

Cost-and-QoS-Based NFV Service Function Chain Mapping Mechanism

Yimin Li

State Key Laboratory of
Networking and Switching
Technology
Beijing University of Posts and
Telecommunications
Beijing, China
ymli@bupt.edu.cn

Lifang Gao

State Grid Hebei Information &
Telecommunication Branch
Shijiazhuang, China
manyupiaoling@126.com

Siya Xu

State Key Laboratory of
Networking and Switching
Technology
Beijing University of Posts and
Telecommunications
Beijing, China
xusiyaxsy@bupt.edu.cn

Qinghai Ou

Beijing Fibrlink Communications
Co., Ltd
State Grid Information &
Telecommunication Group
Beijing, China
ouqinghai@sgitg.sgcc.com.cn

Xinyu Yuan

Tianjin University
Tianjin, China
2530406567@qq.com

Feng Qi

State Key Laboratory of
Networking and Switching
Technology
Beijing University of Posts and
Telecommunications
Beijing, China
qifeng@bupt.edu.cn

Shaoyong Guo

State Key Laboratory of
Networking and Switching
Technology
Beijing University of Posts and
Telecommunications
Beijing, China
syguo@bupt.edu.cn

Xuesong Qiu

State Key Laboratory of
Networking and Switching
Technology
Beijing University of Posts and
Telecommunications
Beijing, China
xsqiu@bupt.edu.cn

Abstract—Network Function Virtualization (NFV) technology decouples network functions from the proprietary hardware by using generalized equipment and software, which lowers the cost of network operator. However, the existing mapping mechanisms in NFV environment can't optimize the cost of deployment and improve the rationality of network resource allocation while ensuring the basic service quality requirements of users. To solve the problem, a mathematical model which looks on the assurance of quality of service and cost optimization is established in this article. The model aims at maximizing the total revenue from service chain deployment in resource-constrained network, and takes the resource demand, end-to-end delay requirement and reliability requirement of the service request as the basic constraints. Furthermore, a greedy algorithm of service chain mapping named GA+LCB is proposed to solve the problem. Simulation results show that compared with other algorithms, GA+LCB can effectively improve the success rate of receiving service requests, reduce the cost in the deployment process and achieve higher deployment benefits while ensuring the QoS requirements.

Keywords—Network Function Virtualization; Service Chaining Mapping; Cost; Quality of Service

I. INTRODUCTION

With the rapid development of the Internet, the user number of information services has increased sharply, and users' requirements for Quality of Service (QoS) have become higher and more differentiated. Users' main service needs have also been transformed from traditional voice services to data communication services such as video and game with high real-time requirements. However, the tight coupling of hardware and software in the traditional communication network leads to

shortcomings such as management difficulties, high operating costs, and poor flexibility [1], making it difficult to meet the users' needs.

With the development of virtualization and cloud computing technologies and the improvements of performance of devices such as generic servers, NFV technology arises. NFV is a next-generation network solution jointly proposed by several major global telecom operators, it integrates various network functions into the generated devices of the underlying network through software and virtualization technology. NFV provides end-to-end network services in the form of a service chain. The service chain defines a specific sequence of virtual network function (VNF) sets, and users' traffic needs to pass through these network function nodes in order.

Fig.1 shows an example of a service chain mapping to the underlying physical network.

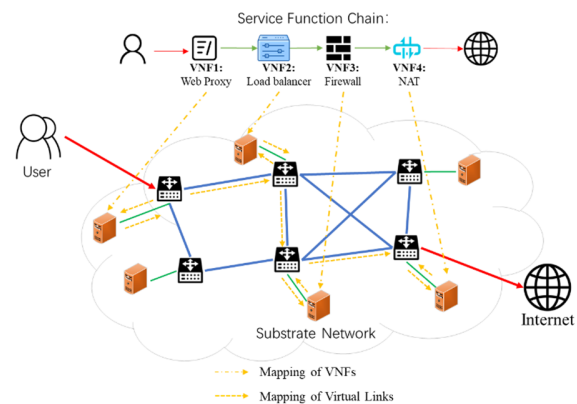


Fig. 1. An example of service chain mapping

Each VNF in the service chain is mapped to different server nodes in the underlying network, and the virtual links connecting each VNF are mapped to the physical links connecting each server node.

Service function chain can be applied in the data center network and the mobile network to perform a series of specific treatment on the data traffic in the network, such as performing stateful inspection and identification of traffic, distributing the traffic for efficient resource utilization, optimizing the use of bandwidth, recognizing and eliminating the security threats, etc. It is of great significance to study the service chain mapping mechanism to ensure the efficient operation of NFV network.

However, most of the current mapping algorithms can only consider one or two aspects of the cost of deployment, the reliability, the end-to-end delay of the service chain, and the rationality of resource allocation. It is difficult for them to achieve multi-objective optimization in the mapping in order to improve the performance of the network, guarantee the quality of service of users and the interests of network operators. In order to solve this problem, this paper has carried out the following contribution:

1. This paper establishes a QoS-and-cost-based service chain mapping model with multi-objective optimization. The model takes into account both network operators' and users' interests and the performance of resource-constrained networks.

2. For the above model, this paper proposes a greedy service chain mapping algorithm GA+LCB (Genetic Algorithm + Least Cost Backup). GA+LCB divides the mapping problem of service chain into two parts: primary mapping and redundant backup.

3. Our genetic algorithm used in the primary mapping presents a load balancing mechanism based on the importance of nodes. It can evaluate the importance of nodes according to the node's reliability, the node's available resources and other factors. As a result, it can reserve more residual resources on the nodes with higher importance in order to deal with the redundant backup and the subsequent service request processing.

