Research on Data Center Load Balancing Technology Based on SDN

Cheng Cheng*
School of Computer and Information Engineering, Henan University, Kaifeng 475004, China

*Corresponding author’s e-mail: 104753200831@henu.edu.cn

Abstract. With the rise of mobile Internet and big data, the continuous expansion of data center network deployment, the trend toward more complex traffic patterns and fine-grained performance management requirements, have brought more and more network performance issues to the fore. As an emerging network architecture, SDN separates the control plane from the data plane, and has the characteristics of centralized control and programmable. The centralized control feature of SDN technology enables the SDN controller to dynamically obtain network topology, link information, node information, and flow information in a timely manner, thereby realizing flexible and fine-grained centralized scheduling of network flows, which is very suitable for the needs of data centers. Due to the high dynamic nature of SDN, how to improve the timeliness of its load balancing and better solve network congestion is very important. Aiming at the problems in data center network applications, this paper studies the SDN-based data center network load balancing management system structure, and discusses the load balancing technology of data center server clusters based on improved ant colony algorithm.

1. Introduction
With the rapid development of cloud computing and big data industries, a batch of interconnected product models have been derived, which has brought huge challenges to the traditional network architecture. As the carrier of large-scale service applications and massive data, data center network performance has received more and more attention, especially how to make full use of data center link bandwidth and server resources to increase network throughput and reduce network delay, etc [1]. With the continuous expansion of network scale, low bandwidth utilization and network congestion problems become more prominent. The new Internet applications put forward higher requirements for the flexibility and scalability of network equipment [2]. On the one hand, the underlying network resources can be centrally controlled through the control layer, and on the other hand, the burden of forwarding equipment can be greatly reduced. The characteristic of SDN network centralized control enables the SDN controller to obtain network topology, link information, node information, and flow information, thereby achieving flexible and fine-grained centralized control of network flows. The world-renowned network equipment manufacturers combine with traditional network equipment to realize their own unique SDN network with brand-new thinking. The core concept of the network is the separation of control and forwarding [3]. The control plane is separated from the network equipment and integrated into the controller, and then the control interface of the forwarding plane is used to control the forwarding equipment. In a large-scale network deployment environment, because the controller needs to manage tens of thousands of forwarding devices, a single controller has limited
control capabilities, and the load pressure is too heavy, and the pathfinding request of the network device cannot be responded in time [4].

2. SDN related technologies

2.1. SDN technology
SDN technology separates and decouples the control logic and forwarding logic in network equipment, so that it is not directly related in structure, and can be expanded and developed on their respective planes. In this network system, a unified forwarding level interface can be established to establish a unified protocol standard for network equipment. SDN concentrates all control logic in the controller part to realize a unified control logic platform. In the SDN network framework, all resources are logical, and physical network devices need to be virtualized. The controller uniformly controls the physical resources of the entire network. The control plane provides corresponding compatible support for multiple protocols [5]. The application layer is a unified interface provided by the control layer, and network applications can control the entire network as long as they call the corresponding interface. In the SDN network architecture proposed by ONF, the network is divided into three layers, namely the application layer, the control layer and the data layer. The application layer is the top layer, responsible for providing services to users; the middle layer is the control layer, responsible for network scheduling; the bottom layer is the data layer, responsible for the transmission of data traffic. The interface between the application layer and the control layer is called the northbound interface, and the interface between the control layer and the data layer is called the southbound interface. A typical SDN network architecture is shown in Figure 1.

2.2. Data center network flow
The increase in the east-west flow between servers requires the data center to have a larger cross-section bandwidth and more redundant links between servers to meet the ever-increasing data center business needs. At the same time, it is necessary to perform load balancing and congestion control on the data center network, because simply increasing the cross-section bandwidth in the east-west direction cannot completely solve the network congestion. Data center data flow is divided into elephant flow and mouse flow. The number of elephant flow is about 8%, but it occupies 90% of the network bandwidth, while the mouse flow size no more than 10KB occupies a lot of small. The network bandwidth is generally a low-latency burst stream and is relatively dynamic. Elephant flow is easy to cause network congestion, so it is the focus of data center network research, while burst flow requires fast response speed and low delay, so it is necessary to ensure the bandwidth requirements for its transmission. There are many burst and sudden flows in the data center. Data center networks have multiple streams with different characteristics that are highly mixed, and different streams have...
different requirements for network performance and resources, which brings greater challenges to data center network load balancing [6].

