Journal of Quantum Computing
DOI: 10.32604/jqc.2022.027560
Article

Tech Science Press

Research on Service Function Chain Orchestrating Algorithm Based on SDN and NFV

Yuning Jia*, Yu Gong and Yifei Wei

Beijing Key Laboratory of Work Safety Intelligent Monitoring, Beijing University of Posts and Telecommunications, Beijing, 100876, China
*Corresponding Author: Yuning Jia. Email: 1073967337@qq.com
Received: 25 January 2022; Accepted: 27 April 2022

Abstract: Software defined network (SDN) and network function virtualization (NFV) have become a new paradigm of a new generation of network architecture. SDN and NFV can effectively improve the flexibility of deploying and managing service function chains (SFCs). By combining SDN and NFV and applying them to the resource orchestration problem of SFC deployment, the three-tier architecture consisting of SDN controller, network function virtualization and physical underlying computing resource layer in the process of heterogeneous network resource mapping is considered. And an optimization algorithm for active control resources based on SDN and NFV is proposed. Firstly, the user's utility is modeled by the multi-standard aggregated multi-criteria utility algorithm, and the optimization goal is transformed into the problem of maximizing the user's utility. Then the controller, based on the algorithm's prediction of the future state and real-time monitoring of the network utilization, makes decisions and issues control commands for the arriving SFC requests, based on which it occupies the underlying resources held by the virtualized network function (VNF). The simulation results show that, compared with the static timing resource allocation algorithm, the active control resource deployment algorithm proposed in the article has better performance in terms of resource utilization, acceptance rate, and user creation utility.

Keywords: SDN; NFV; service function chain; resource allocation

1 Introduction

In order to adapt to the explosive growth of mobile data traffic and a large number of new applications and business needs, operators have deployed large-scale network infrastructure to provide users with ubiquitous network services. At the same time, different access networks such as wired, cellular, and WLAN coexist in the network space around us. On the one hand, huge traffic and heterogeneous characteristics bring difficulties to network management. On the other hand, as the available bandwidth of the network increases, the performance requirements of users for network

services are becoming more and more acute. How to design a solution that can satisfy users' needs for network-perceived network quality [1] under the condition of considering indifferent heterogeneous network connection and rational distribution of heterogeneous network resources has become a key issue in optimizing the utilization of network resources.

Nowadays, SDN and NFV have been widely used in the research of network resource orchestration. With the help of SDN technology, network equipment has changed from a closed mode to an open general equipment mode, which separates the control plane and the data plane, which allows the network to be programmed through open interfaces [2]; With the help of NFV technology, network element functions can be deployed in the form of software on the general server node of the infrastructure, so that network element functions and special hardware devices are separated [3,12]. In [4], the author focuses on the SDN-based 5G cellular network resource allocation solution, describes the methods to solve different resource requirements, and conducts an in-depth analysis and overview of different resource allocation schemes. The author in [5] proposed a GSO-RBFDM model based on dynamic resource pricing to solve the problem of virtual network resource allocation. While realizing dynamic resource allocation, it also optimizes acceptance rate, network cost and link pressure. In [6], the author designed a virtualized resource management framework called GreenVolP based on SDN and NFV orchestration. By managing the number of network devices such as VolP servers and switches, design algorithms that effectively prevent VolP network overloads while minimizing the use of switches and other devices in the network.

At present, the research on network resource mapping and orchestration algorithms for the integration of SDN and NFV is mainly focused on a certain network resource, and the optimization direction is mostly to optimize the overall network resource utilization efficiency and reduce network costs [13,14]. However, in reality we are faced with complex heterogeneous network resources and user groups with different needs and preferences. It has become an inevitable trend to bring users a better perceived service quality. Therefore, the focus of this article is: in the long-term operation process, while ensuring the consumption of physical network resources, as much as possible to meet the needs and preferences of the user side, so as to achieve the goal of maximizing the quality of service for the user by the network.

This paper combines SDN and NFV technologies to further abstract the network, which can form an end-to-end and logically isolated network, and realize a user-centric framework. A user agent corresponding to each user can be generated on a virtual device with connection information retention and user situation awareness capabilities. Through the user agent, the ubiquitous and undifferentiated network connection of users in a heterogeneous network can be realized. User preference feature perception, through flexible resource scheduling management to ensure the maximum utility provided to users. The main contributions of this paper are summarized as follows:

(1) Based on the three-tier architecture composed of SDN controller, network function virtualization, and physical underlying computing resource layer in the process of heterogeneous network resource mapping [15], in view of the resource allocation problem of service function chain deployment, A user-centered framework for maximizing user utility is proposed. Modeling through multi-criteria utility theory transforms the optimization problem of resource scheduling into a problem of maximizing user utility.
(2) Propose a dynamic active control resource deployment algorithm based on request priority and user preference, including two stages: network state prediction and control based on load utilization. The controller makes corresponding decisions on the SFC request and issues control commands based on the real-time monitoring of the utilization rate predicted by the

future state. Based on this, the VNF occupies the resources held by the physical nodes to achieve the purpose of active control and maximum utility.

(3) A series of simulation experiments are used to prove the effectiveness of the proposed algorithm. The simulation results show that in a congested network scenario, the acceptance rate of high-priority requests and the overall user utility benefit are significantly improved. At the same time, this method does not reduce the utilization rate of the network, and ensures the efficiency of resource utilization.

