IJCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Vol.-6, Issue-12, Dec 2018 E-ISSN: 2347-2693

Load Balancing In Mobile Cloud Computing: A Review

C. Arun[1*], K. Prabu[2]

[1,2]Department of Computer Science, Sudharsan College of Arts & Science, Pudukkottai, Tamilnadu, India

Corresponding Author: arunsastha3342@gmail.com, Tel.:+91 9786813342

Available online at: www.ijcseonline.org

Accepted: 08/Dec/2018, Published: 31/Dec/2018

Abstract—cloud computing is said to be the next big boom technology in IT industry infrastructure. It is claimed that it provides new levels of efficiency, flexibility and cost savings of the resources that are used in industries. Mobile Cloud Computing (MCC) is applications, Internet-based data, and related services accessed via smartphones, laptop computers, tablets and other portable devices. By using MCC, the processing and the storage of intensive mobile device jobs will take place in the cloud system and the results will be back to the mobile device. But the mobile cloud computing have some issues like power consumption, bandwidth, mobility and security. Using the mobile devices for accessing the cloud it needs an efficient load balancing technique for offloading the data to the users. In this paper, there is a detailed review on different load balancing techniques which are existing model in cloud analyst tool and some policies by different authors.

Keywords—Mobile Device, Mobile Computing, Cloud Computing, Power Consumption, Bandwidth, Mobility, Security

I. INTRODUCTION

In this era of cloud computing, people leverage cloud services from diverse aspects and enjoy various benefits of cloud computing. Cloud functionalities can be exploited in many ways: infrastructure-as-a-service (IaaS), such as Amazon EC2; platform-as-a-service (PaaS), such as Google App Engine build and deliver web applications; software-as-a-service (SaaS), such as e-mail services (e.g., Hotmail); and web applications (e.g., Google Docs). The increasing commercial adoption of cloud computing is attributable to its advantages over conventional computing, which include reduced cost, easy maintenance, and automatic scaling. Despite the combined advantages of cloud computing, the full potential of mobile cloud computing is far from being fully exploited. Mobile cloud computing (MCC) is a technology that uses computing resources outside of the mobile device. When it comes to mobile handheld apparatus, computing, storage resources of mobile and serious power constraints due to limited battery lifetime are the major contributors leading to a bottleneck. For the efficient processing of large-scale jobs, the mobile device must send jobs to an external cloud server. Offloading means the transfer of data from a computer or digital device to another digital device. Offloading is a solution to augment these capabilities of mobile systems by migrating computation to more resourceful computers, such as servers. This is different from the traditional client–server architecture, whereas thin client always migrates computation to a server. Computation offloading is also different from the migration model used in multiprocessor systems and grid computing, where a process may be migrated for load balancing. The key difference is that computation offloading migrates programs to servers outside the users' immediate computing environment; process migration for grid computing typically occurs from one computer to another within the same computing environment, that is, the grid. Recently, many studies have been conducted on the technology required for offloading processing jobs to external computer resources and for receiving the processed results, thereby overcoming the hardware limitations of the mobile devices. Where the cloud computing architecture uses a server with high computing power, offloading has inherent security problems such as transmission delays and data leakage, which can occur during network-dependent computing data transmission. Therefore, most of the existing studies are focused on the creation of boundaries for security or on the use of relays to improve transmission speed. Such offloading methods can be applied efficiently if the server has sufficient computing power and only the transmission of computing data is required. However, no offloading technologies exist for processing large-scale jobs in an environment where the limited. As an alternative approach to solving the problem of disconnection from the cloud server, the provision of computing services designed only for mobile devices should be researched. In other words, a computing service method is needed that receives resources and services for processing large-scale jobs from other, neighboring, mobile devices and considers the resource state of each mobile device . In this paper we discuss about mobile cloudlet cloud computing architecture in section II, load balancing among cloudlets and load sharing among cloudlets in section III, In section IV significance of virtualization in context to load balancing are discussed, In section V Existing works in load balancing are reviewed.

