

# Minimizing Delay in Network Function Visualization with Quantum Computing

Wenlu Xuan<sup>1</sup>, Zhongqi Zhao<sup>1</sup>, Lei Fan<sup>2</sup> and Zhu Han<sup>1</sup>

*Dept. of Electrical and Computer Engineering<sup>1</sup> and Dept. of Engineering Technology<sup>2</sup>  
University of Houston*

*wxuan@uh.edu, zzhao27@uh.edu, lfan8@uh.edu, zhan2@uh.edu*

**Abstract**—Network function virtualization (NFV) is a crucial technology for the 5G network development because it can improve the flexibility of employing hardware and reduce the construction of base stations. There are vast service chains in NFV to meet users' requests, which are composed of a sequence of network functions. These virtual network functions (VNFs) are implemented in virtual machines by software and virtual environment. How to deploy VMs to process VNFs of the service chains as soon as possible when users' requests are received is very challenging to solve by traditional algorithms on a large scale. Compared with traditional algorithms, quantum computing has better computational performance because of quantum parallelism. We build an integer linear programming model of the VNF scheduling problem with the objective of minimizing delays, and transfer it into the quadratic unconstrained binary optimization (QUBO) model. Our proposed heuristic algorithm employs a quantum annealer to solve the model. Finally, we evaluate the computational results and explore the feasibility of leveraging quantum computing to solve the VNF scheduling problem.

**Keywords**—network function virtualization; virtual network functions; quantum computing; scheduling problem; delay;

## I. INTRODUCTION

In recent years, network function virtualization (NFV) has attracted more attention in the revolution of wireless network technology because it can reduce the cost of deploying hardware and improve network flexibility. For traditional network function technology, specific hardware can only process particular functions, which raises many challenges for machine manufacturing and maintenance, and causes waste while adjusting network functions. For NFV technology, virtual network functions (VNFs) are implemented at universal standard servers, which solves the issues mentioned above. The standard server integrates many types of equipment. Every virtual machine (VM) is allocated on standard commercial servers. Each VM realizes one or more VNFs of the network by software or virtual environments [1], [2]. VNFs, which will be split, can adopt in more than one VM with parallel running to reduce the processing delay. Briefly speaking, the implementation of different VNFs is based on various software instead of specific hardware. Thus, NFV reduces the difficulty of hardware configuration and ameliorates the compatibility of a network. For the realization of NFV, physical machines (PMs) like standard servers are virtualized as one or more VMs, which attaches much importance to how to allocate VMs at PMs under the

consideration of efficiency and costs. This kind of problem is called the VM embedding problem [3]. Another important problem is how to deploy VMs to process VNFs when users' request is received. In this paper, we propose an integer linear programming (ILP) model with delay minimization as the objective function to solve the VNF scheduling problem.

In the NFV network system, VMs are usually located at data centers, and VMs interconnect with others by virtual links. Each VM is configured with specific computing resources. If the user's request needs to process oversized data packages, the processing delay at VMs cannot be ignored. The transmission delay between two VMs, especially in the same data center, can be neglected because the transmission rate is high. For most users and network service providers, the total delay is a crucial aspect of evaluating the performance of a network, which is identified in the service level agreement. A request of users can be realized by the cooperation of several VNFs, and we introduce a service chain to correspond to the request in the NFV-enabled network [4]. A service chain is a sequence of ordered VNFs. The network processes data via VNFs, definitely following the order in the service chain. Consequently, the total delay of processing the request is the whole period that data goes through the corresponding service chain. In this paper, our model aims to optimize the total processing delay of all service chains in the network.

In the real world, the network always receives significant numbers of requests, and it needs to complete the process as soon as possible, and so the desperate need for network operators is an efficient solver that can optimize large-scale resource allocation. Traditional algorithms cannot meet such a requirement, and so we pin the hope on quantum computing. The superiority of quantum computing is based on quantum parallelism, which means a quantum computer searches for possible outcomes simultaneously. As a result, the computation speed of quantum computers is much faster than that of classical computers. It has been proved the speedup of quantum computing in solving certain problems [5]–[8].

Quantum computing takes advantage of quantum properties, such as quantum superposition and quantum entanglement. The first milestone in quantum computation is the no-cloning theorem, which shows the impossibility of copying an unknown quantum state [9]. This is one of the fundaments

of quantum computing and quantum information. In 1980, Benioff brought up a method of simulating quantum systems by Turing machines [10]. In 1982, Feynman introduced the conception of quantum computation, and following brought up the idea that universal quantum computing can be realized through quantum systems following quantum mechanics [11]. In 1985, David Deutsch first proposed a computing paradigm based on quantum mechanics, which is the framework of the modern quantum computer [12]. Many powerful quantum algorithms were proposed by prominent scientists in the past several decades, such as the Deutsch-Jozsa Algorithm [13], Grover's Algorithm [14], Shor's Algorithm [15], and Quantum Approximate Optimization Algorithm [16]. These algorithms demonstrate the great potential of quantum computing in many fields.

In recent years, tech giants, like IBM, Google, Microsoft, and D-wave, took the lead in developing quantum computers. D-wave company employs quantum annealing techniques to construct the quantum annealer. Quantum annealing assists a quantum system in reaching the lowest energy state. Compared with other quantum computing models, e.g. the analog quantum model and universal quantum gate model, quantum annealing technique provides more quantum bits in current industry practice, which means the quantum annealing hardware has more powerful computational performance. The quantum annealer can efficiently solve the quadratic unconstrained binary optimization (QUBO) problem by using the Ising model, which describes the energy state with coupling qubits interaction and externally applied fields [17]. Therefore, the QUBO model can leverage a qubit system via embedding methods to get an optimal solution. It has been proved that many combinatorial optimization problems can be rewritten in the QUBO form, and it facilitates the application of quantum annealing machines [18].

This paper formulates the VNF scheduling problem as an ILP model with the optimization of delay, which is not easy to be solved by classical algorithms. We transfer the ILP model into the QUBO form and propose a heuristic algorithm to solve it using the D-Wave hybrid solver. We study several cases with different parameters and different scales to evaluate the performance of the D-Wave hybrid solver in solving our model. Our key contributions are as follows:

- We propose an ILP model for the VNF scheduling problem, and then we reformulate the model as the QUBO model, which can be solved by the quantum annealing machine.
- We propose a heuristic algorithm to quickly find a feasible solution, which can help strengthen our QUBO model. We also demonstrate the efficiency of our algorithm in multiple experiments.
- We employ quantum computing to solve the VNF scheduling problem. Our work shows the possibility of using quantum computing to allocate resources in NFV.