2 System Model

2.1 Underlying Network Model

Under the business of virtual network resource allocation problem, different users access corresponding service providers (SP) to obtain network services according to business needs. SP constructs corresponding virtual network requests according to business needs, and then based on infrastructure providers (InP) physical network resource constraints, using resource allocation algorithms to allocate suitable physical networks to carry the virtual network [3]. In the network virtualization business system, users can access different service providers according to their own needs and enjoy customized services. When deploying virtual network requests, the SP selects an appropriate InP based on available physical resources, service type, and lease cost. Sometimes it is even necessary to lease the underlying physical network resources of multiple InPs at the same time. The performance of the virtual network resource allocation algorithm directly affects the resource utilization rate of the underlying physical network, the operational effect of network services and the user experience.

The underlying infrastructure network is represented by an undirected graph G = (N, L), where N represents the set of physical nodes, L represents the link between nodes, and node n ($n \in N$) and node p ($p \in N$) The link between is expressed as $l_{n,p}$, and the upper limit of the link bandwidth resource is expressed as $B_{n,p}$. In the virtual mapping process, after the request arrives, according to the available resources of the physical network, a reasonable physical node is selected to perform the mapping of computing resources until the mapping is successful and the physical resources occupied by it are released.

After the underlying physical network is virtualized, it is constructed by multiple physical nodes that can provide physical resources. These physical nodes provide VNFs with various types of resources such as CPU, memory, and disk. At this time, each VNF in the service function chain can be deployed on any physical node, and one physical node can provide services for multiple VNFs located in the same service function chain [3].

The amount of data processed by each VNF determines its own resource requirements. In order to reduce the complexity of the algorithm, this model unifies the multiple resources (CPU, memory, disk) that can be provided by general-purpose processor nodes into computing resources. The virtual nodes in different virtual requests can be mapped to the same physical node. For each node n(n ∈ N), its available resources are:

$$C_N(n) = c(n) - \sum_{\forall n^v \to n} c(n^v) + \sum_{\forall n^v \to n} Rel(c(n^v)) \tag{1}$$

c (n) represents the total amount of computing resources that the node n can provide, and the second term $\sum_{\forall n^v \to n} c(n^v)$ represents the sum of computing power of all virtual nodes n^v mapped to physical node n, the value will only change after the request is accepted and the actual computing resources are allocated. Rel(c (n^v)) in the third item represents the resources released from the virtual node n^v, and the

change of this value depends on whether the corresponding NFV life cycle arrives or not. Therefore, the dynamic change status of the available resource $C_N(n)$ can be captured by the controller in real time, and the next decision can be made based on this.

2.2 SFC Deployment Model

The underlying network can provide VNFs of different protocol layer types. Let $F = \{f_p | p = 1, 2, 3, \ldots P\}$ represent the set of VNFs, where $p = type(f_p)$ represents the type of VNF [9]. It should be noted that each node may not provide all types of VNFs. Therefore, it is assumed that each VNF type has a set of nodes to be deployed [8]. Binary variable $\gamma_{n,p}$ indicates whether node n can deploy $VNFf_p \in F$:

$$\gamma_{n,p} = \begin{cases} 1, & If\ VNFf_p\ can\ be\ deployed\ on\ node\ n \in N \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The SFC request is composed of multiple VNFs, ingress and egress nodes with order constraints [11,12]. Fig. 1 shows the deployment diagram of the service function chain. Let $S = \{s_k | k = 1, 2, 3, \ldots K\}$ represent the set of SFC requests, $s_k = < i_k, e_k, \varphi_k, \tau_k >$ represents the SFC request, where $i_k, e_k \in N$ represent the entry and exit nodes, respectively, and τ_k represents the life cycle of the SFC (the occupancy time of resources).