2.3. Load balancing technology
The development of network technology to this day, load balancing technology has always been an enduring topic. The network load balancing technology is to evenly distribute the client's request tasks to each server in the background server cluster to avoid the situation that some server devices are too busy while the other server devices do not fully exert their processing capabilities. Related technologies that collect, classify, and count related traffic in a distributed network, and use effective load balancing algorithms to determine traffic distribution strategies, and ultimately speed up network traffic processing, optimize overall network performance, and improve user experience. The introduction of load balancing technology can make better use of the resources of each server in the data center server cluster and maximize the processing capacity of the entire server cluster. One kind of load balancing technology is partial load balancing technology. This technology uses limited server resources to deploy load balancing applications on the nodes that are most likely to cause overall performance bottlenecks, and uses a variety of effective load balancing strategies to guide traffic distribution to each Node to improve service performance. The other is the global load balancing technology, which requires more server resources [7]. This solution is to deploy the load balancing application directly to the user side, and distribute the traffic to the nearest effective bandwidth according to the user's request traffic. The most common is CDN technology. There are also many practices in load balancing solutions for data center networks.

3. SDN network load balancing technology based on link bandwidth

3.1. Link bandwidth load balancing system architecture
The characteristics of centralized control of existing SDN network architecture have great advantages for studying how to improve network bandwidth utilization and balance network load, but the current load balancing function based on SDN network architecture is relatively weak, and some functions are not implemented in SDN architecture, Such as congestion detection and elephant flow discovery. The SDN network load balancing technology based on link bandwidth adds congestion detection, elephant flow discovery, and global scheduling algorithms to the original SDN network to achieve network load balancing.

The link bandwidth load balancing architecture is also divided into three levels: the infrastructure layer, the control layer, and the application layer. The load balancing part of the application layer is the core of the architecture, and the infrastructure layer and the control layer provide network flow for load balancing, network topology and path bandwidth information support. The Open Flow switch at the infrastructure layer communicates with the controller through the Open Flow protocol. The control layer provides network services such as topology information, flow table information, network state management, common flow routing calculation, and flow information analysis for the application layer. Efficient acquisition of flow statistics is the key to the generation of network flow views, which includes measurement data collection and measurement data analysis. Flow information can provide basic data sets related to flow status for a variety of upper-level management applications [8]. The controller (topology discovery module) uses the LLDP protocol to discover the network topology. The process is: the controller sends a Packet-Out message to a switch, if the switch port that sends the data packet is also connected to an Open Flow switch, the switch will actively send a Packet-In message to the controller, and the controller will get a piece of link information. The SDN network load balancing architecture based on link bandwidth is shown in Figure 2.
3.2. Elephant flow detection based on application perception
Elephant Streaming has a relatively large impact on the network due to the relatively large amount of data it carries or the long duration. Moreover, the number of elephant streams in the data center network is relatively small and it occupies most of the bandwidth of the network link, so it is large. Image flow is the focus of load balancing. Elephant Streaming found that there are currently problems with detection speed and detection accuracy. Taking into account the rapid changes in the network data flow of the cloud data center, the detection of the elephant flow should be faster, and at the same time, high accuracy should be ensured, and a mechanism for identifying the elephant flow based on the classification of cloud applications should be established; dynamically adjust the elephant flow threshold. Adjust the threshold according to the bandwidth usage between Controller and Switch. If the bandwidth usage between Controller-Switch is low, lower the elephant flow threshold to improve the forwarding efficiency of network data flow; otherwise, increase the elephant flow threshold to avoid the secure channel between Controller-Switch Become a communication bottleneck. The rapid detection framework of elephant flow based on application perception is shown in Figure 3.

4. Data center server cluster load balancing technology based on ant colony algorithm
On the basis of a clear understanding of the basic ant colony algorithm, the research on the data center server cluster load balancing technical solution based on the ant colony algorithm is mainly composed of the SDN controller, the underlying OpenFlow network and the server cluster. The server status monitoring module is mainly composed of a load collection module and a comprehensive load calculation module. The load collection module regularly collects the load information of the servers in the cluster through the SNMP protocol, including the utilization of CPU, memory, network bandwidth, and disk IO, and saves the load information in the database. The comprehensive load calculation module calculates the comprehensive load of the server according to a certain strategy according to the collected load information. According to the comprehensive load information, based on the improved ant colony algorithm, the tasks arriving at the same time are allocated to the corresponding server, and the response of the task is minimized. Time, service cost and load difference between servers.

