NFV Resource Allocation using Mixed Queuing Network Model

Min Sang Yoon, and Ahmed E. Kamal
Department of Electrical and Computer Engineering
Iowa State University, IA, 50010
{my222, kamal}@iastate.edu

***Abstract*—Network resource virtualization emerged as the future of communication technology recently, and the advent of Software Define Network (SDN) and Network Function Virtualization (NFV) enables the realization of network resource virtualization. NFV virtualizes traditional physical middle-boxes that implement specific network functions. Since multiple network functions can be virtualized in a single server or data center, the network operator can save Capital Expenditure (CAPEX) and Operational Expenditure (OPEX) through NFV. Since each customer demands different types of VNFs with various applications, the service requirements are different for all VNFs. Therefore, allocating multiple Virtual Network Functions(VNFs) to limited network resource requires efficient resource allocation. We propose an efficient resource allocation strategy of VNFs in a single server by employing mixed queuing network model while minimizing the customers' waiting time in the system. The problem is formulated as a convex problem. However, this problem is impossible to be solved because of the closed queuing network calculation. So we use an approximation algorithm to solve this problem. Numerical results of this model show performance metrics of mixed queuing network. Also, we could find that the approximate algorithm has a close optimal solution by comparing them with neighbor solutions.**

I. Introduction

In traditional network infrastructure, network functions are implemented by hardware based physical middle box. In Network Function Virtualization (NFV) infrastructure, multiple network functions are assigned to servers or data centers as software based virtual machines. Virtualized Network Functions (VNFs) have logical connections between them and process user requests depending on service chain. The service chain represents sets of VNFs that should be applied to customers' requests. For example, some user requests need to be processed in Firewall, QoS, and WAN opt network functions and some other users will request Firewall, QoS, DDos, Rate limiter network functions. So if we have multiple VNFs in the server, the flows decide which sets of VNFs should be mapped. Since hardware-based physical boxes are substituted to software-based virtual machines, NFV decreases Capital Expenditure (CAPEX) and Operation Expenditure (OPEX) and makes it possible to configure much more flexible network architectures. At the same time, the efficient resource allocation to VNFs is important. Since VNFs receive different requests from customers depending on service chains, it is important to allocate appropriate resource to VNFs. We propose a resource allocation strategy to VNFs in a single server environment. VNFs will be allocated to a single server and our strategy determines the efficient resource to VNFs depending on receiving workload of VNFs.

We employ BCMP mixed queuing network model to analyze the system performance. Arrival rates to VNFs will be assumed as multi-classes Poisson arrival. Each VNF is modeled as a queue. The connectivity of VNFs are considered as mixed queuing network model including multiple open chains and closed chains. Our purpose is minimizing the expected waiting time of service chains. VNFs' routing probabilities to other network functions are assumed to be known based on the history data and will be learned periodically. In order to minimize the maximum expected waiting time of service chains, we formulate the problem as a convex problem with capacity constraints. The optimal service rates will be assigned to each VNFs to minimize the maximum expected waiting time of service chains while guaranteeing the capacity constraints.

In order to calculate the expected waiting time of mixed queuing network, Mean value Analysis (MVA) technique will be employed. However, it is impossible to solve the optimization problem because the closed queuing network calculation. The mean value computation of closed network, expected queuing length, expected waiting time, and expected throughput, requires recursive algorithm from the empty state of network. Also, the service rate of the stations are required to be given when we start the computation. The service rate is unknown decision variable and our purpose is minimizing the expected waiting time of service chain through optimal service rate allocation. So it is not possible to solve the optimization problem in traditional approach. So we propose the algorithm that allocates efficient service rates to VNFs by using approximation approach.

This paper is the first paper studied NFV resource allocation by using queuing network model. Also, the resource allocation in mixed queuing network is not studied before in our knowledge. The rest of this paper is organized as follows. Section II introduces related work on NFV and queuing network models. The system model is described in Section III and the problem formulation will be discussed in Section IV. In order to solve the optimization problem, the approximation algorithm will be described in Section V. The numerical analysis of the system will be presented in Section VI and we will end this paper with conclusions in Section VII.

II. Related Work

The VNF placement problem has been studied widely because of its influence on system cost and efficiency. A VNF placement model is proposed by applying NFV and SDN to

LTE mobile core gateways in [4]. They minimized transport network load overhead while satisfying the data-plane delay budget.

In [5], virtual Deep Pack Inspection (vDPI) function placement problem is studied. Mathieu Bouet *et al.* formulate the problem as an Integer Linear Program (ILP) and minimize the costs.

Sevil Mehraghdam *et al.* suggested a model formalizing the chaining of VNFs using a context-free language and proposed a mapping scheme to the network by using Mixed Integer Quadratically Constrained Program (MIQCP) for finding the optimal placement in [6].

VNF placement problem is also studied in the data center and clouds environment. Ming Xia *et al.* identified the possibility of minimizing expensive optical/electronic/optical conversions for NFV chaining in data centers. They formulated the problem of optimal VNF placement as a Binary Integer Program (BIP) and proposed a heuristic approach to solve the problem in [7].