4. In the redundant backup, a cost-based backup algorithm which jointly considers the advantages of single-path routing and multi-path routing is adopted to reduce the cost of bandwidth. Moreover, when selecting VNF for backup, we will calculate the ratio of the increasable range of the service chain's reliability to the resources required by the VNF in order to improve the efficiency of the backup.

II. RELATED WORKS

Researchers at home and abroad have conducted a series of researches on mapping of service chain, and some achievements have been made. When designing the mapping algorithm of service chain, many researchers take accounts the cost of deployment, the reliability and end-to-end delay of service chain, the acceptance rate of service request, the rationality and effectiveness of resource allocation and other factors

How to map service chain at low cost is a direction discussed by many researches [2-6]. Among them, [2] proposed a heuristic algorithm aiming at minimizing the resource cost of service

chain deployment in elastic optical network, which can intelligently balance the use of IT and spectrum resources according to the setting of resource price in multi-domain environment. [3] aimed to minimize the sum of energy consumption and reconstruction cost of service chain deployment. This paper argued that when carrying out service chain migration, combining VNF and reducing the number of servers can reduce energy consumption, but the migration of VNF will cause data loss, reduce service quality and result in loss of revenue, which needs to be balanced.

The requirement of high reliability for service requests also challenges the mapping of the service chain, so many studies have also explored in the field of reliability [7-10]. The authors of [7] constructed a framework to comprehensively evaluate the reliability of NFV service chain deployment, however, they did not further study how to use this framework for service chain deployment. [8] analyzed the advantages and disadvantages of using single-path routing and multi-path routing in redundant backup, and proposed a heuristic algorithm that uses hybrid routing strategy in backup to minimize the bandwidth cost while meeting the service delay requirements and reliability requirements.

With the increasing demand for response delays, many studies have focused on reducing the delay of the service chain [11-14]. Among them, [12] proposed a neural network-based service chain mapping algorithm that uses the neural network model to predict the traffic flow trend of the service chain in future time, and uses delay optimization algorithm to realize the pre-deployment of the service chain, so as to dynamically optimize the delay of the service chain.

Due to the limited resources of the network, many studies are devoted to designing reasonable resource scheduling and allocation strategies to improve the utilization rate of resources and the reception rate of requests [15-17]. [15] posed a service chain deployment strategy based on resource splitting. Its idea is to reasonably split VNF and virtual links according to the resource allocation of the network, so as to reduce the density of resource allocation and improve the utilization rate of resources. [16] proposed a measurement index named RFD to quantify the resource fragmentation in the network, and designed a heuristic algorithm aiming at minimize the generation of resource fragment to solve the deployment problem.

Each of the above studies provides solutions to the problem of mapping the service chain from different angles. However, the optimization goals of their algorithms are relatively simple, which makes it difficult to improve the performance of resource-constrained networks while ensure both network operators' and users' interests. Therefore, this paper proposes a greedy algorithm of service chain mapping based on cost optimization and QoS assurance, in order to achieve multi-objective optimization in service chain mapping.

III. PROBLEM MODEL

A. Physical Network Model

In this paper, the physical network is represented by a weighted undirected graph $G = (N, L)$. Where N represents the set of physical nodes, each node n_i has certain computational

resources, the total amount of resources is CT_i , and the remaining amount is CR_i . Each node has a certain reliability A_i ($0 < A_i < 1$), and can execute several types of virtual network functions V_i ; L is the set of physical links connecting nodes. Each link $l_i = l_{j,k}$ ($n_j, n_k \in N$) has certain bandwidth resources, the total amount of resources is BT_i , and the remaining amount is BR_i . Each link also has a certain transmission delay D_i . In order to simplify the complexity of the problem, this paper considers that the end-to-end delay of the service chain mainly depends on the transmission delay of the link, while other delays such as processing delay are ignored.

B. Network Service Request model

This paper uses S to represent the set of users' network service requests. Each request s_i in the set has a certain delay requirement d_i , a certain amount of bandwidth requirement b_i and reliability requirement R_i . Each request has a selected entry node o_i ($o_i \in N$) and an exit node t_i ($t_i \in N$) in the network. The service chain of each request consists of a certain number and sequence of VNF. Each VNF $f_{i,j}$ can realize a specific type of network function $v_{i,j}$, and operating the function needs to consume a certain amount of computational resource $r_{i,j}$. The reliability of VNF after instantiation is $a_{i,j}$. After receiving the service request successfully, a certain amount of revenue INC_i can be obtained, and the higher the QoS the user requests and the more resources the user needs, the higher the INC_i will be.

C. Mapping Scheme Model

In this paper, the achievement of required reliability is realized through the adoption of redundancy-based VNF backup placement policies. To indicate the placement of VNF and its backup respectively, variables h and p are introduced. That is:

$$h_{i,j}^k = \begin{cases} 1, \text{the } j^{\text{th}} \text{ VNF } f_{i,j} \text{ of } s_i \text{ is hosted on } n_k \\ 0, \text{otherwise} \end{cases} \quad (1)$$

$$p_{i,j}^k = \begin{cases} 1, \text{the backup of VNF } f_{i,j} \text{ is hosted on } n_k \\ 0, \text{otherwise} \end{cases} \quad (2)$$

In order to indicate whether the VNF adopts backup (each primary VNF may have only one backup) and its backup route, variables ub and bw are introduced respectively (Multi-path backup means that the node selected by the primary VNF is not on the same path as the node selected by the backup VNF).