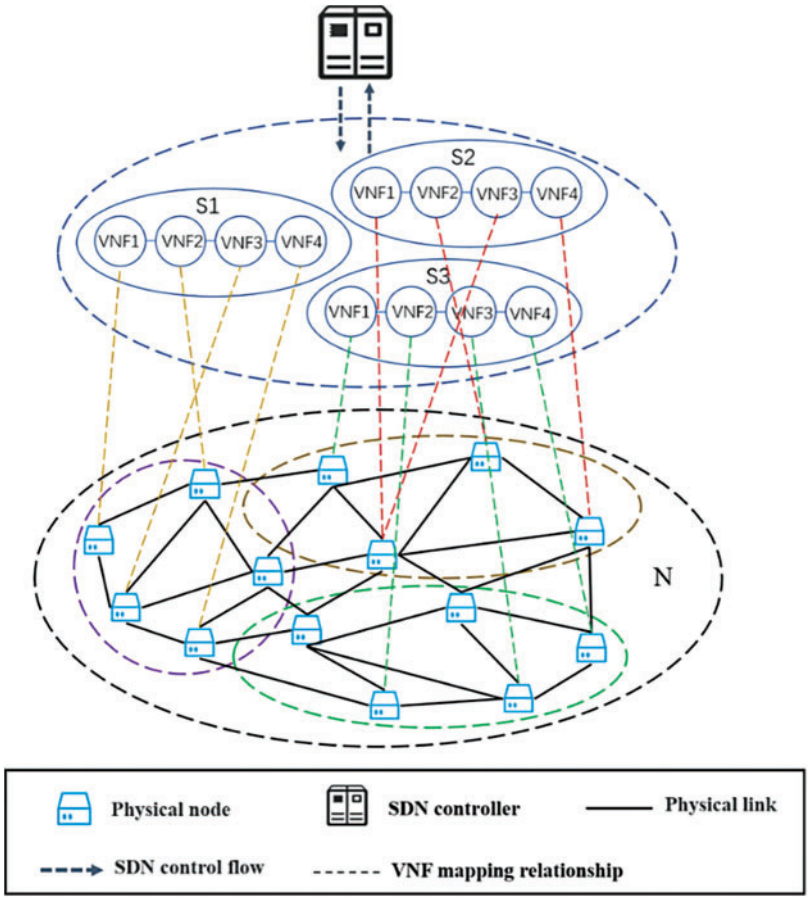

Figure 1: Schematic diagram of service function chain deployment

When the SFC service time exceeds the life cycle, the service ends and the allocated resources are reclaimed [10]. $\varphi_k = \{f_m^k | m = 1, 2, 3, \ldots M^k\}$ represents the ordered VNF sequence of SFC, where $f_m^k \in F$. Finally, define the binary variable $\eta_{m,n}^k \in \{0, 1\}$ to represent the mapping relationship between VNF and physical nodes [12]. If f_m^k is deployed on node n, then $\eta_{m,n}^k = 1$; otherwise, $\eta_{m,n}^k = 0$.

$$\eta_{m,n}^k = \begin{cases} 1, & If\ VNFf_p\ is\ deployed\ on\ node\ n \in N \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Therefore, the number of VNFs deployed in the k-th service function chain can be expressed as:

$$W = \sum_{f_m^k \in k} \sum_{n \in N} \eta_{m,n}^k \tag{4}$$

For the service function chain $s_k \in S$, the data processing volume of the VNF f_m^k composing the SFC is expressed as R_m^k, the f_m^k computing resource demand and its data processing volume are defined as a linear relationship, and the correlation coefficient (data processing demand factor) is expressed by α_m, then the computing resource demand v_m^k of the VNF can be expressed as [13]:

$$v_m^k = \alpha_m^k R_m^k \tag{5}$$

2.3 Optimization Objective

In microeconomics, utility refers to the ability of goods or services to meet human needs. A related term is utility function, which relates to the utility that consumers obtain from goods or services. Different consumers with different user preferences will have different utility values for the same product. The basic idea of the utility theory is that the decision maker will always choose the decision with the greatest utility value.

In a heterogeneous network, for the user side, the utility function is defined as $U(\boldsymbol{x})$, an aggregated multi-criteria utility algorithm based on multiple standards:

$$U(\boldsymbol{x}) = \sum_{i=1}^{n} w_i u_i(x_i), \sum_{i=1}^{n} w_i = 1 \tag{6}$$

where $\boldsymbol{x}$ is the vector of n standards under consideration, w_i is the corresponding weight, $u_i(x_i)$ is the basic utility of standard i, so $U(\boldsymbol{x})$ comprehensively considers n attributes that users care about. This format provides a simple and understandable way to aggregate different basic utilities. But it has some serious limitations that have been studied in [16], so a new form of utility function is proposed in conjunction with the discussion of related issues:

$$U(x) = \prod_{i=1}^{n} [u_i(x_i)]^{w_i}, \sum_{i=1}^{n} w_i = 1 \tag{7}$$

For the single-criteria utility function, select the Sigmoid function:

$$u(x) = \frac{1}{1 + e^{k(x_m - x)}} \tag{8}$$

The utility function used in the early research work is the logarithmic utility function, which is a good approximation of the utility value of the wired communication network. In recent years, the demand for wireless adaptive real-time applications has continued to increase, and the utility function that approximates real-time applications is the sigmoid function [17]. The concave utility function is only suitable for traditional data service modeling, and does not capture the characteristics of the increasingly popular audio and video services on the Internet. That is, corresponding to the

common services that are sensitive to delay and rate, such as streaming video and audio services, reducing the transmission data rate below a certain threshold will cause a significant drop in experience (For example, below a certain bit rate, the quality of audio communication will drop sharply) [18]. Therefore, in order to effectively simulate the changes in user utility value, it is reasonable to use the Sigmoid utility function to model.

Obviously, the value of $u(x)$ is between 0 and 1, the center of the function curve is $(x_m, 0.5)$, the curve can be moved to the left or right by adjusting x_m, and the steepness of the curve can be controlled by the gradient coefficient k, So it can simulate the user's sensitivity to network changes. On the left side of x_m, the curve is convex, and on the right side, the curve is concave. That is to say, when the criterion value is small, as the criterion value increases, the utility value will increase rapidly; and when the criterion value is greater than x_m, as it increases, the utility value will increase slowly.