The rest of this paper is organized as follows. Section II introduces related work about VNF scheduling problems and quantum computing applications. Section III illustrates the NFV system. In Section IV, we describe the ILP model of the VNF scheduling problem and reformulates the ILP model as the QUBO model, and we propose a heuristic algorithm to embed the QUBO onto quantum annealing hardware. Section V shows the case study results of this problem using quantum computing. Finally, Section VI is the conclusion of the whole paper.

## II. RELATED WORK

### A. VNF Scheduling Problem

Because of the importance of the user request processing time, many researchers study the optimal processing delays in the VNF scheduling problem [19]–[21]. In [19], Diez *et al.* implemented NFV into cloud radio access networks, essential in 5G. They took split selection and scheduling into consideration while minimizing traffic delay globally and partially. The results showed that partial optimization is close to the exact optimal solution. Geared *et al.* [20] formulated a complex VNF chaining and placement model considering queuing delay in virtual links and edge clouds. They also analyzed different queuing models in the same situation. Their proposed schemes satisfied the stringent quality of service and meet service-level agreement requirements for both horizontal scaling and vertical scaling. In [21], researchers studied a VNF scheduling problem with minimizing the total delay, including processing delay and transmission delay. They also considered the dynamic virtual link bandwidth while formulating the model. A genetic-algorithm-based method was proposed to get the optimal solution due to the high complexity of the model.

To improve the practicality of proposed models, many researchers tend to study VNF scheduling problems from multiple aspects including minimizing delay, maximizing throughput, optimizing cost, and improving reliability [22]–[24]. [22] presented an integer nonlinear programming model to illustrate the VNF scheduling problem. They developed a heuristic algorithm to achieve the minimum delay in different scenarios. The resiliency of NFV was also studied by constructing a more reliable virtual network. Luizelli *et al.* [23] leveraged the ILP method to formulate the VNF placement and chaining problem. The ILP was designed to minimize delays while guaranteeing the utilization efficiency of resources. In [24], Ren *et al.* presented a delay-sensitive NFV-enabled multicasting problem in mobile edge clouds. They aimed to minimize the implementation cost of the request and to maximize the system throughput. A heuristic algorithm was developed to solve the complex model.

### B. Quantum Computing

Benefited from the technology development of controlling quantum particles and constructing quantum hardware, quantum computation has attracted more attention in recent years.

Some scientists attempted to use quantum computation to solve optimization problems in wireless networks [25]–[27]. Researchers [25] employed quantum computing to solve sizeable multiple-input multiple-output (MIMO) problems in centralized radio networks. They analyzed the performance under different modulations, which demonstrated that quantum computing is generally valuable in these cases. In the binary phase shift keying (BPSK) communication system, quantum annealers can assist the network in serving 48 users with an extremely low bit error rate. Alanis *et al.* [26] developed a non-dominated quantum optimization (NDQO) algorithm for multi-objective routing problems. NDQO illustrated an approximate optimal performance compared with the state-of-the-art evolutionary algorithms. However, they found that the NDQO algorithm was infeasible to search for an available solution with the number of nodes increasing and the number of routes increasing exponentially. Consequently, they proposed a non-dominated quantum iterative optimization (NDQIO) algorithm in [27]. The NDQIO algorithm run on the quantum hardware parallelization framework. It had a good performance of serving routing in wireless multihop networks despite the complexity reduction compared with the NDQO algorithm.

Some researchers also tried to use quantum annealers to solve classical NP-hard problems [28], [29]. In [28], the minimum multicut (MMC) problems were transformed into QUBO formulations using two different methods. They studied a particular case of the MMC problem on the family of random connected trees. The QUBO model of this case was processed at the D-Wave machine to get an optimal solution. Venturelli *et al.* [29] built a simple model for the job-shop scheduling problem with the makespan minimization. They formulated this model in the QUBO form and embedded it on D-Wave chips. Some strategies of fine-tuning parameters and graph-embedding were also presented in this paper. Their results showed that pre-processing using classical algorithms is efficient in this situation.

### III. SYSTEM MODEL

In our system model, the network offers  $K$  types of VNFs to satisfy user's requirements.  $F = \{f^1, f^2, \dots, f^k, \dots, f^K\}$  denotes the set of VNFs and  $f^k$  denotes the  $k^{th}$  type of functions. Any VM can serve one or more than one function, and any function  $f^k$  can be configured on any VM. VMs can be divided into several groups, and for each group, these VMs serve the same VNFs and have the same computing capacity. According to the users' requirement, these functions compose a service chain  $i$  to process data. Thus, we distinguish different VNF instances of the same network functions by denoting them as  $f_{ij}^k$ , which means the  $j^{th}$  function in service chain  $i$  belongs to the  $k^{th}$  type of functions. We assume that each VM can serve one function at a time. There is only one virtual link between any two VMs in the network, and we denote the virtual link

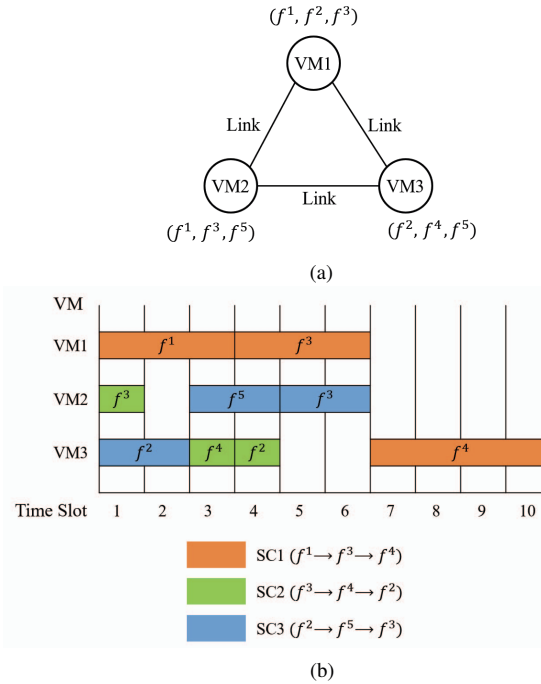


Figure 1. Example of NFW system (a) a simple NFW system (b) a possible arrangement of service chains.

between VM  $m$  to VM  $n$  as  $l_{(m,n)}$ . In our NFW network, all VMs are embedded in PMs located in data centers, and the transmission between any two VMs is high-speed, and so the transmission delay can be neglected. We only consider the minimum total processing delay to scheduling VNFs of service chains.