II. MOBILE CLOUDLET CLOUD COMPUTING ARCHITECTURE

A cloudlet is a computer or a group of computers connected to the Internet and accessible to nearby mobile devices. If the mobile devices do not wish to offload to the cloud due to cost and delay, a nearby cloudlet can be used [1]. Hence, mobile users can meet the demand for interactive response by reduced-delay, single-hop, and high-bandwidth wireless access to the cloudlet [1]. If no cloudlet is found nearby, the mobile device may access the distant cloud or, in the worst possible case, make use of its own resources. Despite the fact that cloudlets successfully deal with the limitations of high WAN latency, they still have two disadvantages [1]. First, mobile users remain dependent on the service provider for providing such cloudlet infrastructure in LAN networks. To alleviate this constraint, a more dynamic cloudlet is created to connect all devices in the LAN network can cooperate in the cloudlet. The second shortcoming of virtual machine (VM)-based cloudlets [1] is the coarse granularity of VMs as an element of allotment. Instead of executing the whole application remotely in the VM, improved performance can be realized by dynamically partitioning the application into components. Moreover, as cloudlet resources are limited, there is a strong probability that the cloudlet runs out of resources when many users run their applications entirely in the cloudlet infrastructure. This limitation can be dealt with if the applications are offloaded in components rather than as a whole. These application components are distributed between the cloudlets. This cloudlet is not fixed; mobile devices can join or leave the cloudlet at runtime. These features eliminate the disadvantages of conventional cloudlets as well as provide a solution to the high WAN latency problem associated with the cloud.

III. LOAD BALANCING AMONG CLOUDLETS AND LOAD SHARING AMONG CLOUDLETS

A cloudlet may be overloaded at some point in time and may share its load with other cloudlets. An overloaded cloudlet may experience delay in providing service or may even provide incorrect results. This situation may hamper the QoS and in turn the QoE. Thus, load balancing among cloudlets is very essential to avoid such circumstances. For load balancing it is important to maintain a load threshold value, which would indicate a maximum load value a cloudlet may have based on its processing capacity of the CPU. A cloudlet cannot accept any load if its current load is at or above the load threshold. Cloudlets with a load below the threshold can accept more load. In this way, the load of each cloudlet can be balanced. The cloudlets can also share the load with other cloudlets. Hence, a cloudlet to which an application is offloaded can split the application into multiple components in such a way that the split components have negligible interdependency. Interdependency refers to the condition in which the output of one component is the input of another component. It is minimized so that the communication between the cloudlets is minimized, and in turn the cost of communication in terms of number of messages transferred between the cloudlets will be less. In addition, the energy required for transferring and receiving messages between cloudlets will be reduced as well. Each component thus obtained will be offloaded to a nearby cloudlet whose load is below the threshold. Each cloudlet will send the result of computation to the cloudlet to which the application was offloaded. This cloudlet will finally combine all the intermediate results to generate the final result and then send it to the requesting device.

a.

IV. SIGNIFICANCE OF VIRTUALIZATION IN CONTEXT TO LOAD BALANCING

The basic definition of Virtualization and load balancing are as follows:

Virtualization: In cloud computing, virtualization is very useful concept which means something which is not real and to create a virtual version of resource, such as a server, storage device, network or even an operating system. It is the software implementation of a computer which will execute different programs like a real machine. Even, something as simple as partitioning a hard drive is considered virtualization because a drive can be partitioned into more than one.