$$ub_{i,j} = \begin{cases} 1, \text{the } j^{\text{th}} \text{ VNF } f_{i,j} \text{ of } s_i \text{ uses a backup.} \\ 0, \text{otherwise} \end{cases} \quad (3)$$

$$bw_{i,j} = \begin{cases} 1, \text{VNF } f_{i,j} \text{ uses multi-path backup} \\ 0, \text{VNF } f_{i,j} \text{ uses single-path backup} \end{cases} \quad (4)$$

This paper also introduces a variable tr to indicate that the node only acts as a traffic forwarding role for the service chain, and does not carry the VNF, that is:

$$tr_{i,j}^k = \begin{cases} 1, n_k \text{ only acts as a forwarding role for } s_i \\ 0, \text{otherwise} \end{cases} \quad (5)$$

To represent the mapping of links, variables sp and mp are introduced respectively such that:

$$sp_i^{j,k} = \begin{cases} 1, l_{j,k} \text{ is selected by } s_i \text{'s main path} \\ 0, \text{otherwise.} \end{cases} \quad (6)$$

$$mp_i^{j,k} = \begin{cases} 1, l_{j,k} \text{ is selected by } s_i \text{'s branch path} \\ 0, \text{otherwise.} \end{cases} \quad (7)$$

D. Reliability Model

This paper argues that the reliability of service chain is mainly affected by hardware factor. After each VNF is hosted on a node, the hardware reliability of the node is obtained by: $a_{i,j} = h_{i,j}^k \times A_k$.

If a VNF uses a backup, its reliability will be:

$$a_{i,j} = 1 - (1 - A_m) \times (1 - A_n), \text{ if } h_{i,j}^m = p_{i,j}^n = 1 \quad (8)$$

Calculation of the reliability of a service chain RA needs to be discussed in two cases.

If each VNF of a chain is placed on different nodes, the reliability of the chain will be:

$$RA_i = \prod_{j=1}^{|s_i|} a_{i,j} \quad (9)$$

If there are two adjacent VNFs hosted on the same node. Then, considering the reliability of two VNF comprehensively, their reliability should be:

$$a_{i,j} \times a_{i,j+1} = A_k, \text{ if } h_{i,j}^k = h_{i,j+1}^k \quad (10)$$

E. Cost Model

To simplify the problem, this paper considers that the deployment cost of the service chain mainly comes from the resource overhead, while energy consumption and other costs are ignored. The cost of resources is the sum of the expense using computing resources and bandwidth resources. The cost of s_i can be calculated as follows:

$$\begin{aligned} COS_i = & EX_c \cdot \sum_{k=1}^{|N|} \sum_{j=1}^{|s_i|} (h_{i,j}^k + p_{i,j}^k) \cdot r_{i,j} \\ & + EX_b \cdot \sum_{m=1}^{|N|} \sum_{n=1}^{|N|} (sp_i^{m,n} + mp_i^{m,n}) \cdot b_i \end{aligned} \quad (11)$$

Where EX_c represents the cost per unit of computational resources and EX_b represents the cost per unit of bandwidth resources.

F. Mapping Constraints

Service chain's mapping needs to follow many constraints. Firstly, the mapping of VNF and virtual links should follow resource constraints, that is, their required resources should not exceed the total amount of resources of nodes and links. Moreover, the type of network function realized by the VNF placed on the node needs to be a type that the node can execute:

$$\sum_{i=1}^{|S|} \sum_{j=1}^{|S|} (h_{i,j}^k + p_{i,j}^k) \times r_{i,j} \leq CT_k, \forall n_k \in N \quad (12)$$

$$\sum_{i=1}^{|S|} (sp_i^{j,k} + sp_i^{k,j} + mp_i^{j,k} + mp_i^{k,j}) b_j \leq BT_{j,k}, \forall l_{j,k} \in L \quad (13)$$

$$v_{i,j} \in V_k, \text{ if } h_{(i,j)}^k = 1 \text{ or } p_{(i,j)}^k = 1, \forall s_i \in S, n_k \in N \quad (14)$$

In order to prevent the VNFs of the same service chain from being placed too centrally on a node, we stipulate that the number of placements should not exceed two. At the same time, each VNF cannot be placed on the same node as its backup. Finally, each primary VNF or backup VNF can only be placed on one node:

$$\sum_{j=1}^{|S|} (h_{i,j}^k + p_{i,j}^k) \leq 2, \forall s_i \in S, n_k \in N \quad (15)$$

$$(h_{i,j}^k + p_{i,j}^k) \leq 1, \forall f_{i,j} \in S_i, s_i \in S, n_k \in N \quad (16)$$

$$\sum_{k=1}^{|N|} h_{i,j}^k \leq 1, \sum_{k=1}^{|N|} p_{i,j}^k \leq 1, \forall f_{i,j} \in S_i, s_i \in S \quad (17)$$

In addition, in order to avoid loops in the path of service chain mapping, nodes are required to only serve as nodes forwarding flow or nodes carrying VNF:

$$tr_i^k + \min \left(\sum_{j=1}^{|S|} (h_{i,j}^k + p_{i,j}^k), 1 \right) \leq 1, \forall s_i \in S, n_k \in N \quad (18)$$

For each requests s_i , in order to ensure that the traffic passes through the VNFs in a predetermined order, we numbers all the nodes (from the entry node to the exit node) according to their order in the main path, which is denoted by an index $I_{i,k} (n_k \in N)$. For any VNF of s_i , it is necessary to ensure that the index of the node hosting it is no larger than the index of node that hosts a VNF after it, that is:

$$I_{i,p} \leq I_{i,q}, \text{ if } h_{i,j}^p = h_{i,k}^q = 1 \text{ and } j < k, \forall f_{i,j}, f_{i,k} \in S_i \quad (19)$$