Workload  $W_{ij}$ , corresponding to the data package required processing, will be processed on VM  $m$ ,  $m \in V_{ij}^k$ . The computing capability of VM  $m$  is denoted by  $C_m$ . Thus, the processing time  $t_{ijm}$  is given by  $W_{ij}/C_m$ . We set the system model as a discrete-time model, and so the controller's working time will be divided into several time slots with the length of  $\Delta T$ .  $T_{ijm}$  is the minimum integer that is equal to or larger than  $t_{ijm}/\Delta T$ . It can be said that the number of time slots occupied by  $t_{ijm}$  is  $T_{ijm}$ .

Fig. 1 is an example of how the controller schedules the VMs to satisfy the request requirements from the users. Suppose that there are three virtual machines, VM1, VM2, and VM3. VM1 can process the functions  $f^1$ ,  $f^2$ , and  $f^3$ . VM2 can process the functions  $f^1$ ,  $f^3$ , and  $f^5$ . VM3 can process the functions  $f^2$ ,  $f^4$ , and  $f^5$ . According to the requests from users, the network receives service chains, SC1 ( $f^1 \rightarrow f^3 \rightarrow f^4$ ) with  $4MB$ , SC2 ( $f^3 \rightarrow f^4 \rightarrow f^2$ ) with  $0.8MB$ , and SC3 ( $f^2 \rightarrow f^5 \rightarrow f^3$ ) with  $2MB$ . In this example, the data size of packages won't change after processed by any VNFs. Every time slot has a length of  $1s$ . We assume that VM1 can process  $1.5MB$  per second, and the computation rate of VM1 is 1.5 times that of VM2 and VM3. The controller arranges VM1 to process  $f^1$  of SC1, and the processing delay is 3 time slots. In SC1, the second

function that needs to be processed is  $f^3$ . Fortunately, VM1 can also process  $f^3$ , and so the controller arranges VM1 to process  $f^3$  because VM1 processes data faster than the other two VMs. The processing delay of  $f^3$  is also 3 time slots. After that, the results of  $f^3$  are transferred to VM3 via the link between VM1 and VM3, and then, the processing delay of  $f^4$  is 4 time slots at VM3. The total delay of SC1 is 10 time slots. The first function of SC2,  $f^3$ , is processed at VM2, which costs 1 time slot. The only VM that can process  $f^4$  is VM3, so the results of  $f^3$  are transmitted to VM3. However, VM3 is occupied by SC3, so SC2 needs to wait until VM3 finishes processing  $f^2$  of SC3. After 1 time slot, VM3 starts to process  $f^4$  of SC2, and the processing time is 0.8s, which means 1 time slot is occupied. After that, the controller arranges VM3 to process  $f^2$  of SC2, and the process occupies 1 time slot. Even though  $f^4$  and  $f^2$  of SC2 are processed at VM3 successively, the whole process is not continuous. At the end of the time slot of serving  $f^4$  of SC2, the controller will determine whether the processing finishes or not, and then arrange other functions to be processed at VM3. The total delay of SC2 is 4 time slots. The first function of SC3 is  $f^2$ , and it is processed at VM3, which costs 2 time slots. After that, the results are sent to VM2 through the link between VM2 and VM3. The second function of SC3 is  $f^5$ , which is processed at VM2 for 2 time slots, and then the third function  $f^3$  is also processed for 2 time slots at VM2. The total delay of SC3 is 6 time slots. The processing delay of all three service chains is 20 time slots, which means 20s. Finally, the controller will evaluate this arrangement and determine whether the total delay of all service chains arrives at the minimum. If not, the controller will rearrange VMs to process these functions.

#### IV. PROBLEM FORMULATION AND ALGORITHM

##### A. ILP Formulation

We develop an ILP model to describe the NVF scheduling problem. This model aims to minimize the total delay of all service chains in the network, and we use the finish time of the last function as the total delay of the corresponding service chain. All notations used in the model and their descriptions are listed in Table I. For the ILP model, the objective function, constraints, and their explanations are listed below.

$$\min_{s_{iJ}} \quad \tilde{s} = \sum_{i=1}^I s_{iJ} \quad (1)$$

$$s_{iJ} = \sum_{m=1}^M \sum_{t=1}^{T_{max}} p_{iJmt} \cdot (t-1) \cdot \Delta T, \quad \forall i. \quad (2)$$

Eq. (1) is the objective function. Eq. (2) shows how to calculate the finish time of any service chain. If  $p_{1J34}$  is equal to 1, it means that the service chain 1 finishes to be processed at the beginning of the 4<sup>th</sup> time slot on VM 3, and so the processing delay of service chain 1 is 3 time slots.

In addition, we have the following constraints.

$$\sum_{m \in V_{ij}^k} x_{ijm} = 1, \quad \forall i, j. \quad (3)$$

$$x_{ijm} = \sum_{t=1}^{T_{max}} z_{ijmt}, \quad \forall i, j, m. \quad (4)$$

Constraint (3) indicates that any function  $f_{ij}^k$  can be processed on only one VM. Notice that in (3) only the VM in the set  $V_{ij}^k$  can be selected. Constraint (4) shows the relationship between  $x_{ijm}$  and  $z_{ijmt}$ . If and only if  $f_{ij}^k$  is allocated to VM  $m$ , this VM can start processing  $f_{ij}^k$  at some point.

$$\sum_{i=1}^I \sum_{j=1}^J y_{ijmt} \leq 1, \quad \forall m, t. \quad (5)$$

Constraint (5) shows that each VM can process at most one function in one time slot. For example, if  $y_{1234} = 1$ , which means that VM 3 processes the second function of service chain 1 in the 4<sup>th</sup> time slot, VM 3 cannot process other functions in this time slot.

$$y_{ijmt} \leq x_{ijm}, \quad \forall i, j, m, t. \quad (6)$$

Constraint (6) indicates the relationship between  $x_{ijm}$  and  $y_{ijmt}$ . If at time  $t$ , VM  $m$  needs to handle function  $f_{ij}^k$ , which means  $y_{ijmt} = 1$ , then  $x_{ijm} = 1$ .