It can be defined as follows:

$$1) \quad when x \le x_m: \qquad u(x) = \frac{1}{1 + e^{k_1 (x_m - x)}}, k_1 > 0 \tag{9}$$

$$2) \quad when x > x_m: \qquad u(x) = \frac{1}{1 + e^{k_2 (x_m - x)}}, k_2 > 0 \tag{10}$$

The upper limit $x_\alpha (0 < x_\alpha < x_m)$ and the lower limit $x_\beta (x_\beta > x_m)$ need to be set. $u(x_\alpha) = u_l$, u_l is a floating-point number close to 0 [16]. $u\left(x_\beta\right) = u_h$, u_h is a floating-point number close to 1. Based on this, the values of k_1 and k_2 can be calculated:

$$k_1 = \frac{\ln\left(\frac{1-u_l}{u_l}\right)}{x_m - x_\alpha}; \quad k_2 = \frac{\ln\left(\frac{1-u_h}{u_h}\right)}{x_m - x_\beta} \tag{11}$$

As for the total benefits provided by the allocation of SFC computing resources for a single user to the user, it is necessary to consider the priority of the user's request and the pickiness of resources to calculate the benefits. Intuitively, high-priority has a larger slope and lower fault tolerance, and a low-priority has a smaller slope and higher fault tolerance. The criteria can be grouped into two types. The timeliness of acceptance of VNF requests and resource allocation in the service function chain can be regarded as a criterion. Whether the user's criticality of the accessed network resources is met in a timely manner can also be regarded as another criterion. Therefore, it is stipulated that if the request is accepted within the time unit, or the resource criticality of a single user's SFC access is satisfied (for example, the preference of cellular network access for security considerations, or the preference for WIFI resources based on cost performance), then the corresponding criterion obtains the maximum benefit, that is, the independent variable of $u(x)$ can take the value x_β.

Otherwise, the corresponding benefit will be lost according to the amount of elapsed unit time, that is, $x = x_\beta - c\left(t_{f_m^k} - t^0_{f_m^k}\right)$ in $u(x)$. In the formula, $t^0_{f_m^k}$ is the initial time when the VNFR request is initiated, and $t_{f_m^k}$ is the time when the request is accepted (the actual mapping resource is allocated).

The above $U(x)$ is a measure of the user benefit result produced by the internal algorithm of the controller for a single SFC. The user benefits generated by a certain VNF within the SFC need to be expressed separately by $u_i(x_i)$, according to the above $\varphi_k = \{f_m^k | m = 1, 2, 3, \ldots M^k\}$, i can be sequentially selected according to the number m of VNF sets of the SFC.

$$U_{s_k} = \prod_{i=1}^{M^k} [u_i(x_i)]^{w_i} \tag{12}$$

Therefore, if the long-term realized total user benefit is calculated, it is necessary to perform cumulative calculation to allocate the total benefit of k SFC computing resources.

$$U_a = \sum_{s_k \in S} U_{s_k} \tag{13}$$

The goal of this article is to fully consider user preferences and the priority of requests for SFC resource allocation. Calculate user request priority and resource criticality according to user characteristics to obtain utility benefits, thereby establishing a user-centric architecture. Therefore, the optimization goal of this paper is modeled as:

$$\begin{aligned}
&\max U_a \\
&\text{s.t.} \\
&\boldsymbol{C}1: \sum_{i=1}^{n} w_i = 1 \\
&\boldsymbol{C}2: \sum_{s_k \in S} \sum_{f_m^k \in \varphi_k} \eta_{m,n}^k v_{m,n}^k \le C_N(n), \forall n \in N \\
&\boldsymbol{C}3: \sum_{fp \in F} \gamma_{n,p} \ge 1 \\
&\boldsymbol{C}4: \sum_{n \in N} \gamma_{n,p} \ge 1 \\
&\boldsymbol{C}5: \sum_{s_k \in S} \sum_{f_m^k \in \varphi_k} \eta_{m,n}^k \eta_{m+1,p}^k v_m^k \le B_{n,p}, n, p \in N, \ n \ne p
\end{aligned} \tag{14}$$

C1 is the user utility calculation coefficient constraint. The calculation of the multi-standard aggregate utility U_{s_k} of each SFC is constrained so that the sum of the coefficients of each single standard is 1; C2 indicates that the computing resources occupied by all VNFs deployed on a node must not be exceeds the total amount of computing resources that this node can provide. Since the binary variable $\gamma_{n,p}$ indicates whether the node can deploy $VNFf_p \in F$, C3 ensures that each node can support at least one type of VNF for deployment, and C4 ensures that every type of $VNFf_p \in F$ can be deployed. C5 indicates that the bandwidth resources of each link in the network will not exceed the upper limit.

3 Algorithm Design

3.1 Process Description

The algorithm in this paper fully considers the decision factors including priority and resource preference when the controller receives the SFC request to allocate the actual resources for each VNF component, and determines the resource allocation plan within the time slice. In this way, a control instruction is issued to the user agent generated at the target node. The whole process is mainly composed of three parts: status collection, decision-making and issuing instructions.

(1) Before making any decision, the controller will collect the underlying network status information, calculate and update the available resources of each node, and collect user preferences.
(2) Then, the controller will execute the decision-making component, which determines the virtual resource allocation plan within each time slice.
(3) Once the allocation plan is determined, the controller will issue a control instruction to the user agent, and the user agent will occupy the underlying virtual resources held by the VNF according to the control instruction of the controller.

3.2 Algorithm Implementation

The input of the algorithm includes: the received VNFR, user characteristics (high and low priority) and resource preference; thus output: whether to accept the request and allocate the actual computing resources, including the selected physical nodes and the resources provided by them. During this period, monitor the information on the successful occupation of node resources, and update the network status in time; monitor the information on the release of node resources in the network, and when the service is terminated, the allocated resources occupied by the service function chain are recovered.