Finally, QoS needs of users should be guaranteed, that is, to meet the requirements of end-to-end delay and reliability:

$$\sum_{j=1}^{|N|} \sum_{k=1}^{|N|} sp_{i,j}^k \times D_{j,k} \leq d_i, \forall s_i \in S \quad (20)$$

$$RA_i \geq R_i, \forall s_i \in S \quad (21)$$

G. Optimization Objective

The optimization objective of this model is as follows:

$$\max g = \max \sum_{i \in S_{SUC}} (INC_i - COS_i) \quad (22)$$

Where S_{SUC} represents the set of service requests that are received successfully. If the success rate of receiving service requests is higher, deployment cost of each service request is lower, g will be larger. Therefore, it can optimize the request receive rate and deployment cost in a balanced way.

To sum up, the optimization problem model of service chain deployment in this paper is shown as follows

optimization objective: $\max g$

constraints: (12)~(21)

IV. ALGORITHM DESIGN

A. Algorithm Overview

The mapping problem of the service chain is a typical NP-hard problem. Therefore, this paper proposes a heuristic greedy algorithm GA+LCB to solve the problem. The overall process of the algorithm is shown in the Fig.2.

The goal of the algorithm is to reasonably schedule and allocate the limited network resources, reduce the deployment cost of operators and improve the reception rate of requests so as to obtain higher deployment benefits. At the same time, the algorithm will ensure the basic service quality requirements and resource requirements of the service.

In this paper, the mapping problem of service chain is divided into two parts: primary mapping and redundant backup. Before mapping, the algorithm will comprehensively sort the service chains according to their delay requirements, revenue and estimated cost, and then the algorithm will process each service chain in turn; In the primary mapping, the algorithm will first use the K shortest path algorithm to get the shortest path set between the entry node and the exit node of the service chain, and the paths will be sorted according to the number of links and the amount of remaining resources on the path to reduce the bandwidth cost and reduce the load on the network; After getting the initial path, the algorithm will place the VNFs on the path based on a genetic algorithm; If the reliability of the service chain after the primary mapping does not meet the requirements, redundant policy is required. In the backup, the algorithm will adopt a cost-based backup algorithm to reduce the cost of backup.

Next, the VNF placement algorithm and backup algorithm used in the GA+LCB algorithm will be introduced in detail.

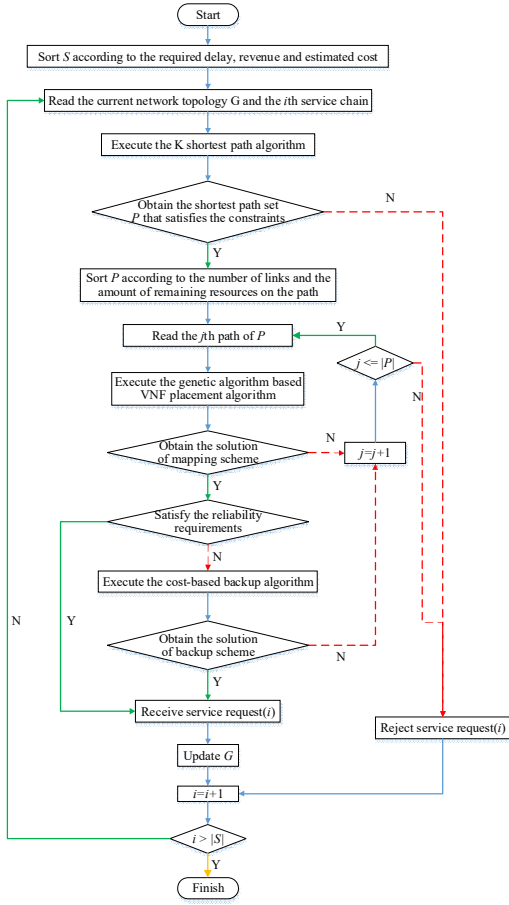


Fig. 2. GA+LCB algorithm

B. Genetic Algorithm Based VNF Placement Algorithm

When an initial path is given, the placement algorithm should place each primary VNF reasonably on multiple nodes on the path to make reasonable utilization of resources on the nodes.

This paper defines an indicator E to comprehensively evaluate the performance of placement schemes:

$$E = a \times \text{nor} \left(|N_R| \times \sum_{k \in N_R} CR_k \right) + b \times \text{nor}(lba) + c \times (1/2 (\text{nor}(\min\{RA_i, R_i\}) + \text{nor}(lba))) \quad (23)$$

Where a, b and c are the scale factors whose sum is 1; nor means dimensionless processing of data; N_R is the set of nodes without resource fragmentation, that is, the nodes whose remaining computing resources are not less than the minimum value of resources required by each VNF and the number of available links (the link whose remaining bandwidth resources are not less than the minimum bandwidth requirement of each service request) is greater than zero:

$$n_k \in N_R, \text{ if } CR_k \geq \min(r_{i,j})$$

$$\text{and } da_k = \sum_{m=1}^{|N|} \min \left\{ 1, \frac{BR_{(k,m)}}{\min(b_i)} \right\} \geq 1 \quad (24)$$

lba is an indicator to measure the degree of load balance, which is defined by the following three formulas:

$$lba = \frac{1}{N_R} \sum_{k \in N_R} \left(\frac{CR_k}{imp_k} - \text{avarage} \left(\frac{CR_k}{imp_k} \right) \right)^2 \quad (25)$$

$$imp_k = w_1 \times \text{nor}(da_k) + w_2 \times \text{nor}(dc_k) + w_3 \times \text{nor}(A_k) + w_4 \times \text{nor}(|V_k|), \quad (26)$$

$$w_1 + w_2 + w_3 + w_4 = 1$$

$$dc_k = \sum_{m=1}^{|N|} \min \left\{ 1, \frac{BR_{k,m}}{\min(b_i)} \right\} \times BR_{k,m} \quad (27)$$