Queuing network model is studied by many researchers. However, the resource allocation problem in queuing network models has not been studied well enough. Optimal server resource allocation problem is presented in [8] only for open queuing network model. Alex Zhang *et al.* suggested nonlinear integer programming model for determining the number of machines in multi-tier server network. They used an open queuing network model average response time to minimize the total number of machines while satisfying the average response time of open queuing network.

The service rate allocation problem is studied in [9] by Onno J. Boxma. He described the model that allocates the optimal service rate in Jackson networks and provided closed form expression for optimal service rate allocation to all stations in the network.

III. System Model

We employ OpenNF controller model proposed in [3]. OpenNF is composed of two layers: flow manger and NFV state manager as we can see in Figure 1. The NFV state manager continuously monitors the state of VNFs and report their states to the flow manager. Then the flow manager sends VNFs status information to SDN switch so that the SDN switches determine flow controls to VNFs. VNFs are allocated to the single server and a hypervisor allocates resources to VNFs depending on their routing and processing status. Since NFV state manager monitors the status of VNFs continuously, the NFV state manger and VNFs generate closed network per service chain. The NFV state manage generates a signal that keep going around VNFs in the same service chain and report VNFs status to NFV state manager. The number of jobs in closed networks are determined depending on how many VNFs are related to service chain and how many classes are related to each VNF. Multiple closed networks can exist on the server depending on service provider's needs and policy.

After receiving monitoring result from the NFV state manager, flow manager sends monitoring results to SDN switches. Then, SDN switches determine the flow control to VNFs. All incoming and departing flows of VNFs are influenced by SDN switches. SDN switches control the arriving flows to VNFs depending on their status and forward flows to all VNFs included in the related service chain of flows. Then, all flows depart the server through the SDN switches after finishing the process in all required VNFs. Therefore, VNFs form open network with SDN switches per service chain to receive incoming flows and send out the departing flows.

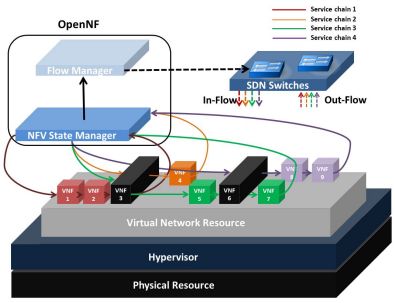

Figure 1. System model

The purpose of this paper is to minimize the maximum expected waiting time of service chains in the single server by assigning optimal service rates to VNFs. Each VNF has different routing probability between VNFs, visiting ratio, and incoming rates by open networks. Some VNFs receive high traffic requests from the SDN switch. On the other hand, some VNFs will not receive frequent requests from users. Thus, it would not be best to assign equal service rates to all virtual network functions. We measure the routing probability or inter-connectivity between network functions continuously and decide the optimal service rates of VNFs.

IV. Problem Formulation

Table 1 shows the notations used in the following formulations.

A. BCMP Mixed Queuing Network Model

We consider a BCMP mixed queuing network model composed of multi-class jobs and multiple stations, VNFs.

The BCMP queuing network contains a finite number, V, of stations having different disciplines: the First-Come First-Served (FCFS), Processor Sharing (PS), Infinite Servers (IS), and Last-Come First-Served (LCFS). FCFS type stations have a negative exponential service time distribution and others can have a general service time distribution. The only restriction of other disciplines is that the service time distribution should have a rational Laplace transform. The service rate of each station is expressed as the average amount of work completed per time unit.

Open class jobs are incoming traffic to the server that runs VNFs. The classes are classified based on types of the service chain. Each service chain requires different levels of Service Level Agreement (SLA). The delay sensitive applications like video streaming or voice call require higher levels of SLA. On the other hand, delay tolerant applications have relatively lower levels of SLA. Thus, each class of open network has

Table I
NOTATIONS

N_S	Population vector of classes
N_C	Station population vector of closed classes
N_O	Station population vector of open classes
s_{ic}	Average amount of loads brought by class c jobs to station i (per visit).
μ_i	Service rate of VNF i. Average amount of work completed per unit time.
C_i	The inverse of service rate of station i ($1/\mu_i$).
$E_i(n)$	Auxiliary function for the calculation of the effective capacity of mixed station i. $E_i(n) = \frac{1}{(1-L_oC_i)^{n+1}}$
L_{Oi}	Load brought to station i by open class jobs.
$\overline{X_{ic}}(N)$	Throughput of class c jobs at station i.
$P_{Ci}(N_C, N_S)$	Marginal length distribution of closed classes.
$P_{Oi}(N_O, N_S)$	Marginal length distribution of open classes.
$\overline{n_{ic}}$	Expected queuing length of closed class c in station i.
$\overline{n_{iC}}$	Expected queuing length of closed classes in station i.
$\overline{n_{io}}$	Expected queuing length of open class o in station i.
$\overline{w_{ic}}$	Expected waiting time of closed class c in station i.
$\overline{w_{io}}$	Expected waiting time of open class o in station i.
$\underline{P}$	NFV policy matrix.

a different average amount of work brought by a single job and it will be considered when we analyze the open networks. Closed networks correspond to monitoring states of VNFs as in Figure 1. Closed networks are as many as the number of service chains.