$$\sum_{t=1}^{T_{max}} y_{ijmt} = T_{ijm} \cdot x_{ijm}, \quad \forall i, j; \quad m \in V_{ij}^k. \quad (7)$$

Constraint (7) ensures that required total time  $T_{ijm}$  for processing function  $f_{ij}^k$  must be satisfied. Notice that in (7) only the VM in the set  $V_{ij}^k$  can be selected because if and only if VM  $m$  can process function  $f_{ij}^k$ ,  $T_{ijm}$  exists.

$$z_{ijmt} + p_{ijmt} \leq 1, \quad \forall i, j, m, t. \quad (8)$$

$$y_{ijm(t-1)} - y_{ijmt} + z_{ijmt} - p_{ijmt} = 0, \quad \forall i, j, m, t. \quad (9)$$

Constraint (8) makes sure that  $z_{ijmt}$  and  $p_{ijmt}$  cannot be equal to 1 at the same time, according to the definition of  $z_{ijmt}$  and  $p_{ijmt}$ . Constraint (9) shows the logical relationship between  $y_{ijmt}$ ,  $z_{ijmt}$  and  $p_{ijmt}$ . For example, suppose that  $y_{1111} = 0$  and  $y_{1112} = 1$ . Eq. (9) constrains that  $z_{1112}$  must equal 1 and  $p_{1112}$  must equal 0. For another example, suppose that  $y_{1111} = 1$  and  $y_{1112} = 0$ . Eq. (9) constrains that  $z_{1112}$  must equal 0 and  $p_{1112}$  must equal 1.

$$\sum_{\alpha=1}^{T_{ijm}} z_{ijm(t-\alpha+1)} \leq y_{ijmt}, \quad \forall i, j, t; \quad m \in V_{ij}^k. \quad (10)$$

$$\sum_{m' \in V_{ij}^k} \sum_{\beta=1}^{T_{max}} p_{ijm(t-\beta+1)} \geq z_{i(j+1)m't}, \quad (11)$$

$$\forall i, j, t; \quad m' \in V_{i(j+1)}^{k'}.$$

Constraint (10) guarantees that once the VM starts processing the function  $f_{ij}^k$ , the VM must process it for required time. Constraint (11) means that the next function of the

Table I. LIST OF NOTATIONS

Notation	Description
$i, i'$	index of service chain; $i, i' \in \{1, 2, \dots, I\}$
$j, j'$	index of the sequence of function in service chain; $j, j' \in \{1, 2, \dots, J\}$
$f^k$	the $k^{th}$ type of functions, $k \in \{1, 2, \dots, K\}$
$F$	the set of all $f^k$
$f_{ij}^k$	the $j^{th}$ function in service $i$ belongs to the $k^{th}$ type of functions
$f_{i(j+1)}^{k'}$	the $(j+1)^{th}$ function in service $i$ belongs to the $(k')^{th}$ type of functions; $k' \in \{1, 2, \dots, K\}$ ; $j \in \{1, 2, \dots, (J-1)\}$
$m, m'$	index of VM; $m \in \{1, 2, \dots, M\}$
$n$	index of VM; $n \in \{1, 2, \dots, M\}$
$V_{ij}^k$	the set of VMs which can serve $f_{ij}^k$
$V_{i(j+1)}^{k'}$	the set of VMs which can serve $f_{i(j+1)}^{k'}$
$l_{(m,n)}$	the virtual link between VM $m$ and VM $n$ ;
$\Delta T$	the length of each time slot;
$t$	the $t^{th}$ time slot; $t \in \{1, 2, \dots, T_{max}\}$
$W_{ij}$	the workload of processing $f_{ij}^k$
$C_{ijm}$	the computing capability of VM $m$ which can serve $f_{ij}^k$
$t_{ijm}$	the time length of processing $f_{ij}^k$ on VM $m$
$T_{ijm}$	the number of time slots occupied by processing $f_{ij}^k$ on VM $m$
$s_{iJ}$	the finish time of processing the last function of service chain $i$
$x_{ijm}$	equals to 1, if VM $m$ is used to process $f_{ij}^k$ ; otherwise, equals to 0
$y_{ijmt}$	equals to 1, if VM $m$ is used to process $f_{ij}^k$ in the time slot $t$ ; otherwise, equals to 0
$z_{ijmt}$	equals to 1, if VM $m$ starts to process $f_{ij}^k$ at the beginning of the time slot $t$ ; otherwise, equals to 0
$p_{ijmt}$	equals to 1, if VM $m$ finishes processing $f_{ij}^k$ at the beginning of the time slot $t$ ; otherwise, equals to 0

service chain must be processed after the processing of the one before it.

$$x_{ijm} = y_{ijmt} = z_{ijmt} = p_{ijmt} = 0, \quad (12)$$

$$\forall i, j, t; \quad m \notin V_{ij}^k.$$

$$\sum_{m \in V_{ij}^k} \sum_{t=1}^{T_{max}} z_{ijmt} = \sum_{m \in V_{ij}^k} \sum_{t=1}^{T_{max}} p_{ijmt} = 1, \quad \forall i, j. \quad (13)$$

Constraint (12) shows that  $x_{ijm}$ ,  $y_{ijmt}$ ,  $z_{ijmt}$ , and  $p_{ijmt}$  must be equal to 0 if the VM cannot process the function  $f_{ij}^k$ . Constraint (12) ensures that for any function  $f_{ij}^k$ , only one  $z_{ijmt}$ , and one  $p_{ijmt}$  can be equal to 1 because the function  $f_{ij}^k$  can be only processed for one time.

### B. QUBO Formulation

D-Wave quantum annealers can only solve the optimization problem in the QUBO formulation. To leverage quantum annealers, we need to transform the ILP model into the QUBO formulation. The definition of QUBO is as follows:

$$\min_x f(x) = x^T Q x, \quad (14)$$

where  $x$  is the vector of binary variables, and  $Q$  is an upper-diagonal matrix or symmetric matrix. As the definition shows above, there is only an objective function and no constraints in the QUBO formulation. All constraints in

Table II. LIST OF CONSTRAINT-PENALTY PAIRS

Constraint	Equivalent Penalty
$x_1 + x_2 = 1$	$P(x_1 + x_2 - 1)^2$
$x_1 + x_2 + x_3 \leq 1$	$P(x_1 x_2 + x_1 x_3 + x_2 x_3)$
$x_1 + x_2 \leq x_3$	$P(x_1 + x_2 - x_3 + \sum_l a_l r_l)^2$
$x_1 + x_2 = b$	$P(x_1 + x_2 - b)^2$

our model must be reformulated into quadratic penalties, and then be added to the original objective function. We choose the value of penalty coefficients according to the influence of original constraints in searching for the optimal solution. The principles of transforming classical constraints as equivalent penalties are listed in Table II, where  $x_1, x_2$  and  $x_3$  are binary variables.  $r_l$  is a binary slack variable.  $a_l$  and  $b$  are constants.  $P$  is the penalty coefficient. The transformed results of (1)-(13) are listed in Appendix appendix A.