The load balancing in this paper is based on the node's importance, the imp is an index used to measure the importance of nodes. In this paper, the importance of nodes is measured by four attributes: the number of available links (da_k), the total available bandwidth of linked links (dc_k), reliability (A_k) and the number of types of network functions the node can carry ($|V_k|$). Nodes of higher importance are more likely to be selected in the backup and appear in the initial path, so the burden on these nodes needs to be reduced (proportionally) to facilitate subsequent processing.

Finally, lba is an index that can measure the rationality of the reliability distribution:

$$lba = \frac{1}{|s_i|} \sum_{j=1}^{|s_i|} \left(\frac{a_{i,j}}{r_{i,j} + \theta} - \text{avarage} \left(\frac{a_{i,j}}{r_{i,j} + \theta} \right) \right) \quad (28)$$

Where θ is a correction factor. The meaning of the above formula is that when placing VNF, we hope that the VNF with more required resources can be placed on nodes with higher reliability to reduce the computational resource cost of backup, since the subsequent backup algorithm tends to select the backup VNF with lower reliability.

In conclusion, the optimization objective of the placement algorithm is: $\max E$.

The placement algorithm will adjust the scale factor according to the current network resources. When the resources are sufficient, the weight of lba in E will be appropriately increased. When resources are tight, the algorithm will pay more attention to resource fragmentation and load balancing.

The genetic algorithm in this paper will use a digital symbol string $\{m_{i,1}, m_{i,2}, \dots, m_{i,|s_i|}\}$ to encode the chromosomes. Each chromosome represents a placement scheme, where $m_{i,j}$

represents the index of the node(selected by the j th VNF of s_i) in the available nodes.

The specific process of placement algorithm is shown in TABLE I.

TABLE I. GENETIC ALGORITHM BASED VNF PLACEMENT ALGORITHM

Input: $G(N, L), \{r_{i,1}, r_{i,2}, \dots, r_{i,s_i}\}$, the k th shortest path $P_{i,k}$; Output: placement scheme $M_{best} = \{m_{i,1}, m_{i,2}, \dots, m_{i,s_i}\}$;
Steps: 1. Check each node on the k th shortest path $P_{i,k}$ to see whether it satisfies the condition $CR_k \geq \min(r_{i,k})$, if so, add it to the available node set N_A . 2. Adjust the sizes of a, b and c based on network resources. 3. Generate the initial population $Group(0)$ with n individuals satisfying the constraint conditions. 4. while $i < i_{max}$ and $t < t_{max}$ do 5. for each individual $g_j \in Group(i)$ do 6. Calculate the fitness e_j ; 7. Record individual with highest fitness as $M_{best}(i)$; 8. end for 9. if $M_{best}(i) = M_{best}(i-1)$ then Set $t = t + 1$; 10. else Set $t = 0$; 11. end if 12. Use roulette algorithms to generate the next generation population $Group(i+1)$; Use $M_{best}(i)$ to replace the individual with the lowest fitness of $Group(i+1)$; 13. for ($j = 0; j < n/2; j++$) do 14. Let g_j and $g_{j+n/2}$ cross at a random point with probability P_C and add the new individuals produced by the hybridization to $Group(i+1)$; 15. end for 16. for each individual $g_j \in Group(i+1)$ do 17. Let g_j mutate with probability P_M ; 18. if g_j does not satisfy the constraint then Delete g_j ; 19. end if 20. end for 21. end while 22. Return M_{best} ;

C. Cost-Based Backup Algorithm

The goal of our backup algorithm is to reduce the resource consumption of backup while making the reliability of service chain meet the requirement.

Our backup algorithm will comprehensively consider the advantages of single-path routing and multi-path routing. Wherein, single-path routing is to extend or change part of the link of the initial path, and complete the placement of the backup VNF on the changed or extended part, which has relatively less bandwidth cost, but will bring a larger extra delay; Multi-path routing is to map an additional path between specific nodes and place the backup VNF on the branch path, that is, the primary VNF and backup VNF are parallel in traffic. In this way, the additional delay is smaller, but the bandwidth cost is relatively high [8]. Most of the existing backup algorithm ([9],[10]) prefers using multi-path routing while single-path routing is always ignored.

The comparison between single-path routing and multi-path routing can be intuitively illustrated by a set of figures (Figure 6). In the case where the reliability of the primary mapping does not meet the requirements, whether the backup is performed using single-path routing or multi-path routing, the reliability can meet the basic requirement (0.95); With single-path routing,

backups consume less bandwidth, while using multi-path routing for backup, the extra delay introduced by backup is smaller, but at the expense of more bandwidth cost.

In order to reduce the consumption of bandwidth resources, we will prioritize single-path routing in redundant backup. Multi-path routing will be taken only if the single-path routing cannot satisfy the delay requirement.