We employ Mean Value Analysis (MVA) technique to analyze the characteristics of the queuing networks. MVA is an efficient method to analyze the product form expressed queuing network model. MVA employs the mean value equation augmented by Little's law. MVA has a significant merit in terms of computational cost compared to other product form queuing network model analysis methods such as joint distribution analysis or convolution algorithm.

Let C denote the set of closed classes, O denote the set of open classes, and M denote the number of service chains, $O = \{1, 2, \ldots, M\}, C = \{M+1, M+2, \ldots, 2M\}$. Since classes are generated as many as service chains for open and closed classes, O and C have the same number of elements. Let $N_S = [N_1, N_2, \ldots, N_M, N_{M+1}, \ldots, N_{2M}]$ be the population vector of each class in the system. The state N_M is a vector including the possible distribution of jobs over the VNFs, $N_M = [N_{M1}, N_{M2}, \ldots, N_{MV}]$. V stands for the number of VNFs in the server. The number of jobs in closed queue is fixed at a certain time point and the number of jobs in an open class denotes the upper bound for the population of jobs.

The marginal length distribution of closed network is described as (3) in [1].

$$P_{Ci}(N_C, N_S) = \sum_{c \subseteq C} s_{ic} C_i \frac{E_i(n_c)}{E_i(n_c - 1)} \overline{X_{ic}}(N_s) P_{Ci}(n_c - 1, N_s - 1) \quad (1)$$

The marginal distribution of closed queue is a recursive equation that is defined in terms of $P_{Ci}(n_c - 1, N_S - 1)$. Therefore, the marginal distribution of closed network can be calculated by using a recursive algorithm starting from the empty space of the closed network.

We can obtain the expected number of jobs of closed class by using the following equation.

$$\begin{aligned}\overline{n_{ic}}(N_S) &= \sum_{n_c=1}^{N_C} n_c P_{Ci}(n_c, N_S) \\ &= s_{ic}\overline{X_{ic}}(N_S) \sum_{n_c=1}^{N_C} n_c C_i \frac{E_i(n_c)}{E_i(n_c - 1)} P_{Ci}(n_c - 1, N_s - 1), \forall c \subseteq C\end{aligned} \quad (2)$$

We substitute $C_i \frac{E_i(N_C)}{E_i(N_C - 1)}$ by EC_i which defined as $EC_i = \frac{C_i}{1 - L_{Oi}C_i}$ in [1]. By using Little's law, the expected waiting time in a closed network can be calculated by:

$$\begin{aligned}\overline{w_{ic}}(N_S) &= \frac{\overline{n_{ic}}(N_S)}{\overline{X_{ic}}(N_S)} \\ &= s_{ic} EC_i \sum_{n_c=1}^{N_C} n_c P_{Ci}(n_c - 1, N_S - 1), \forall c \subseteq C\end{aligned} \quad (3)$$

We rewrite (3) as,

$$\begin{aligned}\overline{w_{ic}}(N_S) &= s_{ic} EC_i \sum_{n_c=1}^{N_C} (1 + n_c - 1) P_{Ci}(n_c - 1, N_S - 1) \\ &= s_{ic} EC_i [\sum_{n_c=1}^{N_c} P_{Ci}(n_c - 1, N_S - 1) \\ &\quad + \sum_{n_c=1}^{N_c} (n_c - 1) P_{Ci}(n_c - 1, N_S - 1)] \\ &= s_{ic} EC_i (1 + \overline{n_{ic}}(N_S - 1_c)), \forall c \subseteq C\end{aligned} \quad (4)$$

$N_S - 1_c$ represents the state that one closed class job is reduced from the state N_S. (4) shows that the expected waiting time of closed class c is affected by the state $N_S - 1_c$.

The marginal queuing length distribution of open network is described in [1].

$$\begin{aligned}P_{Oi}(N_O, N_S) = N_O! \Pi_{o \subseteq O} \frac{(\lambda_{io} s_{io})^{N_O}}{N_O} \sum_{n_c=0}^{N_C} \binom{n_c + n_o}{n_c} \\ \Pi_{j=n_c+1}^{n_c+n_o} C_i [E_i(n_c)]^{-1} P_{Ci}(n_c, N_S)\end{aligned} \quad (5)$$

(5) includes the marginal distribution of closed network, which means that the marginal distribution of open network is affected by the marginal distribution of closed network. Therefore the distribution of open network can be calculated after we obtain the marginal distribution of closed network in the state of N_C. The expected number of jobs in open network

can be described by using (5) as described in [1].

$$\begin{aligned}\overline{n_{io}}(N_S) &= \sum_{n_o=1}^{\infty} n_o P_{Oi}(n_o, N_S) \\ &= \lambda_{io} s_{io} \sum_{n_c=0}^{N_C} (n_c+1) EC_i P_{Ci}(n_c, N_S), \forall o \subseteq O\end{aligned} \tag{6}$$

From Little's law, the expected waiting time of open network can be calculated like below:

$$\begin{aligned}\overline{w_{io}}(N_S) &= s_{io} EC_i \sum_{n_c=0}^{N_C} (n_c+1) P_{Ci}(n_c, N_S) \\ &= s_{io} EC_i (1 + \overline{n_{iC}}(N_S)), \forall o \subseteq O\end{aligned} \tag{7}$$

So the expected waiting time of open network can be described based on the expected length of closed network at state N_S.