### C. Proposed Algorithm

We propose a heuristic algorithm to employ the D-Wave solver to solve our model. Due to the limitation of qubits on the D-Wave QPU server, cases with too many variables cannot be solved. On the one hand, to let the solver handles as many variables as possible, we turn to the D-Wave hybrid solver, which employs classical computation to assist

quantum annealing and can accept at least one thousand variables for this optimization problem. On the other hand, since the value of  $T_{max}$  has an effect on the number of variables, we could reduce the range of  $t$ , which means find a feasible  $T_{max}$ , to reduce the number of variables for solving more complex cases by the hybrid solver. In our system model, the range of  $t$  is the working time of the NFV system controller, which means that  $T_{max}$  is sufficiently large, and the controller has the freedom to determine how to schedule VNFs. However, the objective of our model is to minimize delays of service chains, and we don't need to provide such a long time tolerance for the solver to schedule VNFs. If we set a big value to  $T_{max}$ , it will bring a lot of variables to our model, and then we need much more qubits to help solve our QUBO model. Therefore, we leverage a greedy algorithm to assist us in finding a reasonable  $T_{max}$ . For the proposed greedy algorithm, we rearrange all VNFs in service chains to a service chain, and every function  $f_{ij}^k$  will be allocated to VM  $m$ , which processes this function for the shortest time. We set the total processing delay by this greedy algorithm to  $T_{max}$ , and then the QUBO model is embedded in the quantum annealing hardware by an algorithm. The penalty coefficients of the QUBO model play an important role for the hybrid solver in searching for optimal solutions. Since the penalty coefficient needs to be sufficiently large compared with other values in the QUBO model, before setting up the penalty, we evaluate the maximum value that the objective function can reach. In all case studies, the penalty coefficients are set to about 100 times the value of the maximum objective value. After the penalty pre-processing, the penalty could be fine-tuned according to the output, and so we can reach a more suitable penalty set. Finally, we use a D-Wave hybrid solver through the proposed algorithm to solve the QUBO model. The whole proposed algorithm is presented in Algorithm 1.

---

#### Algorithm 1

---

**Input:** parameters,  $I, J, M$ ; the functions in service chain  $i, f_{ij}^k$ ; the set of VMs which can process  $f_{ij}^k, V_{ij}^k$ ; the NFV network;

**Output:**  $\tilde{s}, x_{ijm}, y_{ijmt}, z_{ijmt}, p_{ijmt}$ ;

- 1: Set the value of  $T_{max}$ : run the proposed greedy algorithm to get a feasible  $T_{max}$ ;
  - 2: Set the value of penalty coefficients;
  - 3: Eqs. (15)-(26): transform from (1)-(13);
  - 4: The QUBO model: add all terms in (15)-(26) to the right hand side of (15);
  - 5: Embedding the QUBO model onto the quantum annealing hardware;
  - 6: **return**  $\tilde{s}, x_{ijm}, y_{ijmt}, z_{ijmt}, p_{ijmt}$ ;
- 

## V. EXPERIMENT

We study the cases with different parameters  $I, J$ , and  $M$ , and different service chains, and analyze the performance of

the quantum annealer under different cases.  $\Delta T$  is equal to 1s in all cases. If the hybrid solver cannot output a solution, we will increase the value of  $T_{max}$  until the hybrid solver can provide a feasible solution. All results are listed in Table III. In Table III, the third column shows the results of the proposed greedy algorithm, which are feasible solutions of our model and can be used to evaluate the outputs of the D-Wave hybrid solver. The fourth column lists the solutions given by the D-Wave hybrid solver, which are the total processing delays of all service chains. The fifth column presents the processing delays of the most time-consuming service chain in each case given by the D-Wave hybrid solver. The sixth column shows the average processing time of each VM, which impacts the costs of resources. The longer the average processing time, the more cost of electricity. The seventh column is the sizes of matrix  $Q$  denoted in the QUBO formulation definition in (14). As the matrix  $Q$  size increases, the solver needs to employ more qubits, and the difficulty of solving the problem increases. In Table III, we can find that for case 10 and case 12, the longest delay given by the D-Wave hybrid solver is longer than the delay given by the proposed greedy algorithm. It means that the solution given by the hybrid solver cannot be the optimal solution, which shows the hybrid solver cannot solve our model on such a large scale.

Fig. 2 and Fig. 3 are the probability distribution of the results for running 50 times for each case. For all cases,  $\Delta T$  is equal to 1s. Fig. 2 shows the most time-consuming service chain processing delays, and Fig. 3 shows the total processing delays of all service chains. In Fig. 2, we can find a higher probability of achieving optimal solutions when the matrix  $Q$  size is small. With the increase of matrix  $Q$  size, the highest probability of the longest delay moves to  $T_{max}$ . It means that the difficulty of finding the optimal solution increases as the matrix  $Q$  size increases. In Fig. 3, we can find that for case  $a$ , it is effortless to find the optimal solution by the hybrid solver. For the case  $a$  and case  $b$ , the sizes of matrix  $Q$  are small, and solutions are concentrated. For other cases, the sizes of matrix  $Q$  are larger, and solutions are more dispersed. Table IV shows the hybrid solver running time and the QPU working time for each case. The hybrid solver spends a much longer time on finding a feasible solution for case  $f$ . Unfortunately, the solver only has the success rate of 4% to solve case  $f$ , which means if you use the hybrid solver to study case  $f$  100 times, you can only get a feasible solution 4 times. In other words, it is difficult for the hybrid solver to solve this case due to such a large matrix  $Q$  size, which corresponds to the results in Fig. 2 and Fig. 3.