Active control resource deployment based on request priority and user preference is the main focus of this algorithm. Assuming that the controller knows the average arrival rate of SFC requests, it can predict the next state of the network by considering the arrival probability of the newly added SFC in the next time unit. Then, according to the prediction of the next state, the resource utilization rate of the next time unit is calculated. Based on the comparison between this value and the network utilization threshold, the algorithm selects and decides whether to temporarily not process a low-priority request. Therefore, the active control resource deployment algorithm proposed in this paper includes two stages: network state prediction and control based on load utilization. The flow chart of the SFC deployment algorithm is shown in Fig. 2.

1) Network state prediction: At this stage, the controller estimates the next state of the network according to the arrival probability of the upcoming request in the next time unit [7]. The next state of the network is predicted as:

$$\tilde{S}_{T+1} = \rho S_{T+1} + (1 - \rho) S_T \tag{15}$$

where ρ represents the probability of arriving at a new request in the next time unit, and the first term on the right side of the equation represents the arrival of a new request, and the needed resources are allocated according to the average amount of requested computing resources. The second term on the right side of the equation represents that the network status does not change: the probability that there is no request in the next time unit (1-ρ). The left side represents the predicted next state (next time unit) of the underlying network.

2) Control based on load utilization: Calculate the resource $\sum_{n \in N} C_N(n)$ according to the predicted next state $\tilde{S}_{T+1}$, and calculate the network resource utilization as:

$$U = 1 - \frac{\sum_{n \in N} C_N(n)}{\sum_{n \in N} c(n)} \tag{16}$$

In the formula, U is the ratio of network allocation resources to total resources. If the percentage of network utilization is greater than the pre-selected threshold, it is possible that in the next state, the upcoming high-priority VN request will be rejected. Therefore, as will be mentioned below, in this case, the proposed active control resource deployment algorithm will actively not process a low-priority VNFR temporarily.

Since the algorithm solves the problem of VNF deployment and orchestration under dynamic changes in business requirements, the value will only change when the request is accepted and the actual computing resources are allocated. The essence of dynamic resource changes is the arrival of new requests and the release of allocated resources. When new requests arrive in each time slot, the deployment decision of the algorithm will be triggered. The dynamic changes of available resources can be captured by the controller in real time, based on which the next decision can be made.