B. Optimal Service Rate Allocation

The purpose of this paper is minimizing the maximum expected waiting time of service chains in the limited resource. Each service chain includes different types and number of VNFs. Therefore, minimizing the maximum expected time of service chains will guarantee the SLA of NFV.

VNFs receive different numbers of requests depending on service chain types and routing probability. In order to minimize the expected waiting time of a service chain, more resources will be allocated to VNFs with relatively heavy traffic. We consider resource allocation in a single server. The resource allocation can be easily managed by the hypervisor depending on VNF requirements.

We assume a policy matrix $\underline{P}$, which is a $M \times V$ matrix that shows policies of each service chain. Each row represents the policy of service chains and each column corresponding to VNFs. If $P_{ij} = 1$, i_{th} service chain needs to be processed in j_{th} VNF. The convex optimization problem to minimize the maximum expected waiting time service chains can be formulated like below. Since $\underline{P}$ is $M \times V$ matrix and $\vec{w}$ is $V \times 1$ vector, $\underline{P}\vec{w}$ gives $M \times 1$ vectors which is equal to the expected waiting time of each service chain. Thus, the maximum of the vector will give the maximum expected time of service chains.

$$Minimize_{\mu_i} \quad max_M(\underline{P}\vec{w})_{1\times M}$$

$$Subject\ to$$

$$C1: \sum_{i \subseteq V} \mu_i \leq \overline{C}$$

$$C2: \mu_i \geq \sum_{o \subseteq O} s_{io}\lambda_{io} + \sum_{c \subseteq C} s_{ic}\overline{n_{ic}}(N_S), \forall i$$

$\overline{C}$ denotes the total capacity of the server. The total service rates allocated to VNFs cannot exceed the total capacity of the server in C1. C2 represents the minimum service rate of VNF i. For the stability of the system, the minimum service rates of each VNF should be greater than open and closed classes job requests. $\vec{w}$ is the vector that denotes the expected waiting time of each VNF including open and closed networks, $\vec{w} = (w_1, w_2, \ldots, w_V)$.

$$\overrightarrow{w_i} = \sum_{j \subseteq O \cup C} \overline{w_{ij}}(N_S) \tag{8}$$

$\overline{w_{ic}}(N_S)$ and $\overline{w_{io}}(N_S)$ are defined in (4) and (7). $\overline{w_{ic}}(N_S)$ includes $\overline{n_{ic}}(N_S - 1_c)$, the expected number of jobs in closed network at $N_S - 1_c$ state, and $\overline{w_{io}}(N_S)$ includes $\overline{n_{ic}}(N_S)$ the expected number of jobs in closed queue at N_S state. However, calculating the expected number of jobs in closed queue requires recursive computation starting from the empty state of closed queue as mentioned in the previous section. The expected queuing length calculation in closed network requires the knowledge of service rate of VNFs. In other words, the minimization of waiting time through service rate allocation cannot be accomplished because we cannot calculate the expected waiting time of closed network without knowledge of service rates, decision variables, of service stations, VNFs. So we employ the Schweitzer Core algorithm described in [2]. The Schweitzer Core algorithm estimates the expected queuing length of closed network by using a iterative algorithm. The detailed process of the algorithm will be explained in section V.

The convexity of the problem can be simply proved. The objective function is a linear combination of $\overline{w_{ic}}(N_S)$ and $\overline{w_{io}}(N_S)$ for arbitrary c and i depending on service policy. If we can prove that the individual terms $\overline{w_{ic}}(N_S)$ and $\overline{w_{io}}(N_S)$ are convex, the objective function is a convex function. $\overline{n_{ic}}(N_S)$ and $\overline{n_{ic}}(N_S - 1_c)$ will be estimated as a constant by the Schweitzer Core algorithm. Equation (5) can be expressed as:

$$\begin{aligned}\overline{w_{ic}}(N_S) &= s_{ic} \frac{C_i}{1 - L_{O_i} C_i} (1 + \overline{n_{ic}}(N_S - 1_c)) \\ &= \frac{s_{ic}}{\mu_i - L_{O_i}} (1 + \overline{n_{ic}}(N_S - 1_c))\end{aligned} \tag{9}$$

If we find the second derivative of $\overline{w_{ic}}(N_S)$, $\frac{\partial^2 \overline{w_{ic}}(N_S)}{\partial \mu_i^2}$, we can obtain the equation below.