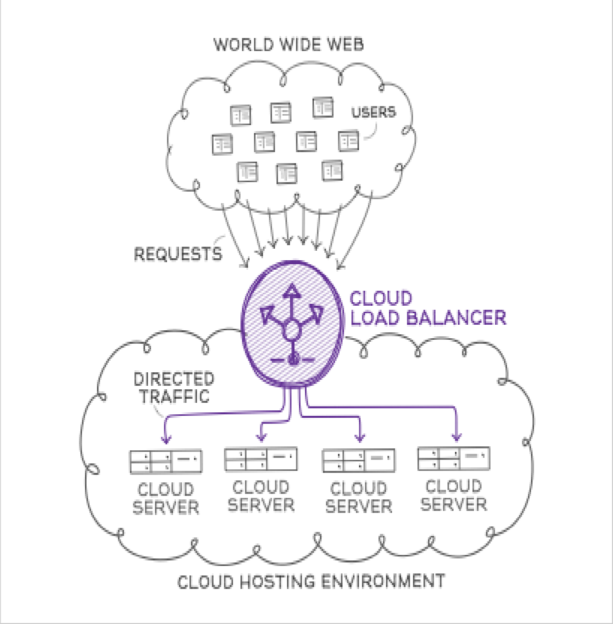

Figure 1.Load Balancing

Load Balancing in Cloud Computing: Load balancing is one of the critical aspects in cloud computing environment that can significantly improve resource utilization, performance and save energy by properly assigning/reassigning computing resources to the incoming requests from users in fig.1. Load balancing takes care of two important things primarily to make sure of the availability of Cloud resources and secondarily to enhance the performance [2]. As technology is growing faster, there are huge amount of users on internet so, to manage and fulfill their requirements, load balancing comes into the picture which ensure that workload is spread equally to all of the available servers without any delay to accomplish higher user satisfaction and maximum throughput with minimum response time [3]. This will ensure,

- Resources are easily available on demand.
- Resources are efficiently utilized under condition of high/low load.
- Reduced energy consumption in case of low load, when the usage of the CPU cycles and memory falls below a certain threshold.
- Reduction in the resource usage cost.

Load balancing helps in the allocation of computing resources to achieve proper resource utilization. High resource utilization with proper load balancing helps in minimizing resource consumption. It helps in implementing scalability and avoiding Bottlenecks. Load balancing techniques help networks and resources by providing a maximum throughput with minimum response time. Load balancing is dividing the traffic between all servers, so data can be sent and received without any delay with load balancing.

Load Balancing is classified in two key approaches based on decisions making process: Static and dynamic load balancing algorithms.

Load balancing strategies for clouds: Load balancing algorithms can be broadly categorized into static and dynamic load balancing algorithms in fig 2.

i. **Static Load Balancing Algorithms:** Static algorithms are much simple as compared to the dynamic algorithms. It must require knowledge of global status of distributed system and does not consider the current state or behavior of a node while allocating the load to the available nodes. It divides the traffic equivalently among all available servers or VMs. It is used when the computational and communication requirements of a problem are known a priori. In this case, the problem is partitioned into tasks and the assignment of the task-processor is performed once before the parallel application initiates its execution.

ii. **Dynamic Load Balancing Algorithms:** Dynamic load balancing is more flexible than the static and they doesn't rely on prior knowledge but depends on current state of the system. In a distributed system, dynamic load balancing has two different ways: distributed and non-distributed. In the distributed one, algorithm is executed by all nodes present in the system and the task of load balancing is shared among these servers. The interaction among nodes to achieve load balancing can take two forms: cooperative and non-cooperative. In the first one, the nodes works side-by-side to achieve a common objective which means is to improve the overall response time, etc. In the second form, each node works independently toward a goal local to it [4].

V. RELATED WORKS

L. Shakkeera et. al., [5] proposed QoS and Load Balancing Aware (QALBA) approach formulates task scheduling using Enriched-Look ahead HEFT algorithm (E-LHEFT). It utilizes MAUI (Mobile Assistance Using Infrastructure) architecture to execute the compute-intensive tasks. E-LHEFT algorithm modifies the processor selection phase of

LHEFT algorithm using task grouping and uses the Pareto principle for effective load balancing of Physical Machine (PM). This strategy saves the battery level of the mobile device and reduces makespan with less latency and achieves load balancing between cloud resources. A.Singha et. al., [6] proposed an Autonomous Agent Based Load Balancing Algorithm (A2LB) which provides dynamic load balancing for cloud environment. This mechanism has been implemented and found to provide satisfactory results.

K.Singh et. al., [7] proposed a new algorithm is proposed load balancing for mobile clouds has been presented. The results have shown that the load balancing during job placement plays a critical role in energy consumption of cloud computing environment.