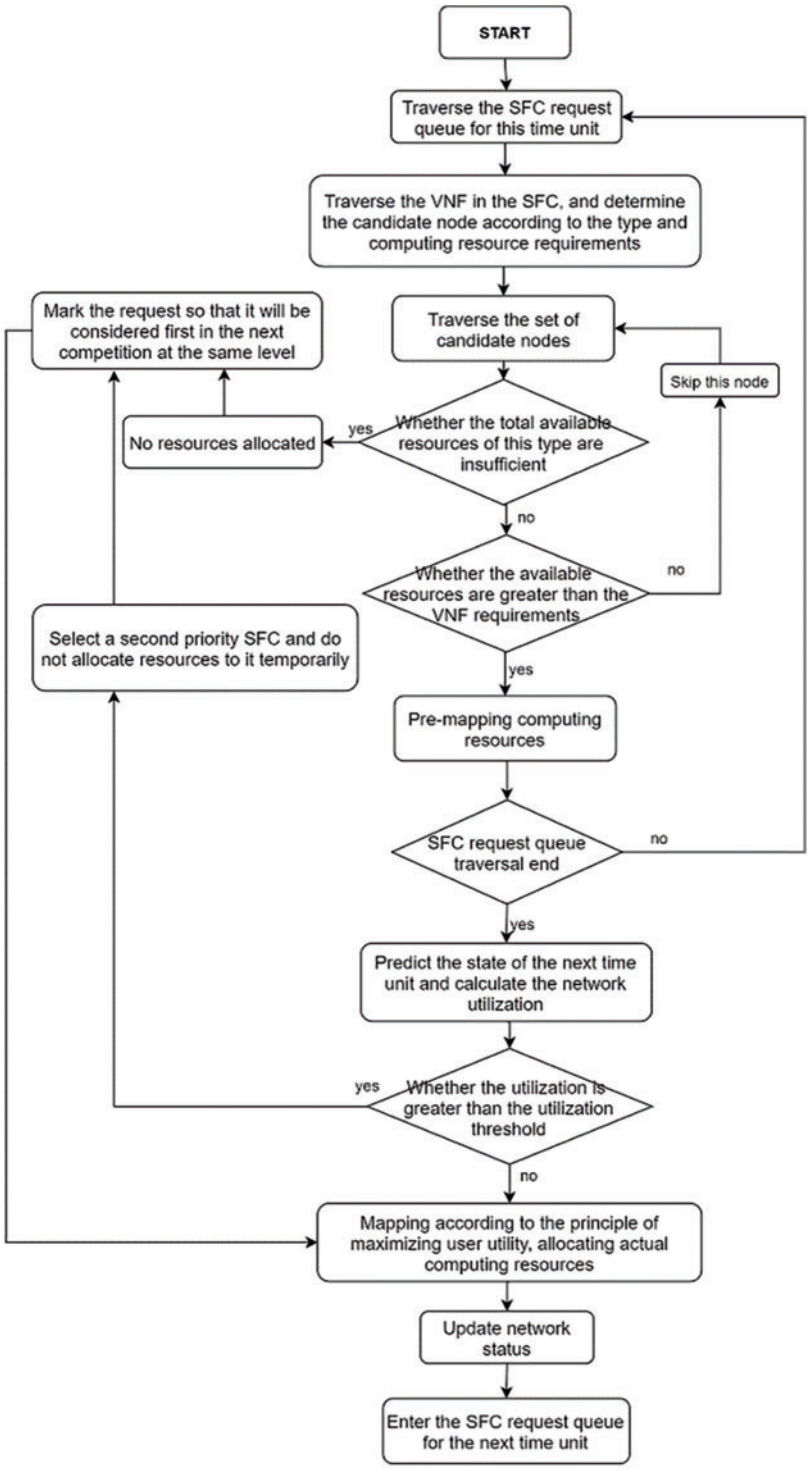

Figure 2: Algorithm flow chart

After receiving a new request in this time slot, the algorithm will sequentially traverse the SFC request queue within the time unit and traverse the VNFs in the service chain in order. By evaluating the types of resources required by each VNF, a series of candidate nodes that can be deployed are selected. If the remaining available resources of the candidate node are sufficient to provide the mapping requirements of the VNF, the calculation resource pre-mapping is performed, that is, the process of assuming that the request is accepted and calculating the possible network resource utilization rate in the future. If the prediction result shows that even if all the requests are accepted, the network resource utilization rate of the next time slot will not exceed the threshold, then the actual resource allocation is performed. If it is found that the future utilization rate may exceed the preset threshold, select a low-priority SFC to temporarily not allocate available resources for it, and mark the request so that it will be prioritized in the next competition of the same level. This is to prevent the same or several requests from being discarded consecutively, which will not meet expectations. After the resource allocation is over, the algorithm updates the network state and starts to process the SFC request of the next time slot, and starts to traverse the newly arrived request queue.

4 Simulation Results and Analysis

In order to evaluate the method proposed in this article, we chose to use Python 3.6.9 for simulation and run on an Ubuntu 18.04 virtual machine with 4.0GB RAM and a 2-core processor. As shown in formula 17, the simulation assumes that the arrival rate of VNF requests follows the Poisson distribution P [19], and the values of λ for different priority requests are different, and the selection of numerical values was based on relevant studies [19,20] and the network scale simulated in this paper. The average arrival rate of high-priority requests is λ_h, and the average arrival rate of low-priority requests is λ_l. Choose to estimate the next state of the network based on the Poisson distribution arrival rate of high-priority VN requests.

$$P(X=i)=\binom{n}{i}\left(\frac{\lambda}{n}\right)^{i}\left(1-\frac{\lambda}{n}\right)^{n-i}=\frac{e^{-\lambda}\lambda^{i}}{i!} \tag{17}$$

And all the simulation data are shown in Tab. 1 as follows:

Table 1: Simulation data

Parameter	Value
Number of nodes	12
Types of VNFs that can be deployed	2
Computing resources provided by the node	U (300, 500)
Number of VNF types	3
VNF resource requirements	U (1, 8)
Data processing demand factor	U (0.8, 1.2)
SFC life cycle	180
High priority request arrival rate	P ($\lambda_l = 1.0$)
Low priority request arrival rate	P ($\lambda_h = 2.0$)
Total time	1500

In order to verify the performance of the proposed active control resource deployment algorithm, the following four performance indicators are considered:

(1) Acceptance rate of high-priority VNF requests: According to the purpose of resource management in this article, in order to increase the number of high-priority VN requests accepted in the network, we calculated the ratio of accepted high-priority VN requests to total high-priority VN requests.
(2) Resource utilization: The average ratio of the sum of the used resources of the network to the total available resources is measured.
(3) User utility: The algorithm in this paper aims to meet the needs and preferences of users as much as possible, so as to maximize the network's service quality for users. The measurement of user utility can intuitively reflect the improvement of the algorithm to the user's perception quality and the advantages of the algorithm.
(4) Utility load ratio: In order to better evaluate the utilization value of the algorithm to the unit load, the ratio between utility and load intensity is calculated and measured in real time.

Fig. 3 shows the comparison between a network that uses active control resource deployment algorithm and a network that passively accepts requests for resource allocation. In the case of active control, the increase in the acceptance rate of high-priority requests can approach 1, and it has stabilized over time. As for the network that passively accepts requests for resource allocation, the degree of anti-congestion is poor, and can only be maintained between 0.8–1.0. This significant improvement confirms the effectiveness of the method based on active admission control and greatly improves the high-priority admission rate of the network.