## VI. CONCLUSION

In this paper, we formulate the VNF scheduling problem as an ILP model with the optimization of delay and transfer the ILP model into the QUBO form, which can be solved by the quantum annealing machine. We propose a heuristic

Table III. SIMULATION RESULTS

Case	Parameters	Result of the greedy algorithm (s)	Objective Solution (s)	The Longest Delay (s)	Average Processing Time for Each VM (s)	Matrix Q Size
1	$I = 2, J = 2, M = 2$	7	8	5	4.0	(280, 280)
2	$I = 2, J = 2, M = 2$	9	11	6	5.5	(338, 338)
3	$I = 2, J = 2, M = 2$	5	6	4	3.0	(200, 200)
4	$I = 2, J = 3, M = 2$	12	16	9	8.0	(662, 662)
5	$I = 2, J = 3, M = 2$	12	20	11	10.0	(662, 662)
6	$I = 2, J = 3, M = 2$	19	36	19	18.0	(1012, 1012)
7	$I = 3, J = 2, M = 2$	14	17	11	8.5	(732, 732)
8	$I = 3, J = 2, M = 2$	10	16	9	8.0	(540, 540)
9	$I = 3, J = 2, M = 2$	13	22	12	11.0	(684, 684)
10	$I = 3, J = 3, M = 2$	15	49	18	24.5	(1462, 1462)
11	$I = 3, J = 3, M = 2$	13	29	13	14.5	(1173, 1173)
12	$I = 3, J = 3, M = 2$	14	39	15	19.5	(1266, 1266)

Table IV. TIME CONSUMING AND SUCCESS RATE

Case	Average QPU access time (s)	Average solver run time (s)	Success rate
a	0.065	2.993	100%
b	0.065	2.997	64%
c	0.063	2.998	36%
d	0.061	2.994	100%
e	0.064	2.997	58%
f	0.063	3.630	4%

algorithm to solve the QUBO formulation, using the D-wave hybrid solver. We report and analyze the solutions of several cases under different settings. Our work shows the possibility of using a quantum computer to allocate resources in NFV. From the results of the case study, we can find that the performance of the hybrid solver is better in cases with fewer variables. There is less probability to achieve the optimal solution in cases with more variables.

## ACKNOWLEDGMENT

This work is partially supported by US NSF CNS-2107216, EARS-1839818, EPCN-2045978 and Toyota.

## REFERENCES

- [1] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network function virtualization: challenges and opportunities for innovations," *IEEE Commun. Mag.*, vol. 53, no. 2, pp. 90–97, Feb. 2015.
- [2] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, "Network function virtualization: state-of-the-art and research challenges," *IEEE Commun. Surv. Tut.*, vol. 18, no. 1, pp. 236–262, 1st Quart. 2016.
- [3] L. Zhang, H. Zhang, L. Yu, H. Xu, L. Song, and Z. Han, "Virtual resource allocation for mobile edge computing: a hypergraph matching approach," in *Proc. IEEE GLOBECOM*, Waikoloa, HI, Dec. 2019.
- [4] A. M. Medhat, T. Taleb, A. Elmangoush, G. A. Carella, S. Covaci, and T. Magedanz, "Service function chaining in next generation networks: state of the art and research challenges," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 216–223, Oct. 2017.

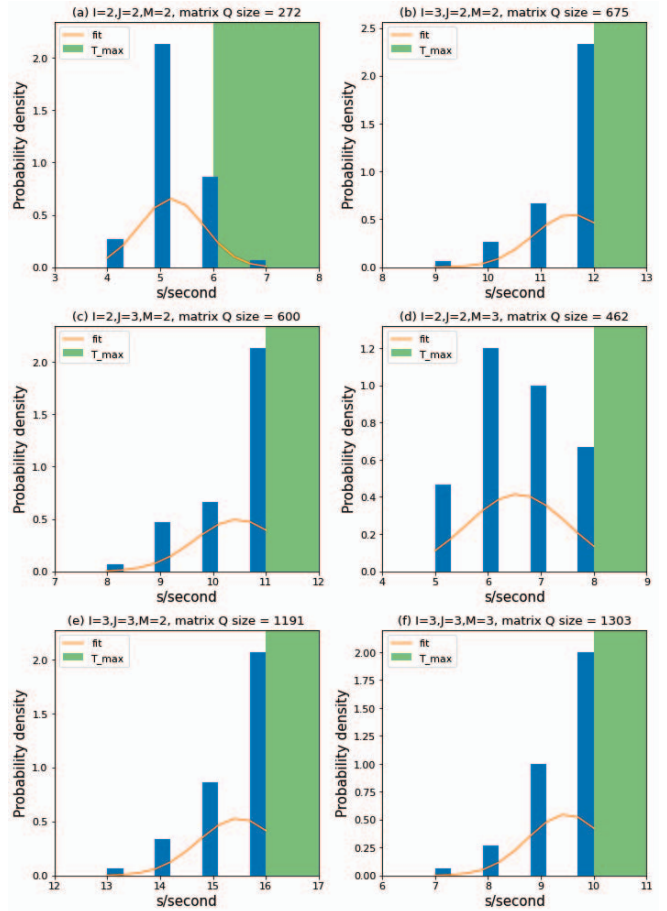


Figure 2. Histograms of the most time-consuming service chain processing delays given by the D-Wave hybrid solver (a) case a (b) case b (c) case c (d) case d (e) case e (f) case f.

- [5] J. Roland and N. J. Cerf, "Quantum search by local adiabatic evolution," *Phys. Rev. A*, vol. 65, no. 4, p. 042308, Mar. 2002.
- [6] A. M. Childs, R. Cleve, E. Deotto, E. Farhi, S. Gutmann, and D. A. Spielman, "Exponential algorithmic speedup by a quantum walk," in *Proc. 35th annu. ACM symp. Theory comput.*, San Diego, CA, Jul. 2003.
- [7] I. Hen, "Period finding with adiabatic quantum computation,"