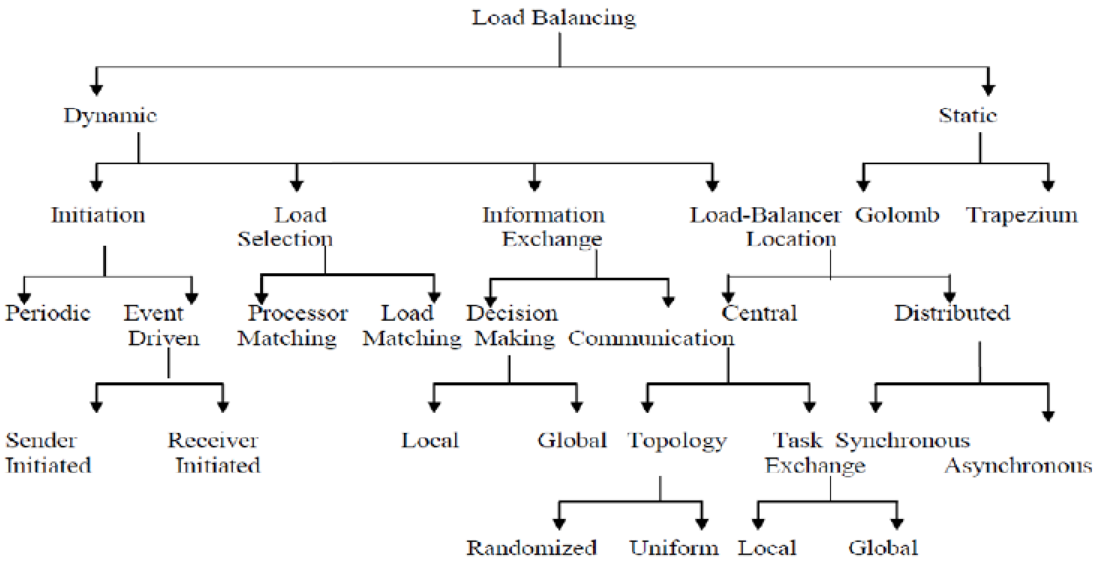

Figure 2.Types of Load Balancing

X.Wei, et. al., [8] proposed architectures into four categories. X.Wei, et. al., [8] presented a Hybrid Local Mobile Cloud Model (HLMCM) by extending the Cloudlet architecture. Then, after formulating the application scheduling problems in HLMCM and bringing forward the Hybrid Ant Colony algorithm based Application Scheduling (HACAS) algorithm. Mobile devices are the higher data consumption when on mobile network data. To solve this problem, use cloud computing to mitigate these large applications and use less data. Integrating in a mobile cloud system to allocate and store these applications will allow for the mobile devices to conserve battery and memory by avoiding large computational processes. I. Cushman et. al., [9] proposed present a new framework that will allow for a smart load balancer to efficiently allocate resources to increase application processing speed for data and request response of memory stored by mobile devices in a secure manner. B.Kim et. al., [10] proposed adaptive mobile resource offloading (AMRO) for processing large-scale jobs using mobile resources without a cloud server. AMRO is applied in a mobile cloud computing environment based on collaborative architecture. A load balancing scheme with efficient job division and optimized job allocation is needed because the resources for mobile devices will not always be provided consistently in this environment.

D.Yao et. al., [11] presented an energy efficient task scheduling strategy (EETS) to determine what kind of task with certain amount of data should be chosen to be offloaded under different environment also evaluated the scheduler by using an Android smartphone. It was achieved 99% of accuracy to choose the right action in order to minimize the system energy usage. C.Chen et. al., [12] proposed a Heterogeneous Mobile Cloud (HMC) computing design that efficiently utilizes the communication and computation resources to support data storage and data processing services in a group of mobile devices. Each mobile device may have different energy, communication and computation capabilities, but our Mobile Storage & Processing System (MSPS) ensures that: i) the communication and computation tasks are executed in an energy efficient manner, ii) task allocation considers device heterogeneity and achieves system-wide load balancing, and iii) the stored data are fault-tolerant.

J.Grover et. al., [13] proposed Agent Based Dynamic Load Balancing (ABDLB) approach in which mobile agent plays very important role, which is a software entity and usually defined as an independent software program that runs on behalf of a network administrator. R.Hasan et. al., [14] modified dynamic energy-aware cloudlet-based mobile cloud computing model (MDECM) was introduced for energy cost awareness in load balancing based on the service rate and energy of the mobile users. To develop the performance and availability of the MCC services B.Zhou et. al., [15] proposed a code offloading framework, namely mCloud. It contains of mobile devices, public cloud services and nearby cloudlets. A context-aware offloading decision algorithm is proposed to selecting wireless medium for deliver code offloading decisions at runtime and suitable cloud resources for offloading.