In addition, unlike most existing backup algorithms, when selecting a VNF for backup, we will not simply select the VNF with the lowest reliability. We will comprehensively consider the increasable range of service chain's reliability and the compute resources needed by the VNF, and select the VNF with the maximum ratio of former to latter for backup so as to improve the efficiency of backup. Moreover, when deciding the distribution of VNF's backup along single/multi-path, the node with fewer links and more remaining resources will be chosen for backup placement in order to release the burden of network.

The specific steps of our backup algorithm are shown in TABLE II.

TABLE II. COST-BASED BACKUP ALGORITHM

Input: $G(N, L), \{r_{i,1}, r_{i,2}, \dots, r_{i,s_i}\}, b_i, d_i, R_i, RA_i$, initial placement scheme $M_{best} = \{m_{i,1}, m_{i,2}, \dots, m_{i,s_i}\}$, the k th shortest path $P_{i,k}$; Output: backup scheme $B_{best} = \{b_{i,1}, b_{i,2}, \dots, b_{i,s_i}\}$;
Steps: 1. while $RA_i < R_i$ do 2. Set the selection factor $a_{max} = 0$, the label of VNF being selected for backup $sel = -1$; 3. for $f_{i,j} \in s_i$ do 4. Calculate the ratio of the increasable range of the service chain's reliability to the resource required by $f_{i,j}$ when selecting $f_{i,j}$ for backup; 5. if $a_{i,j} > a_{max}$ and $f_{i,j}$ has not been selected for backup 6. then Set $a_{max} = a_{i,j}$; Set $sel = j$; 7. end if 8. end for 9. if a node that satisfies the constraint condition and has the largest remaining resource can be found on the mapping path between $f_{i,sel-1}$ and $f_{i,sel+1}$; 10. then Place the backup; Update G ; continue ; 11. end if 12. Perform the single-path routing based backup strategy: a. Get the shortest path set between $f_{i,sel}$ and $f_{i,sel+1}$; b. Traverse the path set, maintain a priority queue Q based on the number of path-based links (the less the better) and the remaining resources of the node (the more the better); c. if Q is not empty then Pop up the vertex of Q and place the backup on it; Update the mapping path; Update G ; continue ; end if d. Perform similar steps(a~c) between $f_{i,sel}$ and $f_{i,sel-1}$; 13. Perform the multi-path routing based backup strategy: a. Get the shortest path set between $f_{i,sel-1}$ and $f_{i,sel+1}$; b. Traverse the path set, maintaining a priority queue Q based on the number of path-based links and the remaining resources of the node; c. if Q is not empty then Pop up the vertex of Q and place the backup on it; Update the mapping path; Update G ; end if 14. end while 15. Return B_{best} ;

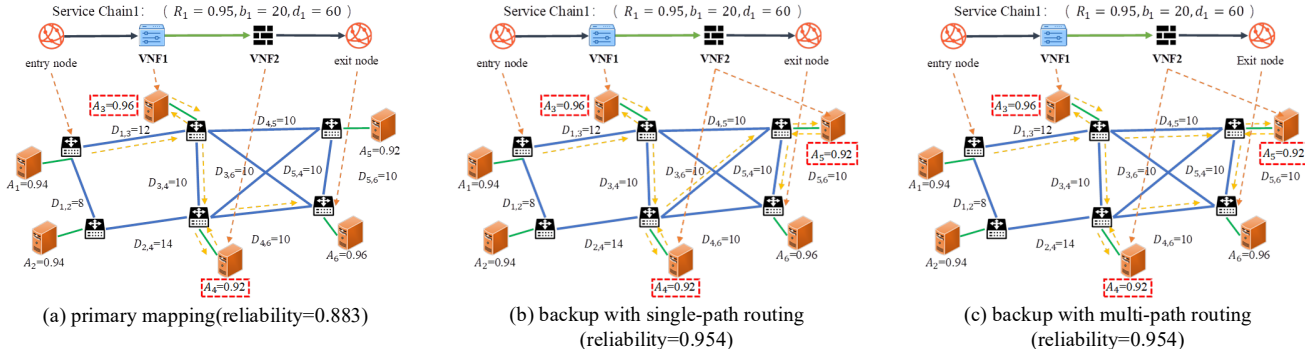


Fig. 3. Illustrative comparison between singlepath routing and multi-path routing

V. SIMULATION

We will use MyEclipse and MATLAB to verify the algorithm. Specific parameters of the simulated network topology and service chain are shown in TABLE III

TABLE III. PARAMETER SETTINGS OF THE SIMULATION

	Parameter	Value
Physical Network	Number of Physical Nodes	100
	Number of Physical Links	200
	Per Unit of Computation Resources' Cost	1
	Per Unit of Bandwidth Resources' Cost	0.2
	Computational Resources of Each Node	50~150
	Reliability of Each Node	0.9~0.99
	Number of Types of Network Function Use by The Services in S	6, Represented by {a, b, c, d, e, f} Respectively
	Number of Types of Network Functions That Each Node Can Execute	2~6, Selected among {a, b, c, d, e, f}
	Bandwidth Resources of Each Link	250~500
	Delay of Each Link (ms)	Selected from Two Ranges [5, 15] and [10, 20]
Service Request	Number of Service Requests	100~500
	Entry Node (Index)	Selected among {1,2,11,81,91,92}
	Exit Node (Index)	Selected among {9,10,20,90,99,100}
	Number of VNFs of Each Service Chain	2~6
	Delay Requirement	Selected from Three Ranges [60,100], [100,140] and [140,180]
	Bandwidth Requirement	Selected among {15,20,25}
	Reliability Requirement	Selected among {0.95,0.96,0.97,0.98,0.99}
	Computational Resources Required by Each VNF	Selected among {2,3,4,5,6}
	Type of Network Function That Each VNF Can Realize	Selected among {a, b, c, d, e, f}
	Revenue from Deployment	30~150, Depends on The Resource Requirements and QoS Requirements of Service Request

In order to verify the effectiveness of the algorithm, three algorithms widely used as comparison objects ([8]-[10], [17]) are selected in this paper for comparison.

