Kondo, D., Walfredo, C.: Host load prediction in a Google compute cloud with a Bayesian model. Proceedings in the international conference for high performance computing, networking, storage and analysis, SC, pp. 1–11 (2012)
4. Caprarescu, B.A., Calcavecchia, N.M., Di Nitto, E., Dubois, D.J.: Sos cloud: Self-organizing services in the cloud. In: Bio-inspired models of network, information, and computing systems, vol. 87, pp 48–55. Springer, Berlin, Heidelberg (2012)
5. Khan, A., Yan, X., Shu, T., Anerousis, N.: Workload characterization and prediction in the cloud: A multiple time series approach. In: Proceedings of the IEEE Network Operations and Management Symposium, Maui, HI, USA NOMS NOMS 2012, pp. 1287–1294 (2012)
6. Keller, G., Tighe, M., Lutfiyya, H., Bauer, M.: DCSim: A data Centre simulation tool. In: Proceedings of the 2012 8th international conference on network and service management, and 2012 workshop on systems virtualization management, CNSM-SVM 2012, Las Vegas, NV, USA, pp. 385–392 (2012)
7. Gmach, D., Rolia, J., Cherkasova, L., Kemper, A.: Workload analysis and demand prediction of enterprise data center applications. In: Proceedings of the IEEE 10th international symposium on workload characterization, IISWC, pp. 171–180, Boston, MA, USA (2007)
8. Hoffmann, G.A., Trivedi, K.S., Malek, M.: A best practice guide to resource forecasting for computing systems. IEEE Trans. Reliab. 56(4), 615–628 (2007)
9. Kalbasi, A., Krishnamurthy, D., Rolia, J., Dawson, S.: DEC: Service demand estimation with confidence. IEEE Trans. Softw. Eng. 38(3), 561–578 (2012)
10. Liu, Z., Wynter, L, Xia C, Zhang F, "Parameter inference of queueing models for it systems using end-to-end measurements". Perform Eval. 63(1), 36–60 (2006)
11. Ostermann, S., Plankensteiner, K., Prodan, R., Fahringer, T.: Groudsim: An event-based simulation framework for computational grids and clouds. In: Proceedings of the conference on parallel processing, Euro-Par 2010, Ischia, Italy, pp. 305–313 (2010)
12. RamMohan, N.R., Baburaj, E.: Resource allocation techniques in cloud computing-research challenges for applications. In: Fourth international conference on computational intelligence and communication networks (2012)
13. Kalbasi, A., Krishnamurthy, D., Rolia, J., Richter, M.: MODE: Mix driven on-line resource demand estimation. In: Proceedings of the 7th international conference on network and services management, international federation for information processing, pp. 1–9 (2011)
14. Casale, G., Cremonesi, P., Turrin, R.: Robust workload estimation in queueing network performance models. In: Proceedings of Euromicro PDP, pp. 183–187 (2008)
15. Vignesh, V., Sendhil Kumar, K.S., Jaisankar, N.: Resource management and scheduling in cloud environment. Int. J. Sci. Res. Publ. ISSN 3(6), 2250–3153 (2013)
16. Wei, G., Vasilakos, A.V., Zheng, Y., Xiong, N.: A game-theoretic method of resource allocation for cloud computing services. In Springer, J. Supercomput. pp. 252–269 (2010)
17. Bacigalupo, D., van Hemert, J., Chen, X., Usmani, A., Chester, A., He, L., Dillenberger, D., Wills, G., Gilbert, L., Jarvis, S.: Managing dynamic enterprise and urgent workloads on clouds using layered queuing and historical performance models. Simul. Model. Prac. Theory
19, 1479–1495 (2011)
18. Desnoyers, P., Wood, T., Shenoy, P.J., Singh, R., Patil, S., Vin, H.M.: Modellus: Automated modeling of complex internet data center applications. TWEB 6(2): 8 (2012)
19. Shimonski, R.: Windows 2000 & windows Server 2003 clustering and load balancing, p. 2.
McGraw-Hill Professional Publishing, Emeryville, CA, USA (2003)
20. Brian, A.: Load balancing in the cloud: tools, tips, and techniques. A Technical white paper in
Solutions Architect, Right Scale
21. Ye, Z., Bouguettaya, A., Zhou, X.: QoS-aware cloud service composition based on economic models. In: Proceedings of the 10th international conference on service-oriented computing, ICSOC'12, Shanghai, China, pp. 111–126 (2012)
22. De Saram, S.L., Perera, S., Jayewardene, M.: QoS aware load balancing in multi-tenant cloud environments, published in Int. J. Next Gener. Comput. 4(1) (2013)
23. Jhawar, R., Piuri, V.: Fault tolerance management in IaaS clouds. In: Proceedings of 2012
IEEE first AESS European Conference on Satellite Telecommunications, ESTEL 2012, Rome, Italy, pp. 1–6 (2012)