When a user initiates a request to the IP address segment of the server cluster, the SDN controller decides to allocate the requested task to the appropriate server in the cluster according to the improved ant colony algorithm scheduling strategy. The SDN controller inserts a new rule into the flow table of the switch that receives the request. According to this rule, the improved ant colony algorithm scheduling module is used to obtain the IP address and MAC address of a server from the cluster. This...
address is the flow. After the change is completed, the SDN switch forwards the request packet to the new destination server according to the rules. After the server processes the request, the SDN controller changes the source IP and MAC address of the response packet header to the original destination IP and MAC address of the request packet, ensuring that the response packet received by the client always comes from the host corresponding to the destination IP address and MAC address at the time of the request. The SDN data center cluster load balancing architecture based on the improved ant colony algorithm is shown in Figure 4.

![Figure 3. The rapid detection framework of elephant flow based on application perception](image1.png)

![Figure 4. SDN data center cluster load balancing architecture based on improved ant colony algorithm](image2.png)

Establish a graph model for the scheduling problem of batch tasks in server clusters, solve them through ant colony algorithm, and allocate batch tasks reasonably through certain strategies, shorten the overall response time of batch tasks, improve server cluster processing performance, and achieve better Load balancing effect. Server resource scheduling, according to certain strategies to reasonably allocate tasks to servers for execution, the allocation plan should ensure that the execution time of the entire system is the shortest, the cost is the lowest, and the overall load of the system is the most balanced. Each node is identified by a task, a processor unit, and a pheromone. The two elements of the task and the processor unit can estimate the impact of the scheduling on load fluctuations. The pheromone represents the number of ants passing through the node, that is, the number of scheduling. Servers and tasks are divided into ranks. Since we stipulate that the same task cannot be processed by multiple servers, nodes in the same row represent the scheduling of tasks by the same server. Since one server can handle multiple tasks, the same nodes in the row are all connected by arcs, and the nodes in the same column represent the scheduling of the same task on different servers. The task scheduling process based on the improved ant colony algorithm is shown in Figure 5.
5. Conclusion
The rise of SDN networks provides new research ideas for the study of network load balancing. Its centralized control feature can globally obtain network topology, flow information, and link bandwidth information. It provides different solutions for elephant flow discovery, congestion control, and flow scheduling. Implementation methods, these methods are difficult or impossible to achieve in traditional networks. The controller cluster load balancing system is an effective solution to the uneven load problem caused by the controller expansion scheme. Although the SDN network framework has great advantages in the research of load balancing, the original SDN network only provides basic network functions such as link discovery, routing calculation, and flow table distribution. In the research of this article, the data center network flow, SDN network load balancing technology based on link bandwidth, and data center server cluster load balancing technology based on ant colony algorithm are discussed. These are all effective ways to solve the problem of SDN-based data center load balancing. Due to time reasons, this article did not conduct more detailed experimental verification and large-scale real environment implementation. In the follow-up research, the related work needs to be further refined.

References
[1] Wang Y and Kuang Y W 2015 SDN-based cloud center dynamic load balancing method. Journal of Guilin University of Electronic Technology, 35(04): 321-324.
[2] Huang X M 2015 Research and implementation of load balancing technology for SDN network controller. Nanjing University of Posts and Telecommunications.
[3] Wen K Y 2016 Data center load balancing and flow control technology research. Nanjing University.
[4] Hu R C 2017 Research and implementation of cluster-oriented multi-SDN controller load balancing technology. Southeast University.
[5] Lang W M Yao J F and Zhao Y F 2018 Research on Load Balancing Technology of Big Data Center. Telecommunications Express, (04):1-4.
[6] Ruan C H 2018 Server load balancing design based on SDN technology. Journal of Beihua University (Natural Science Edition), 19(03): 405-409.
[7] Hu T 2018 Research on Multi-controller Load Balancing Technology in Software Defined Network. Strategic Support Force Information Engineering University.
[8] Zhong Y 2019 Technical research of SDN load balancing based on ant colony algorithm. Microcomputer Applications, 35(05): 65-67+77.