Changboka et. al., [16] proposed a context model based on ontology in MCC to deliver services to mobile devices based on context-awareness data and distributed IT resources. It offer more correct modified services and manage distributed resources. M. Rahimi et. al., [17] proposed a new framework model namely a location-time workflows (LTW) for mobile applications. User mobility patterns are translated to a mobile service usage patterns and an efficient heuristic algorithm called MuSIC also proposed. The proposed model decreases 35% in price and 25%lower delays and power in the public cloud.

D.Parmar et. al., [18] proposed a mechanism to find a cloudlet for computation loading in a decentralized environment. This identification is classified into two phases. One is cloudlets within Wi-Fi range of the mobile device that are identified without connecting to the cloudlets. Second is selection of the ideal cloudlet for offloading. D.Egbe et. al., [19] proposed a novel context based service discovery algorithm for ad hoc mobile cloud computing using MANETs. The algorithm uses a Search All Pick One Algorithm (SAPOA) to discover the closest service based on the provider's context. This discovery approach is obtaining high quality of service in response time.

L.Chunlin et. al., [20] proposed a model namely multiple context based service scheduling algorithm for satisfy when vast number of mobile requests, increase mobile user's quality of services in experiences and decrease system overheads. The proposed contains of mobile user quality of services optimization sub problem and cloud resources allocation optimization sub problem. The proposed algorithm scheduled the resources according to the context information. D.Kovachev et. al., [21] proposed a model for the mobile users and multimedia context with the help of ontologies. The cloud computing offers service as infrastructure to the mobile clients for the complex semantic multimedia tasks.

VI. CONCLUSION

As cloud computing is picking up prevalence, an imperative inquiry is the means by which to ideally convey programming applications on the offered base in the cloud. Cloud computing is a utility to deliver services and resources to the users through high speed internet. Especially in the setting of mobile computing where programming parts could be offloaded from the cell phone to the cloud, it is vital to improve the association. One of the major issues in Mobile Cloud Computing is load balancing. Load balancing helps in the efficient utilization of resources and hence in enhances the performance of the system. This paper presents a concepts and review of Mobile Cloud Computing along with load balancing.

REFERENCES

[1]. M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, The case for VM-based cloudlets in mobile computing, IEEE Pervasive Computing, 8(4), 14–23, 2009.

[2]. KalaiSelvi B. Mary L. (2014, August). A Survey of Load Balancing Algorithms using VM. , International Journal of Advancements in Research & Technology: IJOART 2014, pp 68-76, ISSN: 2278-7763

[3]. Kaur R. And Luthra P. (2012) .Load Balancing in Cloud Computing. Association of Computer Electronics and Electrical Engineers: ACEE 2014, pp. 375-381, ISSN: 1899-0142, DOI:02.ITC.2014.5.92

[4]. Panwar R., Mallick B. (2015, May) .A Comparative Study of Load Balancing Algorithms in Cloud Computing. , International Journal of Computer Applications: IJCA 2015, pp.33-37, ISSN: 0975 – 8887, DOI: 24, May 2015

[5]. L. Shakkeera and Latha Tamilselvan ," QoS and load balancing aware task scheduling framework for mobile cloud computing environment", Int. J. Wireless and Mobile Computing, Vol. 10, No. 4, 2016 309.

[6]. A.Singh, D.Juneja and M.Malhotra," Autonomous Agent Based Load Balancing Algorithm in Cloud Computing", International Conference on Advanced Computing Technologies and Applications (ICACTA-2015), Procedia Computer Science 45 (2015) 832 – 841

[7]. K.Singh," Energy Efficient Load Balancing Strategy for Mobile Cloud Computing ",International Journal of Computer Applications (0975 – 8887) Volume 132 – No.15, December2015 6

[8]. X.Wei, J.Fan, Z.Lu and K.Ding," Application Scheduling in Mobile Cloud Computing with Load Balancing", Journal of Applied Mathematics,Volume 2013.