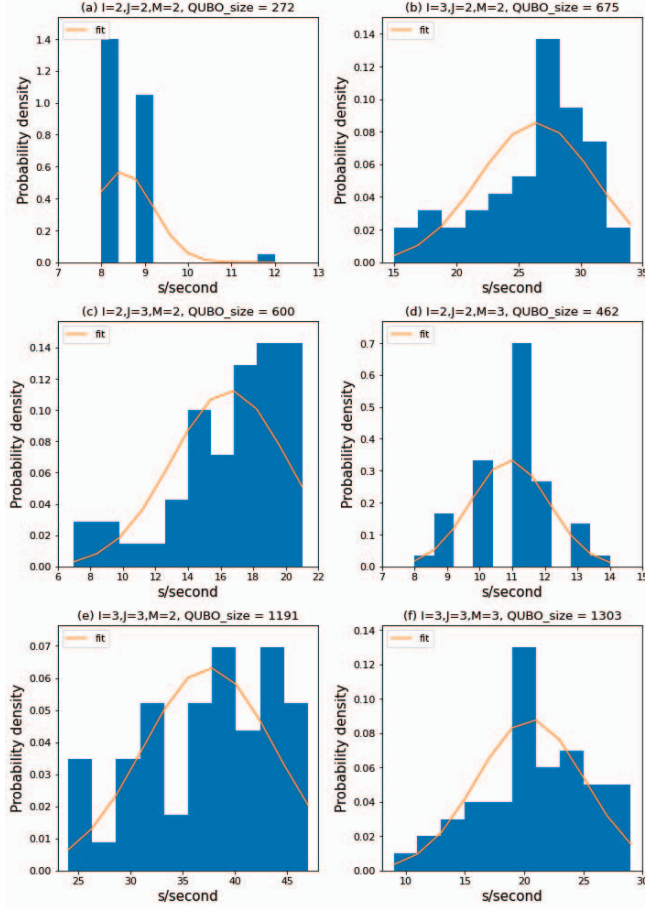


Figure 3. Histograms of the total processing delays of all service chains given by the D-Wave hybrid solver (a) case a (b) case b (c) case c (d) case d (e) case e (f) case f.

*EPL*, vol. 105, no. 5, p. 50005, Mar. 2014.

- [8] R. D. Somma, D. Nagaj, and M. Kieferová, "Quantum speedup by quantum annealing," *Phys. Rev. Lett.*, vol. 109, no. 5, p. 050501, Jul. 2012.
- [9] W. K. Wootters and W. H. Zurek, "A single quantum cannot be cloned," *Nature*, vol. 299, no. 5886, pp. 802–803, Oct. 1982.
- [10] P. Benioff, "The computer as a physical system: a microscopic quantum mechanical hamiltonian model of computers as represented by turing machines," *J. Stat. Phys.*, vol. 22, no. 5, pp. 563–591, May 1980.
- [11] R. P. Feynman, "Simulating physics with computers," *Int. J. Theor. Phys.*, vol. 21, pp. 467–488, Jun. 1982.
- [12] D. Deutsch, "Quantum theory, the Church–Turing principle and the universal quantum computer," *Proc. Royal Soc. Lond. A*, vol. 400, no. 1818, pp. 97–117, Jul. 1985.
- [13] D. Deutsch and R. Jozsa, "Rapid solution of problems by quantum computation," *Proc. Royal Soc. Lond. A*, vol. 439, no. 1907, pp. 553–558, Dec. 1992.
- [14] L. K. Grover, "A fast quantum mechanical algorithm for database search," in *Proc. 28th Annu. ACM Symp. Theory Comput.*, Philadelphia, PA, May 1996.
- [15] P. W. Shor, "Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer," *SIAM J. Comput.*, vol. 26, no. 5, pp. 1484–1509, Oct. 1997.
- [16] E. Farhi, J. Goldstone, and S. Gutmann, "A quantum approximate optimization algorithm," *arXiv preprint arXiv:1411.4028*, Nov. 2014.
- [17] E. Ising, "Beitrag zur theorie des ferromagnetismus," *Z. Physik*, vol. 31, no. 1, pp. 253–258, Feb. 1925.
- [18] F. Glover, G. Kochenberger, and Y. Du, "Quantum bridge analytics I: a tutorial on formulating and using qubo models," *4OR*, vol. 17, no. 4, pp. 335–371, Dec. 2019.
- [19] L. Diez, V. Gonzalez, and R. Aguero, "Minimizing delay in NFV 5G networks by means of flexible split selection and scheduling," in *IEEE 90th Veh. Technol. Conf.*, Honolulu, HI, Sep. 2019.
- [20] R. Gouareb, V. Friderikos, and A.-H. Aghvami, "Virtual network functions routing and placement for edge cloud latency minimization," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 10, pp. 2346–2357, Sep. 2018.
- [21] L. Qu, C. Assi, and K. Shaban, "Network function virtualization scheduling with transmission delay optimization," in *IEEE/IFIP Netw. Oper. Manag. Symp.*, Istanbul, Turkey, Jul. 2016.
- [22] S. Yang, F. Li, R. Yahyapour, and X. Fu, "Delay-sensitive and availability-aware virtual network function scheduling for NFV," *IEEE Trans. Serv. Comput.*, pp. 1–14, Jul. 2019.
- [23] M. C. Luizelli, L. R. Bays, L. S. Buriol, M. P. Barcellos, and L. P. Gaspary, "Piecing together the NFV provisioning puzzle: efficient placement and chaining of virtual network functions," in *IFIP/IEEE Int. Symp. Integr. Netw. Manag.*, Ottawa, ON, Canada, May 2015.
- [24] H. Ren, Z. Xu, W. Liang, Q. Xia, P. Zhou, O. F. Rana, A. Galis, and G. Wu, "Efficient algorithms for delay-aware NFV-enabled multicasting in mobile edge clouds with resource sharing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 31, no. 9, pp. 2050–2066, Mar. 2020.
- [25] M. Kim, D. Venturelli, and K. Jamieson, "Leveraging quantum annealing for large MIMO processing in centralized radio access networks," in *Proc. ACM SIGCOMM*, Beijing, China, Aug. 2019.
- [26] D. Alanis, P. Botsinis, S. X. Ng, and L. Hanzo, "Quantum-assisted routing optimization for self-organizing networks," *IEEE Access*, vol. 2, pp. 614–632, Jun. 2014.
- [27] D. Alanis, P. Botsinis, Z. Babar, S. X. Ng, and L. Hanzo, "Non-dominated quantum iterative routing optimization for wireless multihop networks," *IEEE Access*, vol. 3, pp. 1704–1728, Sep. 2015.
- [28] W. Cruz-Santos, S. E. Venegas-Andraca, and M. Lanzagorta, "A QUBO formulation of minimum multicut problem instances in trees for D-wave quantum annealers," *Sci. Rep.*, vol. 9, no. 17216, pp. 1–12, Nov. 2019.
- [29] D. Venturelli, D. Marchand, and G. Rojo, "Job shop scheduling solver based on quantum annealing," in *Proceedings of the 11th Workshop on Constraint Satisfaction Techniques for Planning and Scheduling*, London, UK, Jun. 2016, pp. 25–34.