$$\frac{\partial^2 \overline{w_{ic}}(N_S)}{\partial \mu_i^2} = \frac{2 s_{ic}}{(\mu_i - L_{O_i})^3} (1 + \overline{n_{ic}}(N_S - 1_c)) \tag{10}$$

For the convexity condition, $\frac{\partial^2 \overline{w_{ic}}(N_S)}{\partial \mu_i^2} \geq 0$, $\mu - L_{O_i}$ should be greater than zero. The constraint C2 restricts the minimum service rates VNFs. C2 includes the requirements of open and closed networks, so the convexity condition is satisfied with C2. Using the same approach, the convexity condition for open classes, $\frac{\partial^2 \overline{w_{io}}(N_S)}{\partial \mu_i^2} \geq 0$, are satisfied for all VNFs. Constraint C1 is a linear combination of μ_i. The Right Hand Side (RHS) of C2 is constant. Thus, the problem is convex problem.

V. Heuristic Algorithm for the Mixed Queuing Networks Model

A. The Schweitzer Core Algorithm

The heuristic algorithm is suggested because the MVA algorithm requires a complete solution for all population in the

closed network. For example, if we assume there are N_C jobs in a closed class, the MVA algorithm requires the complete solution for all possible populations from $(0, 0, \ldots, 0)$ to N_C about all stations, VNFs. This process demands high level of computation and storage complexity. The Schweitzer Core algorithm approximates the expected number of jobs in the closed network based on an iterative approximation technique.

Initial distribution of jobs is approximated by distributing jobs uniformly to related stations, (11). Since classes denote the service chains in our model, the population of closed classes will be uniformly distributed to associated VNFs depending on service chain policy.

$$\overline{n_{ik}}(N_S) = \frac{N_k}{\sum_{j \subseteq V} \underline{P}(k,j)} I(\underline{P}(k,i)), \forall k \subseteq C, i \subseteq V \quad (11)$$

$I(\underline{P}(c,i))$ is an indicator function, which is 0 when $\underline{P}(c,i) = 0$ and 1 when $\underline{P}(c,i)$=1 for all c and i.

The population of $N_S - 1_c$, one c class job is reduced from N_S state, state is estimated proportional to the N_S state population vector in [2], where N_C is the population distribution state of closed network. When one c class job is reduced from N_S, the expected queuing length of other classes are not changed from N_S state in the first equation of (12). However, the expected queuing length of c class job is reduced proportional to the c class job, N_c from N_S state in the second equation of (12).

$$\overline{n_{ik}}(N_S - 1_c) = \begin{cases} \overline{n_{ik}}(N_S) & for\ k \neq c, k \subseteq C \\ \frac{N_c - 1}{N_c} \overline{n_{ik}}(N_S) & for\ k = c, k \subseteq C \end{cases} \quad (12)$$

Since $\overline{n_{ic}}(N_C)$ is approximated from (11), $\overline{n_{ic}}(N_C - 1_c)$ also can be estimated through (12). After obtaining $\overline{n_{ic}}(N_C - 1_c)$, we calculate more exact expected queuing length by using (2) and (4), $(\overline{n_{ic}}(N_S) \rightarrow \overline{n_{ic}}(N_S - 1_c) \rightarrow \overline{n_{ic}}(N_S))$. Then, we repeat this iteration until the queuing length difference between sequential iterations converges. The Core Schweitzer algorithm is described in Algorithm 1. $\overline{n_{ic}}(N_S)$ is approximated by using (11), which means that the number of jobs in closed queue are uniformly distributed to related service chains. Thus, we can compute the number of jobs in closed queue of $n_{ic}(N_S - 1_c)$ state by using (12) (line 3). Then more accurate expected queuing length of $n_{ic}(N_S)$ state is computed again by using (3), (4), and Little's law (line 4). If the expected queuing length difference between t and $t+1$ iterations are greater than the evaluation factor, we increase the iteration t to $t+1$ and go back to step 1. If the termination condition is satisfied, we finish the algorithm and compute the expected queuing length, throughput, and waiting time of the closed networks of N_S state (line 5 - line 11).

Algorithm 1 The Core Schweitzer Algorithm

1: INPUT: Closed classes c, the number of VNFs V, N_c for all classes c, $\overline{n_{ic}}(N_S)$ for each VNF, s_{iC}, and EC_i
2: Initialize iteration t to 1.
3: **Step1.** Compute approximate $\overline{n_{ic}}(N_S - 1_c)$ from (12) $\forall c$.
4: **Step2.** Compute approximate $\overline{n_{ic}}(N_S)$ by using (3), (4), and $\overline{n_{ic}}(N_S - 1_c)$ $\forall k$.
5: **if** $max_{\forall i,c} \frac{(\overline{n_{ik}}(N_S)^t - \overline{n_{ic}}(N_S)^{t-1})}{N_S} \geq \frac{1}{(4000+16|N_S|)}$ **then**
6: Superscript t designates iteration, then set $t = t+1$ and goto step 1.
7: **else**
8: Go to final step.
9: Increase t to t+1.
10: **end if**
11: **Final Step.** Compute throughput estimates $\overline{x_{iC}}(N_S)$.
12: Output: $\overline{n_{ic}}(N_S)$, $\overline{x_{ic}}(N_S)$, $\overline{w_{ic}}(N_S)$, $\forall c \subseteq C$

B. Approximate Optimal Service Rates Allocation in Mixed Queuing Network

The optimization problem could not be solved because it includes $\overline{n_{ic}}(N_S)$ and $\overline{n_{ic}}(N_S - 1_c)$. Since two terms can be obtained through recursive algorithm which requires service rates of VNFs, unknown decision variables, it has not been possible to solve the optimization problem.

