Load Balancing in cloud computing – A hierarchical taxonomical classification
-Summary-

Written by Shahbaz Afzal and G. Kavitha, it is an article that tries to present in a detailed and encyclopedic manner, the different types of load balancing techniques, drawing the lines between the advantages and limitations of the existing methods. By doing so, the authors engage into a discussion as to what the future holds for load balancing algorithms and what are some of the insights that should be held for the future in this branch of cloud computing.

The article opens with the definition of the Cloud Computing and its main "giant players", as the paper describes them, the main companies that are already using it and adopting it more and more by day: Google, Amazon, Microsoft, Oracle, etc. Load balancing is also defined in this section of the article, together with the summary of the four main contributions of the paper:

1. It tries to explore the main causes for the problem of load unbalancing in cloud computing
2. It outlines an analysis over the already existing approaches in the process of load balancing and the way they have been used in Cloud Computing.
3. It provides a thorough classification of the different techniques, methods strategies and algorithms, all used in load balancing.
4. The article offers an analysis over the different challenges that have been faced by researchers while trying to develop efficient algorithms.

Throughout the next section, the authors provide a brief overview over a two-level architectural load balancing model and how it is used in imbalanced clouds for solving the problem of load shedding. The architecture is explained as follows: the request generator oversees creating user tasks that require computing resources for their execution, the data center controls the management of the tasks, the first level of the load balancer distributes the workload among the PMs by balancing it to the associated VMs, and the second level load balancer stabilizes the workload across the VMs of PMs.
The main activities involved in load balancing are also presented in this chapter:

- User task requirements identification
- VM resource details identification
- Task Scheduling
- Allocation of resources
- Migration

The next section, as the title implies, outlines other related work that has been done in the field of cloud computing, crediting authors and papers for their work in the field of load balancing and cloud computing alike.

The 'Research methodology' section outlines the way authors tackled the research done for the writing the paper and what are the main questions the section tries to answer:

1. "What causes the load unbalancing problem?"
2. "Why load balancing is the need of hour in cloud computing?"

3. “Does load balancing consider the evaluation of single objective or multi-objective functions?”
4. “What is time complexity of load balancing algorithm?”

While other papers classify the load balancing algorithms in their own way, the authors concluded that there is also a need for a classification that best gives the perspective over where algorithms stand in taxonomy, thus giving the purpose of the whole article. This is the focus point of the fifth section of the article, that also presents the various criteria used for the multiple classifications of the algorithms: “nature of the algorithm”, “state of the algorithm”, “trait used for load balancing”, “type of load balancing” and “technique used in load balancing”.

- By the nature of the algorithm, the authors divide the types of algorithms into **proactive based approaches** (which acts according to both the change and the cause of the change), depicted in Table 1, and **reactive based approaches** (which acts in response to a situation rather than controlling it), as depicted in the second Table.
- Considering the state of the algorithm (criteria which is the most widely used for classifications), LB algorithms are classified as **static, dynamic and hybrid** (a mix between the first two). In static LB algorithms, the traffic is isolated uniformly across the servers by the algorithm, which is already aware of the task complexity and system resources. Dynamic LB algorithms, on the other hand, are more complex than the static ones, and they are already in control of the traffic flow at run time, while overseeing the change of a running task’s state at any point in time.
- Judging by the trait used for the LB, the algorithms can be associated as **scheduling** algorithms and **allocation** algorithms. Each policy plays a crucial role in the monitoring of the performance and the management of the resources in cloud. Furthermore, each policy is divided into three subsequent activities similarly from one policy to the other (Task S/All, VM S/All, Resource S/All)
- The functionality aspect distributes the algorithms in hardware LB alg. and Elastic LB alg., each one in charge of the hardware or software (containers, IP Addresses, etc.) aspects of the system.
- Based on the technique used in LB, algorithms are classified as heurists and meta-heuristics techniques, and optimization techniques, designed to find, either the most simple and sufficient method to find a solution to the given problem, or to find the optimal solution that can be generalized for future problems.

During the sixth section of the article, the authors explore and outlines the results they have achieved by the classifications mentioned above on different LB algorithms in Cloud Computing. While giving percentages for each of the category explained above, they perfectly describe the algorithmic system of LB algorithms.

“Conclusion and future work” is the final section of the article, and, as the name suggests, the scientists create a short summary over what they have observed and tried to present in this article.
Using 35 articles and considering the first criteria presented in the article, used for the classification, the most important conclusion to which they have reached is that 100% of proactive approaches are dynamic, but not all dynamic approaches need to be proactive (i.e., they can represent reactive approaches as well).
Another important statistical conclusion revealed by the studies is that, while trying to determine the actual performance of an algorithm, 80% of the examples chosen by the

authors does not consider the complexity of the algorithm as a major turning point in that performance.
As some final thoughts, the authors finish their work by deducing that, by applying efficient and sophisticated algorithms along the metrics of the quality of the services, and correctly evaluating algorithms in each situation, the gaps that currently exist in the load balancing process can be significantly reduced, guiding future researchers, at the same time, to deal with the already lasting problems in Cloud Computing industry.