[9]. I. Cushman, M.Sadi, L.Chen and R. Haddad," A Framework and the Design of Secure Mobile Cloud with Smart Load Balancing", DOI: 10.1109/MobileCloud.2017.41

[10]. B.Kim, H.Byun, Y.Heo and Y.Jeong ," Adaptive Job Load Balancing Scheme on Mobile Cloud Computing with Collaborative Architecture", DOI: 10.3390/sym9050065

[11]. D.Yao, C.Yu, H.Jin, and J.Zhou," Energy Efficient Task Scheduling in Mobile Cloud Computing", DOI: 10.1007/978-3-642-40820-5_29.

[12]. C. Chen, R.Stoleru and G.ie," Energy-efficient Load-balanced Heterogeneous Mobile Cloud", DOI: 10.1109/ICCCN.2017.8038422.

[13]. J.Grover, S.Katiyar ,"Agent Based Dynamic Load Balancing in Cloud Computing", DOI: 10.1109/ICHCI-IEEE.2013.6887799

[14]. R.Hasan and M.Mohammed ,"A Krill Herd Behaviour Inspired Load Balancing of Tasks in Cloud Computing "*Studies in Informatics and Control*, ISSN 1220-1766, vol. 26(4), pp. 413-424, 2017.

[15]. B.Zhou, A.V.Dastjerdi, R.N. Calheiros, S.N.Srirama, and Rajkumar Buyya,"mCloud: A Context-aware Offloading Framework for Heterogeneous Mobile Cloud," IEEE Transactions on Services Computing vol: 10, issue: 5, 2017, pp: 797 - 810.

[16]. Changboka, H.Chang, H.Ahn and Euiin Choi," Efficient Context Modeling Using OWL in Mobile Cloud Computing", Energy Procedia, vol:16, Part B, 2012, pp:1312-1317.

[17]. M. Rahimi, N.Venkatasubramanian, and A. Vasilakos,"MuSIC: Mobility-Aware Optimal Service Allocation in Mobile Cloud Computing", International Conference on Cloud Computing (CLOUD), DOI: 10.1109/CLOUD.2013.100

[18]. D.Parmar,A. Kumar, A.Nivangune ,P.Joshi and U.Rao,"Discovery and Selection Mechanism of Cloudlets in a Decentralized MCC environment", International Conference on Mobile Software Engineering and Systems (MOBILESoft), 2016, DOI: 10.1109/MobileSoft.2016.017

[19]. D.Egbe, M.B. Mutanga and M.O. Adigun," Service Discovery In Ad-Hoc Mobile Cloud: Contemporary Approaches And Future Direction", Journal of Theoretical and Applied Information Technology, August 2016. Vol:90,pp:101-117.

[20]. L.Chunlin,Y.Xin, Z.Yang and L.Youlong,"Multiple Context Based Service Scheduling for Balancing Cost and Benefits of Mobile Users and Cloud Datacenter Supplier in Mobile Cloud", Computer Networks, Vol:122, July 2017, pp: 138-152.

[21]. D.Kovachev and R. Klamma," Context-aware Mobile Multimedia Services in the Cloud", International Workshop of the Multimedia Metadata Community on Semantic Multimedia Database Technologies, 2009.

Authors Profile

Mr. C. Arun received his M.Sc degree from Bharathidasan University, Trichy, Tamilnadu and M.Phil from Bharathiyar University, Coimbatore, Tamilnadu, India. He is doing his Ph.D in Sudharsan College of Arts & Science, Pudukkottai, Tamilnadu, India. Mobile Computing and Cloud computing are his interested area in research. He has published more than 12 technical papers at various National / International Conferences and Journals.

Dr. K. Prabu received his MCA and M.Phil from Annamalai University, Chidambaram, India. He received his Ph.D Degree in Computer Applications from Manonmaniam Sundaranar University, Tirunelveli, India. He is now working as an Assistant Professor in PG& Research Department of Computer Science, Sudharsan College of Arts &amp; Science, Pudukkottai, Tamilnadu, India. He is a Reviewer of 06 National/International Journals. His Research interested is Adhoc Networks, Wireless &amp; Mobile Computing, and Wireless Sensor Networks. He has published more than 50 technical papers at various National / International Conferences and Journals. He is a life member of ISTE, IACSIT, and IAENG.