We employ the initial estimation of the Schweitzer Core algorithm to estimate the expected queuing length of $\overline{n_{ic}}(N_S - 1_c)$ and $\overline{n_{ic}}(N_S)$ states. $\overline{n_{ic}}(N_S)$ is estimated by (11) and $\overline{n_{ic}}(N_S - 1_c)$ will be obtained by (12) with the estimated value of $\overline{n_{ic}}(N_S)$. By substituting $\overline{n_{ic}}(N_S)$ and $\overline{n_{ic}}(N_S - 1_c)$ in (4) and (7), the optimization problem can be solved from approximate mean queuing length of closed classes.

We propose Algorithm 2 that computes the approximate optimal service rate allocation in mixed queuing networks. In step 1, the approximate expected queuing length of N_S and $N_S - 1_c$ states are obtained in closed classes network by using (12) and (11). Iteration value k is set to 1 (line 2 - line 5). Since the $\overline{n_{ic}}(N_S)$ and $\overline{n_{ic}}(N_S - 1_c)$ are estimated by (12) and (11), we can solve the convex optimization problem that finds optimal service rate allocation by substituting $\overline{n_{ic}}(N_S)$ and $\overline{n_{ic}}(N_S - 1_c)$ in (4) and (7) in step 2 (line 6 - line 9). By using optimally allocated service rates, we can compute more exact expected queuing length by using the Schweitzer Core Algorithm. Improved expected queuing lengths $\overline{n_{ic}}(N_S)$ and $\overline{n_{ic}}(N_S - 1_c)$ are obtained through the Schweitzer Core Algorithm with inputs: $\overline{n_{ic}}(N_S)$ and $\overline{n_{ic}}(N_S - 1_c)$, and $(C_i^*)_k$ in step 3 (line 10 - line 13). The optimization problem is initially solved with approximate expected queuing length. Therefore, the resource allocation in kth iteration is not optimal allocation of $k+1$th expected queuing length. In step 4, we solve the convex optimization problem with $k+1$th iteration's expected queuing length and obtain new optimal service rate allocation, $(C_i^*)_{k+1}$. Since the service rate is changed, the expected queue length of networks will be changed as well. So the algorithm goes back to step 3 and repeat the calculation of the expected queuing length and solving optimization problem (line 14 - line 17). We repeat this process until k reaches k_{max} iterations (line 18 - line 23). In our experiments, this algorithm could achieve convergence within 10 iterations. After finishing the algorithm, we select the optimal service rate allocation of iteration k which has the minimum expected waiting time of service chains through all iterations. The complexity of algorithm can be calculated

Algorithm 2 Service Rate Allocation Algorithm in Mixed Queuing Network

INPUT: s_{io}, s_{ic}, L_{Oi}, N_C, k_{max}.
Step1:
  Initialize k=1.
  Obtain approximate expected queuing length of closed classes, $\overline{n_{ic}}(N_S)_k$ by using (11) in N_S state $\forall$ c, i.
  Compute approximate expected queuing length of closed classes $N_S - 1c$ state, $\overline{n_{ic}}(N_S - 1_c)_k$ by using (12), $\forall$ c, i,.
Step2:
  Substitute $\overline{n_{ic}}(N_S)_k$ and $\overline{n_{ic}}(N_S - 1_c)_k$ in (4) and (7) and solve the optimization problem.
  Optimal service allocations, $(C_i^*)_k$, are obtained for all VNFs.
  Compute the maximum expected waiting time of service chains, $max_{(C_i^*)_k}(\underline{P}\vec{w})_k$ at k_{th} iteration.
Step3:
  Increase the iteration k= k+1.
  Use the Schweitzer Core Algorithm with optimal service rates, $(C_i^*)_{k-1}$, to compute exact expected queuing length of closed classes.
  Obtain advanced $\overline{n_{ic}}(N_S)_k$ and $\overline{n_{ic}}(N_S - 1_c)_k$.
Step4:
  Solve the optimization problem by substituting $\overline{n_{ic}}(N_S)_k$ and $\overline{n_{ic}}(N_S - 1_c)_k$ in (5) and (8).
  Obtain new service rate allocation, $(C_i^*)_k$.
  Compute the maximum expected waiting time of service chain, $max_{(C_i^*)_k}(\underline{P}\vec{w})_k$
Step5:
if k=k_{max} **then**
    Terminate Algorithm.
else
    Go back to step 3.
end if
OUTPUT: $k^* = argmin_k[max_{(C_i^*)_k}(\underline{P}\vec{w})_k]$, $(C_i^*)_{k^*}$, $max_{(C_i^*)_{k^*}}(\underline{P}\vec{w})_{k^*}$

as $O(optimization\ solving\ time + (K_{max} - 1)(VC^2 + optimization\ solving\ time))$.