## APPENDIX

For the QUBO formulation, there is only an objective function. Therefore, all constraints in the ILP model must be reformulated into quadratic penalties, and then we add them to the original objective function to form the QUBO formulation. All penalties and their explanations are listed below. In all penalty terms, all penalty coefficients, denoted as  $P$  with subscripts, are sufficiently large positive constants.

We combine (1) and (2), and have



$$\min_{p_{ijmt}} \tilde{s} = \sum_{i=1}^I \sum_{m=1}^M \sum_{t=1}^{T_{max}} p_{ijmt} \cdot (t-1) \cdot \Delta T. \quad (15)$$

$$P_{1ij} \left( \sum_{m \in V_{ij}^k} x_{ijm} - 1 \right)^2, \quad \forall i, j. \quad (16)$$

$$P_{1ijm} \left( \sum_{t=1}^{T_{max}} z_{ijmt} - x_{ijm} \right)^2, \quad \forall i, j, m. \quad (17)$$

Eq. (16) is transformed from (3). We find that constraint (3) only allows one  $x_{ijm}$ , which  $m \in V_{ij}^k$ , is equal to 1. In (16), if more than one  $x_{ijm}$  is equal to 1, these terms will add a huge value to the objective function. Thus, the optimizer will avoid this situation. This is why constraints (3) can be transformed as (16). Eq. (17) is transformed from (4). Constraint (4) only allows the situation that  $x_{ijm} = 1$  and  $\sum_{t=1}^{T_{max}} z_{ijmt} = 1$  and the situation that  $x_{ijm} = 0$  and  $\sum_{t=1}^{T_{max}} z_{ijmt} = 0$ . Eq. (17) has the same effect on the model. Therefore, we transform constraint (4) into the terms in (17). We transform (5) to the following equation.

$$P_{mt} \left( \sum_{i \neq i' \text{ or } j \neq j'} (y_{ijmt} \cdot y_{i'j'mt}) \right), \quad \forall m, t. \quad (18)$$

Constraint (5) shows that either or neither  $y_{ijmt}$  can be equal to 1. If any two  $y_{ijmt}$  are equal to 1 in (18), these terms will lead the solution away from the minimum. Therefore, constraint (5) is equivalent to the terms in (18). We transform (6) to the following equation.

$$P_{1ijmt} \left( y_{ijmt} - x_{ijm} + r_{1ijmt} \right)^2, \quad \forall i, j, m, t. \quad (19)$$

We need to add slack variables to convert the inequalities in (6) into equalities. We only add one binary slack variable to constraint (6) because the difference between the right hand side and the left hand side must be equal to or less than 1. If  $y_{ijmt} = 1$  and  $x_{ijm} = 0$ , these terms will add a huge value to the objective function. Thus, the optimizer will avoid this situation. This is why constraint (6) can be transformed as the terms in (19).  $r_{1ijmt}$  is a binary slack variable. We transform (7) to the following equation.

$$P_{2ijm} \left( \sum_{t=1}^{T_{max}} y_{ijmt} - T_{ijm} \cdot x_{ijm} \right)^2, \quad \forall i, j; m \in V_{ij}^k. \quad (20)$$

Constraint (7) ensures that the number of  $y_{ijmt}$  valued 1 is  $T_{ijm}$  when  $x_{ijm} = 1$ . If the number of  $y_{ijmt}$  valued 1 is not equal to  $T_{ijm}$  when  $x_{ijm} = 1$ , the terms in (20) will add a large value to the objective function. Thus, the optimizer will avoid this situation. This is why constraint (7) can be transformed as the terms in (20). We transform (8) to the following equation.

$$P_{2ijmt} (z_{ijmt} \cdot p_{ijmt}), \quad \forall i, j, m, t. \quad (21)$$

Constraint (8) precludes the situation that both  $z_{ijmt}$  and  $p_{ijmt}$  are equal to 1. All terms in (21) have the same effect as constraint (8), and so we transform (8) as the terms in (21). We transform (9) to the following equation.

$$P_{3ijmt} \left( y_{ijm(t-1)} - y_{ijmt} + z_{ijmt} - p_{ijmt} \right)^2, \quad \forall i, j, m, t. \quad (22)$$

$$P_{4ijmt} \left( \sum_{\alpha=1}^{T_{ijm}} z_{ijm(t-\alpha+1)} - y_{ijmt} + r_{2ijmt} \right)^2, \quad \forall i, j, t; \quad m \in V_{ij}^k. \quad (23)$$

$$P_{ijm't} \left( z_{i(j+1)m't} - \sum_{m \in V_{ij}^k} \sum_{\beta=1}^{T_{max}} p_{ijm(t-\beta+1)} + r_{ijm't} \right)^2, \quad \forall i, j, t; \quad m' \in V_{i(j+1)}^{k'}. \quad (24)$$

Eq. (23) is equivalent to (10).  $r_{2ijmt}$  is a binary slack variable. Eq. (24) is transformed from (11). We need to add slack variables to convert the inequalities in (11) into equalities. We only add one binary slack variable to constraint (11) because the difference between the right hand side and the left hand side must be equal to or less than 1.  $r_{ijm't}$  is a binary slack variable. We transform (12) and (13) to the following equations.

$$P_{3ijm} \cdot x_{ijm}^2 + P_{5ijmt} \cdot y_{ijmt}^2 + P_{6ijmt} \cdot z_{ijmt}^2 + P_{7ijmt} \cdot p_{ijmt}^2, \quad \forall i, j, t; \quad m \notin V_{ij}^k. \quad (25)$$

$$P_{2ij} \left( \sum_{m \in V_{ij}^k} \sum_{t=1}^{T_{max}} z_{ijmt} - 1 \right)^2 + P_{3ij} \left( \sum_{m \in V_{ij}^k} \sum_{t=1}^{T_{max}} p_{ijmt} - 1 \right)^2, \quad \forall i, j. \quad (26)$$

To form the QUBO formulation, all terms in (16)-(26) need to be added to the right hand side of (15).