The first comparison algorithm FF+LCB (First-Fit + Least Cost Backup) adopts the First-Fit algorithm for initial VNF deployment. Its basic idea is to select the first deployable node for each VNF on the shortest path from the entry node to the exit node, and the backup algorithm is the same as GA+LCB.

The second comparison algorithm, LLG+LCB (Least Load Greedy + Least Cost Backup), will use the Least Load Greedy algorithm for initial VNF deployment. It prefers to select the node with the most remaining resources for each VNF on the initial path, and the backup algorithm is the same as GA+LCB

The third comparison Algorithm, GA+LDB (Genetic Algorithm + Least Delay Backup), will only take the multi-path routing strategy in backup, and in the selection of nodes for backup placement, it will select the node with the smallest delay of the path, so as to reduce the delay of the Backup.

In order to verify the effectiveness of the algorithm, the simulation will use the following indicators:

1. The total benefit of the deployment, which is the optimization objective of our model, will be used as the primary indicator;
2. Number of received service requests;
3. The average resource cost per VNF, which is the total resource cost of the deployment divided by the number of VNFs of the successfully mapped service chain;

Next, we will analyze the performance of each algorithm.

Firstly, we 'd like to talk about the time complexity of our algorithm. The time complexity of GA+LCB mainly depends on the VNF placement parts. The complexity of our GA is $O(MG(N+V))$ (N is the number of nodes on the path, V is the amount of VNF in the service chain, M is the number of iteration of GA and G is the size of GAs population) while LLG's complexity is $O(NV)$ and FF's complexity is $O(N)$, but GA can get a better placement scheme. Compared with the enumeration method, Our GA can solve the MILP problem in polynomial time, and the solution is not far from the optimal solution in performance.

Secondly, we compare the deployment benefits of each algorithm, as shown in Fig.4. It can be seen from the figure that GA+LCB algorithm designed in this paper can always obtain the highest deployment benefit. With the increase of the number

of requests, GA+LCB's advantage becomes more and more obvious. When the number of requests reaches 500, GA+LCB outperforms the second-best algorithm by 17%. This is because GA+LCB fully considers the rationality and effectiveness of resource allocation and the cost overhead in the service chain mapping, so it has achieved better comprehensive performance. The worst performance of FF+LCB is due to its simple strategy used in the initial VNF placement.

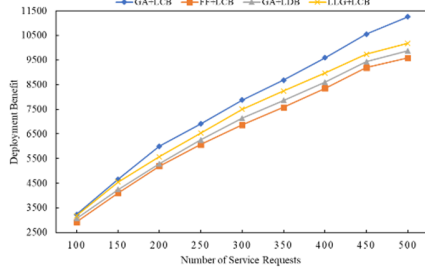


Fig. 4. Deployment Benefit vs Number of Service Requests

Next, we compare the number of received service requests, as shown in Fig.5. the advantage of GA+LCB is not obvious in the case of fewer requests, since the resources in the network are still sufficient. With the increase of requests' number, the advantages of GA+LCB gradually increase, since it fully considers the resource fragmentation issues and load balancing in primary mapping, and tries its best to reduce the network's load when performing the backup. LLG+LCB performs second best since it locally optimizes the network load problem during the primary mapping; GA+LDB and FF+LC perform the worst, because the former consuming a lot of bandwidth resources in backup aggravates the load of the links, while the latter doesn't consider the load of the network or resource fragmentation in primary mapping.

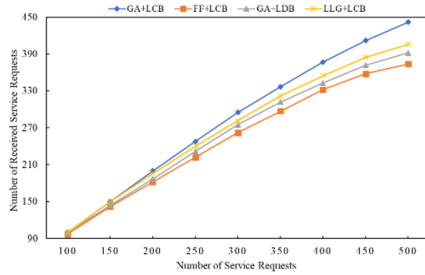


Fig. 5. Number of Achieved Requests vs Number of Service Requests

Fig.6 shows the performance of each algorithm in computational cost: GA+LCB and GA+LDB behave better than the other two algorithms because they try to improve the reliability of the primary mapping and the rationality of the reliability distribution, thus reducing the computational resources that may be consumed by backup. As the number of service requests increases, their advantage decreases gradually. This is because when resources are tight, the placement algorithm will reduce the proportion of reliability factor in fitness function so as to pay more attention to the influence of resource fragmentation and load balancing.

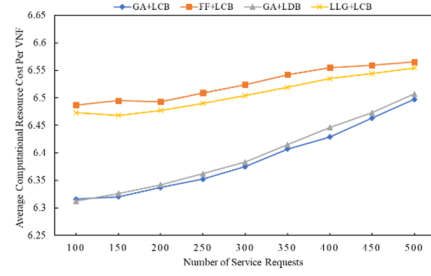


Fig. 6. Computational Cost vs Number of Service Requests

Finally, we compare the bandwidth cost of each algorithm, as shown in Fig.7: One can see that GA+LCB performs best, and LLG+LCB takes the second; The latter consumes more bandwidth resources when the number of service requests increases; GA+LDB and FF+LCB performed the worst; Because the former takes the multi-path routing with higher bandwidth cost during the backup, while the latter's strategy in the primary mapping increases the network's burden, improving the probability of successful deployment of service chain on the later selected paths(the path with more links), which increases the bandwidth cost.

