Review of „Issues and Challenges of Load Balancing Techniques in Cloud Computing: A Survey: P. Kumar and R. Kumar" by Amalia Stefanescu

The following paper has the purpose to present the main concepts derived from the research done by P. Kumar and R. Kumar on the topic of issues and challenges in cloud computing when it comes to load balancing techniques. Load balancing has an important purpose in the cloud since it enables data centers to work efficiently without the issue of overloading or loss of performance.

Due to its practicability load balancing has seen an increased popularity in the last years which led developers to search for even more efficient algorithms that are able to deliver desired results. The paper reviewed is divided in three main topics: a study of load balancing techniques, a classification of them and the issues they pose followed by an overview of the research areas that are going to see an improvement in terms of load balancing in the future.

At its core, load balancing uses available resources in order to distribute workload evenly and efficiently, avoiding infrastructure blockers and providing continuous service in the cloud. The resource management is done fairly between tasks while also providing priority of execution based on necessity. In terms of cloud computing we can distinguish two main categories: based on services offered and location. The former helps in minimizing the need for configuring basic infrastructure while the latter helps organisations and individuals to manage the visibility of their infrastructure providing security and data sharing control. In order to reduce costs and improve response time, cloud computing is continuously developing techniques for optimum resource management.

The paper written by P. Kumar and R. Kumar researches the main challenges in load balancing after an intensive study of various scientific articles published on the topics of resource scheduling, execution cost, system performance and performance monitoring, quality of service, energy consumption, metrics, consumers and so on.

Among the issues of load balancing in the cloud a few are selected as primary among which are mentioned the following: geographical distributed nodes, algorithm complexity, storage management, load-balancer scalability, virtual machine migration, single point of failure, and heterogenous nodes.

Geographical distributed nodes raises the issue of resource distribution across big distances. The farther apart that the nodes are, the more difficult it is for load balancing algorithms to spacially distribute the nodes. However, some algorithms are designed for small regions and do not take into account issues that arise due to large distances between the users and the resources they try to access, such as network delay or loss of communication.

Algorithm complexity is an issue that can affect all layers of development, thus it is mentioned in the case of cloud computing as well. In order to design efficient systems, the algorithms used have to be readable and must not pose high implementation challenges. Complex algorithms have the disadvantage of diminishing performance and that can lead to timeout errors.

Storage management has seen major improvements along the years with the cloud infrastructure, thus giving up the manually managed storage systems that took a lot of effort and costs to maintain. The cloud storage systems are able to store any type of data the users needs while also being readily accessible. However, an issue that is becoming more and more apparent is caused by the amount of data stored in the cloud which is increasing very fast.

The disadvantage of a lot of data stored in cloud is that it needs to be accessed just as fast while also maintaining the consistency. In order to do this, a replica of the data is stored as well, which leads to duplicates that take up space redundantly. Partial replication, on the other hand, solces the issue of duplicate data but it increases the complexity and tends to fail due to dataset availability issues. For this reason, there's a need for implementing load balancing techniques that use partial replication and take into consideration the data distributed alongside the application.

Load-balancer scalability reffers to quick access of the users to application/ services that are contained in the cloud. The technique has to take into consideration any infrastructure changes in order to efficiently scale up or down, some of the changes in demand being related to storage, high computations and system shape.

Virtual machine migration enables the use of multiple independent virtual machines on the same physical machine. In order to load balance the distribution of computation once a virtual machine gets overloaded, such an algorithm is used in order to migrate to other machines that may be distantly located.

Single point of failure is a challenge for non-distributed load balancers that delegate the task to a single node of the application. In case that node fails, then the resources may become overloaded leading to crashes in the system. For this reason, there should be multiple nodes in order to make sure that the tasks can be carried over to a different node once one has failed.

Heterogenous nodes are a need caused by the continuous changing in demand and user requirements in the technical world. These changes have to be distributed efficiently across the available resources and they must be responsive, thus research has gone into developing a heterogenous environment for the cloud which deviated from the early studies into homogenous nodes.

Existing load balancing techniques that try to solve some of the issues are primarily based on the state of the system and divided into static and dynamic techniques. The static approach is desirable for non-distributed system that use a small number of resources because it is fast and easy to implement. The resource load is divided before the execution time and the smaller tasks are executed before the bigger ones. However, they are executed using the round robin approach in order to avoid starvation. On the other hand, dynamic load balancing techniques take into acccount the state of the system prior to distributing the tasks and take a decision based on this. They are more suited to the cloud ecosystem because in case of overloading the load is transferred to a different machine which is underloaded. These techniques monitor the load of the system continuously and improves performance by calculating and redistributing workload. Some examples of this technique are Agent-based load balancing and Ant Colony optimization.

In conclusion, the paper reviews load balancing techniques and their challenges in order to identify better ways to improve new algorithms and researches in the future. Based on the already discovered techniques, more research is needed into improving the performance of the load balancers while also reducing the high energy consumption needed in order to process big amounts of data.

References:

Pawan Kumar and Rakesh Kumar. 2019. Issues and Challenges of Load Balancing Techniques in Cloud Computing: A Survey. *ACM Comput. Surv*. 51, 6, Article 120 (February 2019)