CHAPTER 6

A LOAD BALANCING SCHEME FOR EDGE DATA CENTRE IN FOG COMPUTING

Mohd Husaini Mohd Fauzi, Sharifah Hafizah Syed Ariffin, and Muhammad Ariff Baharudin

6.1 INTRODUCTION

An edge data centre is a data centre located at the edge of the core network. Currently, the cloud data centre is hosted in the cloud for data storage. It can be used for various applications like extensive data analysis or a simple read-and-write database.

The increasing number of connected devices will generate more packets into the network. A high number of packets in the network will create increased traffic. An application hosted in the cloud data centre is physically far from its user. So due to the high traffic in the network, it will delay the user's use.

The purpose and requirements of the application will determine if hosting the application in the cloud is suitable. Suppose the application requires fast interaction between the user and the application. In that case, the application is impractical to be hosted in the distant cloud. Otherwise, the application can be hosted in the cloud. Suppose the application is a storage-dependent type that requires fast data processing. In that case, it must have a low-latency connection to the cloud data centre. What if the nearest cloud data centre could not satisfy the latency requirement? This situation is where the edge data centre is required.

The characteristics of a cloud data centre are that they are always available and have extremely high bandwidth, processing power, and scalable storage capacity. Meanwhile, an edge data centre has limited bandwidth, limited storage, and limited computing capacity, but it has location awareness. An edge data centre must have location awareness because it is meant to serve any applications in its area. The area can be small as a building, district, or country region, depending on how big its resource is to handle the load. Anything outside or far from an edge data centre should not take because it will not fulfil the purpose of having an edge data centre.

Two types of networks define access to a network: private and public. The public allows anyone to have access to the network. In contrast, the private will allows a specific user access to the network. Suppose an edge data centre is deployed in a private network. In that case, only the member connected to that network can use the data centre features.

In comparison, the edge data centre in the public network is open to any application nearby. This mode allows the edge data centre to have an influx of requests that can cause overload and directly create an effect similar to the denial of services (DOS). This overload can disrupt the services.

6.2 LOAD BALANCING CONTROL WITH SWITCHES

An edge data centre on the public network will have an issue of overload requests coming from nearby applications. There must be coordination between the edge data centres to exchange their load to reach a suitable load level and avoid overloaded. Load balancing is not a new solution to a load-managing problem. It is always there to try to maximise the usage of devices that are always underutilised. It is also used to improve the performance of a device that overloads with work.

Some of the past work already tries to implement a load-balancing method. The following section will describe related work in the research community for a load management problem.

6.2.1 Controller Placement Affects Load Balancing

Compared to a traditional network, a software-defined network (SDN) would take the management away from the individual router. Each router would deliver the packet from a source to a destination. However, the path or route would be decided by a controller where the management part is. The management or controller will handle multiple routers at a time and control the traffic flow in the network. The router will reroute all packets received according to the controller's rules and policies.

According to the work by Isong et al. (2020), a controller's load depends on the number of switches controlled by a controller. An overloaded controller can cause queueing delay, or a new request not being served.

6.2.2 Switch Migration in Software-defined Network

One of the SDN components is a programmable switch. The switch will conduct the routing part, where a controller sets the policy. A controller can control multiple switches. However, a switch can be controlled by one controller at a time.

The more switches connected to a controller, the more load, the controller will have. This condition will happen if all switches have the same load, but each switch's load depends on the end device connected to the switch. Some switches carry more load than the other switch, which is not the same all the time. This situation is called a dynamic load, which varies depending on time, place and network status.

Liu et al. (2021), proposes a highly efficient switch migration method to balance the network. According to him, the current switch migration efficiency is low because the load balancing performance of controllers does not improve significantly after migration. It also produces extra migration costs.

Adekoya et al. (2021), is proposing an improved switch migration algorithm because two frameworks, Dynamic and Adaptive Load