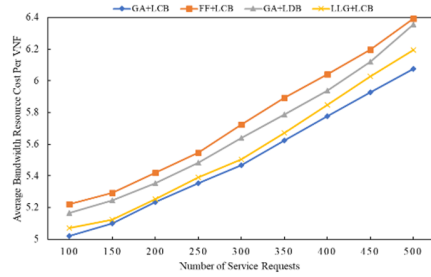


Fig. 7. Bandwidth Cost vs Number of Service Requests

VI. CONCLUSION

In order to reasonably and effectively allocate the limited physical resources in NFV network, and to ensure the basic QoS requirements of users and the interests of network operators, we propose a mathematical model oriented to cost optimization and QoS assurance. Furthermore, a greedy algorithm for service chain mapping GA+LCB is designed to solve the problem. GA+LCB uses the genetic algorithm to solve the VNF placement scheme in the primary mapping. In the case that the reliability of the service chain after primary mapping does not meet the requirements, a cost-based backup algorithm is taken to make redundant backup of the service chain. Simulation results show that GA+LCB algorithm outperforms other algorithms by about 17% in terms of deployment benefit.

ACKNOWLEDGMENT

This work is supported by the Science and Technology Project of State Grid Corporation of China (SGHEXT00YJJS1900 050) Fundamental Research Funds for the Central Universities under Grant 2019RC07.

REFERENCES

- [1] Weilin Zhou, Yuan Yang, Mingwei Xu, "Network Function Virtualization Technology Research," *Journal of Computer Research and Development*, vol. 55, no. 4, pp. 675-688, Apr. 2018.
- [2] Y. Wang, P. Lu, W. Lu and Z. Zhu, "Cost-Efficient Virtual Network Function Graph (vNFG) Provisioning in Multidomain Elastic Optical Networks," in *Journal of Lightwave Technology*, vol. 35, no. 13, pp. 2712-2723, 1 July 1, 2017.
- [3] V. Eramo, E. Miucci, M. Ammar and F. G. Lavacca, "An Approach for Service Function Chain Routing and Virtual Function Network Instance Migration in Network Function Virtualization Architectures," in *IEEE/ACM Transactions on Networking*, vol. 25, no. 4, pp. 2008-2025, Aug. 2017.
- [4] M. Zeng, W. Fang and Z. Zhu, "Orchestrating Tree-Type VNF Forwarding Graphs in Inter-DC Elastic Optical Networks," in *Journal of Lightwave Technology*, vol. 34, no. 14, pp. 3330-3341, 15 July 15, 2016.
- [5] J. Wu, Z. Zhang, Y. Hong and Y. Wen, "Cloud radio access network (C-RAN): a primer," in *IEEE Network*, vol. 29, no. 1, pp. 35-41, Jan.-Feb. 2015.
- [6] N. Huin, B. Jaumard and F. Giroire, "Optimal Network Service Chain Provisioning," in *IEEE/ACM Transactions on Networking*, vol. 26, no. 3, pp. 1320-1333, June 2018.
- [7] J. Liu, Z. Jiang, N. Kato, O. Akashi and A. Takahara, "Reliability evaluation for NFV deployment of future mobile broadband networks," in *IEEE Wireless Communications*, vol. 23, no. 3, pp. 90-96, June 2016.
- [8] 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," in *IEEE Transactions on Network and Service Management*, vol. 14, no. 3, pp. 554-568, Sept. 2017.
- [9] J. Fan, et al., "GREP: Guaranteeing reliability with enhanced protection in NFV," in *Proc. ACM SIGCOMM Workshop HotMiddleBox*, London, U.K., Aug. 2015, pp. 13-18.
- [10] W. Ding, H. Yu and S. Luo, "Enhancing the reliability of services in NFV with the cost-efficient redundancy scheme," *2017 IEEE International Conference on Communications (ICC)*, Paris, 2017, pp. 1-6.
- [11] J. Liu, Y. Li, Y. Zhang, L. Su and D. Jin, "Improve Service Chaining Performance with Optimized Middlebox Placement," in *IEEE Transactions on Services Computing*, vol. 10, no. 4, pp. 560-573, 1 July-Aug. 2017.
- [12] C. Ouyang, Y. Wei, S. Leng and Y. Chen, "Service chain performance optimization based on middlebox deployment," *2017 IEEE 17th International Conference on Communication Technology (ICCT)*, Chengdu, 2017, pp. 1947-1952.
- [13] R. Cziva and D. P. Pezaros, "Container Network Functions: Bringing NFV to the Network Edge," in *IEEE Communications Magazine*, vol. 55, no. 6, pp. 24-31, June 2017.
- [14] N. Ma, J. Zhang and T. Huang, "A model based on genetic algorithm for service chain resource allocation in NFV," *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, Chengdu, 2017, pp. 607-611.
- [15] Liu Liu, Hongfang Yu, "Virtual network function service chain mapping based on resource splitting algorithm," *Application Research of Computers*, vol. 33, no. 8, pp. 2440-2445, Aug. 2016.
- [16] Z. Zhu, H. Lu, J. Li and X. Jiang, "Service Function Chain Mapping with Resource Fragmentation Avoidance," *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Singapore, 2017, pp. 1-6.
- [17] Liang Wei, Tao Huang, Jiao Zhang, Jiang Liu, Yunjie Liu, "Service chain mapping algorithm based on reinforcement learning," *Journal on Communications*, vol. 39, no. 1, pp. 90-100, Jan. 2018.