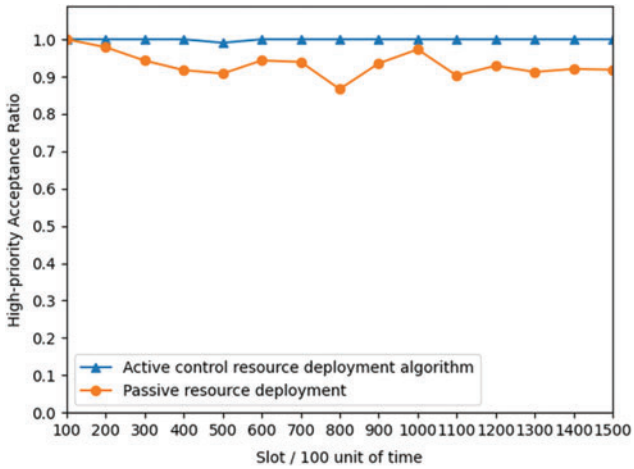

Figure 3: Comparison of acceptance rate of high-priority requests

As shown in Fig. 4, the average ratio of the sum of the used resources of the network to the total available resources in the time unit is measured. Comparing the two utilization curves, it can be seen that the utilization of network resources using the active control algorithm is close to the situation of passive network resource allocation. That is to say, the algorithm proposed in this paper can effectively use network resources, it will not waste resources because of the emphasis on user-centered framework, and it will not reduce the effective utilization of resources because of the resource reservation in the prediction mapping stage.

Fig. 5 shows the user utility value using the active control resource deployment algorithm and passively accepting requests to allocate resources. Calculating the long-term total user benefits requires cumulative calculation to allocate the total revenue of k SFC computing resources. As shown in the figure, it is not difficult to find that the user utility value created by the former is significantly higher than that of the latter, and the gap is more significant over time. It shows that in the long-term operation process, under the condition of ensuring the consumption of physical network resources, the algorithm can meet the needs and preferences of the user side as much as possible, so as to achieve the purpose of maximizing the quality of service of the network to the user.

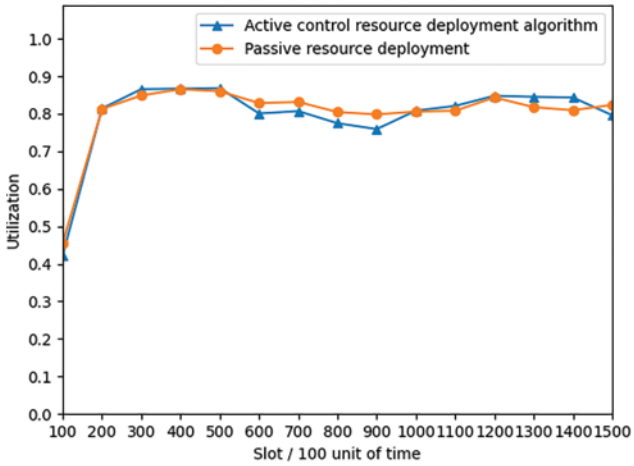

Figure 4: Comparison of network resource utilization

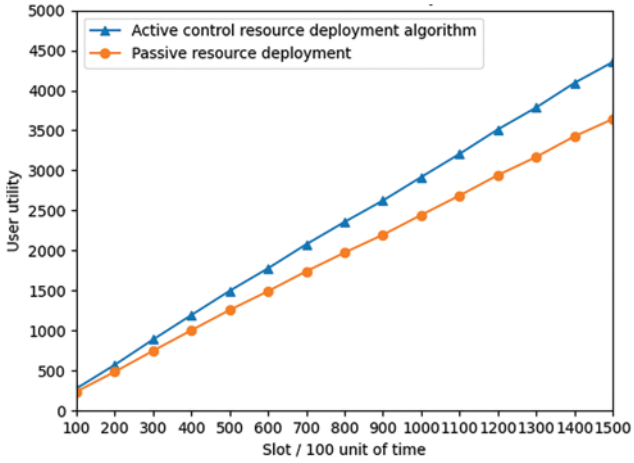

Figure 5: Comparison of created user utility

Fig. 6 shows the change trend of the average utility load ratio over time under the two schemes, the ratio between utility and load intensity can better evaluate the utilization value of the algorithm for unit load. It can be seen that compared with passive allocation, the proposed algorithm has great advantages in improving the utilization value of unit load.

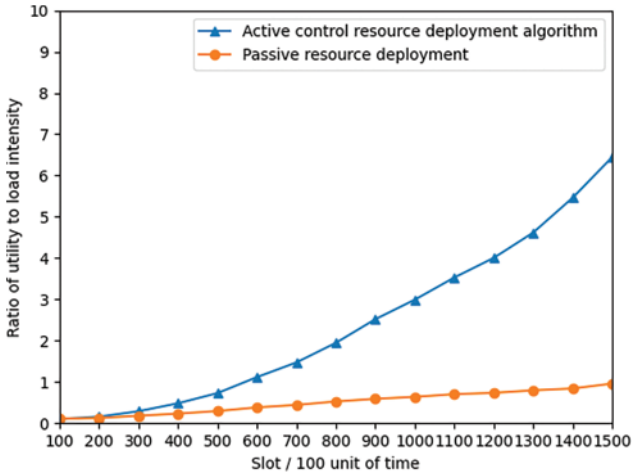

Figure 6: Comparison of utility and load intensity ratio

5 Conclusion

Based on SDN and NFV technology, this paper abstracts the network to form an end-to-end and logically isolated network, realizes a user-centric framework, and proposes an active control resource deployment algorithm that maximizes user utility. In the process from receiving the SFC request to allocating actual resources for each VNF component, the controller takes the lead to consider the SFC demand on a user basis, and fully considers decision factors including priority, resource preference, and network prediction. Through flexible resource scheduling management to ensure the maximum utility provided to users. The simulation results show that, compared with the strategy of passively accepting VNF requests and allocating resources, our proposed solution performs better, can use network resources more effectively, and create greater value for network users.