VI. Numerical Results

The proposed algorithm is implemented in MATLAB and CVX is used to solve the optimization problem. In our numerical analysis model, eight types of VNFs are assigned to a single server: NFV State Manager, Firewall, QoS, Distributed Denial of Service (DDos), WAN opt, Instruction Detection System (IDS), Load Balancer, and Rate Limiter. There are four types of service chains including different VNFs. The system model is shown in Figure 2. Since NFV State manager works for monitoring the states of VNFs, NFV State manager is only related to closed networks. Therefore, all closed networks require NFV State Manager but this is not included in open networks. Arrivals to each VNF are generated randomly per class according to a Poisson distribution with $\lambda = 100$. The

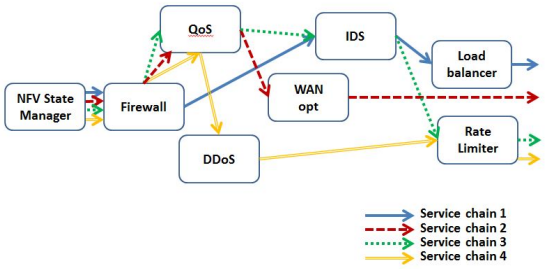

Figure 2. Simulation Model

average amount of loads per visit, s_c, is randomly provided by normal distribution with $\mu = 20$ and $\sigma = 5$ per classes. The number of jobs in closed networks is also randomly given by Normal distribution with $\mu = 10$ and $\sigma = 5$. The iteration parameter k_{max} is set to 30 and the total service rate of the server is basically set to 60,000 per one minute. Figure 3 shows

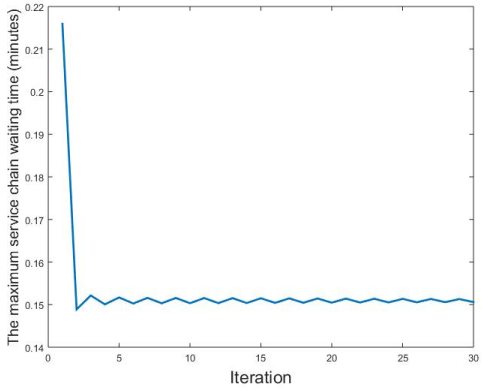

Figure 3. The maximum expected waiting time of service chain

the performance of the algorithm. As the algorithm is repeated, the expected queuing length is stabilized from the initially approximate queuing length. As iterations are repeated, the closed class queuing length is accurately computed for each VNF and service rates are optimally allocated based on these queuing length. So we could see that the expected waiting time difference between iterations becomes smaller with iterations. The Y-axis presents the maximum expected waiting time of service chains. We selected the iteration k^* which has the minimum value of the objective function value and $(C_i^*)_{k^*}$ to be the approximate optimal solution. Figure 4 represents the maximum expected waiting time of service chains as we increase the capacity of the server. As we increase the capacity of the server, we could reduce the expected waiting time of service chains. We can observe that the closed network has more impact on waiting time of the system than the open network. Figure 5 (a) shows the approximate optimal service rate allocation to each VNF when the maximum server capacity is set to 60000. Figure 5 (b) presents the minimum service rate requirements of open and closed classes workloads. NFV State Manager (NSM) does not receive any open classes loads, but the third highest service rate is assigned

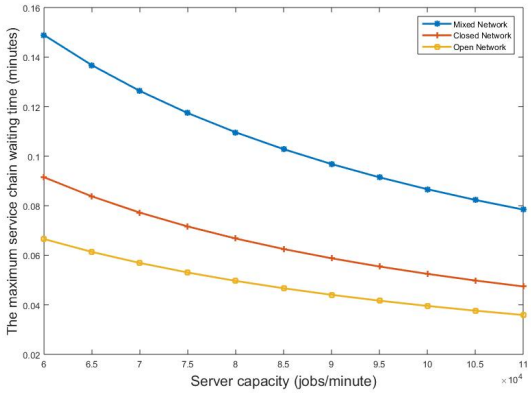

Figure 4. The maximum expected waiting time of service chains with increase in capacity of server

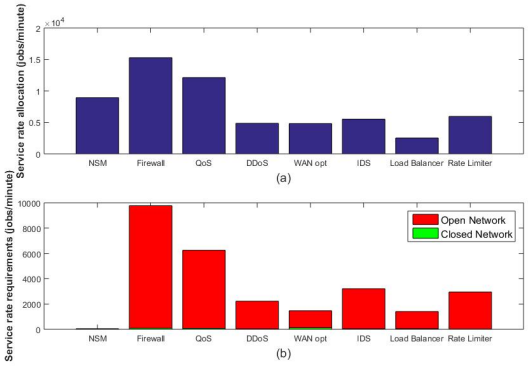

Figure 5. Service rate allocation to VNFs

to NSM. Also, the high service rate is assigned to WAN opt because WAN opt receives the high workloads from the closed classes even though WAN opt receives relatively low workload by open network. Therefore, we can observe that the workloads of closed classes have a significant impact on the resource allocation of the mixed queuing network. In order