Acknowledgement: This work was supported by the National Natural Science Foundation of China (61871058).

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Z. Lu, T. Lei and X. Wen, "SDN based user-centric framework for heterogeneous wireless networks," *Mobile Information Systems*, vol. 1, no. 1, pp. 1–9, 2016.

[2] Y. X. Zhao, Y. Y. Chen, R. L. Jian and L. Q. Yang, "A resource allocation scheme for SDN-based 5G ultra-dense heterogeneous networks," in *Proc. 2017 IEEE Globecom Workshops (GC Wkshps)*, Singapore, pp. 1–6, 2017.

[3] A. A. Barakabitze, A. Ahmad and R. Mijumbi, "5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges," *Computer Networks*, vol. 167, no. 1, pp. 106984, 2020.

[4] S. K. Tayyaba and M. A. Shah, "Resource allocation in SDN based 5G cellular networks," *Peer-to-Peer Networking and Applications*, vol. 12, no. 1, pp. 514–538, 2019.

[5] X. C. Xiao, X. W. Zheng, Y. Wei and X. C. Cui, "A virtual network resource allocation model based on dynamic resource pricing," *IEEE Access*, vol. 8, no. 3, pp. 160414–160426, 2020.

[6] A. Montazerolghaem, M. H. Yaghmaee and A. Leon-Garcia, "Green cloud multimedia networking: NFV/SDN based energy-efficient resource allocation," *IEEE Transactions on Green Communications and Networking*, vol. 4, no. 3, pp. 873–889, 2020.

[7] S. Shakeri, S. Parsaeefard and M. Derakhshani, "Proactive admission control and dynamic resource management in SDN-based virtualized networks," in *Proc. 2017 8th Int. Conf. on the Network of the Future (NOF)*, London, UK, pp. 46–51, 2017.

[8] S. Mehraghdam, M. Keller and H. Karl, "Specifying and placing chains of virtual network functions," in *Proc. 2014 IEEE 3rd Int. Conf. on Cloud Networking (CloudNet)*, Luxembourg, Luxembourg, 2014.

[9] L. Qu, C. Assi, K. Shaban and M. J. Khabbaz, "A reliability-aware network service chain provisioning with delay guarantees in NFV-enabled enterprise datacenter networks," *IEEE Transactions on Network and Service Management*, vol. 14, no. 3, pp. 554–568, 2017.

[10] W. R. Ding, H. F. Yu and S. X. Luo, "Enhancing the reliability of services in NFV with the cost-efficient redundancy scheme," in *Proc. 2017 IEEE Int. Conf. on Communications (ICC)*, Paris, France, pp. 1–6, 2017.

[11] Y. Kanizo, O. Rottenstreich, I. Segall and J. Yallouz, "Optimizing virtual backup allocation for middleboxes," *IEEE/ACM Transactions on Networking*, vol. 25, no. 5, pp. 2759–2772, 2017.

[12] S. Yang, F. Li, S. Trajanovski, R. Yahyapour and X. M. Fu, "Recent advances of resource allocation in network function virtualization," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 2, pp. 295–314, 2020.

[13] A. A. Z. Ibrahim, F. Hashim, N. K. Noordin, A. Sali, K. Navaie *et al.,* "Heuristic resource allocation algorithm for controller placement in multi-control 5G based on SDN/NFV architecture," *IEEE Access*, vol. 9, no. 1, pp. 2602–2617, 2020.

[14] H. Li, L. H. Wang, X. M. Wen, Z. M. Lu and J. Y. Li, "MSV: An algorithm for coordinated resource allocation in network function virtualization," *IEEE Access*, vol. 6, no. 1, pp. 76876–76888, 2018.

[15] N. N. Ma, J. Zhang and T. Huang, "A model based on genetic algorithm for service chain resource allocation in NFV," in *Proc. 2017 3rd IEEE Int. Conf. on Computer and Communications (ICCC)*, Chengdu, China, pp. 607–611, 2017.

[16] Q. T. Nguyen-Vuong, N. Agoulmine, E. H. Cherkaoui and L. Toni, "Multicriteria optimization of access selection to improve the quality of experience in heterogeneous wireless access networks," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 4, pp. 1785–1800, 2012.

[17] A. Abdel-Hadi and C. Clancy, "A utility proportional fairness approach for resource allocation in 4G-LTE," in *Proc. 2014 Int. Conf. on Computing, Networking and Communications (ICNC)*, Honolulu, HI, USA, 2014.

[18] J. W. Lee, R. R. Mazumdar and N. B. Shroff, "Non-convex optimization and rate control for multi-class services in the internet," *IEEE/ACM Transactions on Networking*, vol. 13, no. 4, pp. 827–840, 2005.

[19] S. Stepanov and M. Stepanov, "Modeling of open flow-based SDN node with taking into account the differences of serving TCP and UDP traffic streams," in *Proc. 24th Conf. of Open Innovations Association (FRUCT)*, Moscow, Russia, pp. 415–421, 2019.

[20] K. Sood, S. Yu and Y. Xiang, "Performance analysis of software-defined network switch using *M*/*Geo*/1 model," *IEEE Communications Letters*, vol. 20, no. 12, pp. 2522–2525, 2016.