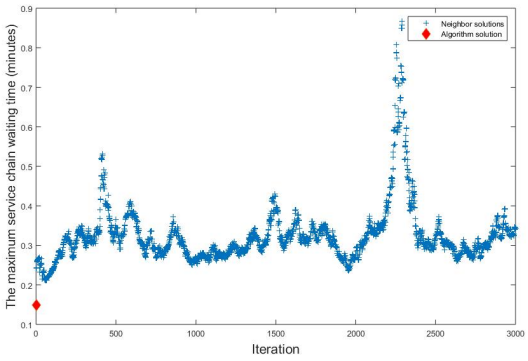

Figure 6. Neighbor solutions

to verify the effectiveness of the algorithm, we compared the approximate optimal solution with neighbor solutions. In every iteration, we search neighbor solution which has different service rate allocation and closed classes jobs distribution. Since the service rate is optimized under a given closed classes jobs distribution, it is required to change the closed class job distribution as well to find the neighbor solution. Two VNFs are randomly selected, and five percentage of service rates and the number of closed class jobs are exchanged between randomly selected VNFs. In Figure 6, we could observe that none of the neighbor solutions has a better objective function value than the algorithm solution.

VII. Conclusions

In this paper, NFV service chains are modeled as a BCMP mixed queuing network and formulated a convex problem to minimize the expected waiting time of service chains. Since the problem can not be solved in the mixed queuing network because the expected queuing length cannot be calculated without knowledge of service rate, an approximate service rate allocation algorithm is proposed. In numerical analysis, we could observe that closed class networks have more influence on the expected waiting time of the system. So high service rates are assigned to VNFs receiving heavy requests from closed class networks.

References

[1] S.C. Bruell, G. Balbo, and P.V. Afshari, *"Mean value analysis of mixed, multiple class BCMP networks with load dependent service stations"*, Performance Evaluation, 4, (1984), pp.241-260.

[2] K.M. Chandy and D. Neuse, *"Linearizer: A heuristic algorithm for queuing network models of computing systems"*, Magazine Communications of the ACM, vol 25, issue 2, Feb 1982. pp.126-134.

[3] A. Gemer-Jacobson, R. Viswanathan, C. Prakash, R. Grandl, J. Khalid, S. Das, A. Akella, *"OpenNF: enabling innovation in network function control"*, Proceedings of the 2014 ACM conference on SIGCOMM, pp. 163-174.

[4] A. Basta, W. Kellerer, M. Hoffmann, H.J. Morper, and K. Hoffmann, *"Applying NFV and SDN to LTE mobile core gateways, the functions placement problem"*, in Proc. 4th Workshop AllThingsCellular, New York, NY, USA: ACM, 2014, pp. 33-38.

[5] M. Bouet, J. Leguay, and V. Conan, *"Cost-based placement of VDPI functions in nfv infrastructure"*, in Proc. IEEE 1st IEEE Conf. NetSoft, Apr. 2015, pp.1-9.

[6] S. Mehraghdam, M. Keller, and H. Karl, *"Specifying and placing chains of virtual network functions"*, in Proc. IEEE 3rd Int. Conf. CloudNet, Oct. 2014, pp. 7-13.

[7] M. xia, M. Shirazipour, Y. Zhang, H. Green, and A. Takacs, *"Network function placement for NFV chaining in packet/optical datacenters"*, J. Lightw. Technol., vol. 33, no. 8, pp. 1565-1570, Apr 2015.

[8] A. Zhang, P. Santos, D. Beyer, and H. Tang, *"Optimal Server Resource Allocation Using an Open Queuing Network Model of Response Time"*, available: http://www.hpl.hp.com/techreports/2002/HPL-2002-301.pdf.

[9] O. J. Boxma, *"Static Optimization of Queuing Systems"*, WSSIAA 5, 1995, pp. 1-16.

[10] R. Mijumbi, J. Serrat, J. gorricho, N. Bouten, and F. D. Turck, *"Network function virtualization: state-of the art and research challenges"*, IEEE Communications survey & tutorials, vol. 18, No. 1, first quarter 2016, pp. 236- 262.

[11] M. Reiser and S. S. Lavenberg, *"Mean-value analysis of closed multichain queuing networks"*, Journal of the association for computing machinery, vol. 27, No. 2, Apr 1980, pp. 313-322.

[12] F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios, *"Open, closed, and mixed networks of queues with different classes of customers"*, Journal of the association for computing machinery, vol. 22, No. 2, Apr. 1975, pp. 248-260.

[13] R. Szabo, M. Kind, F. J. Westphal, H. Woesner, d. Jocha, and A. Csaszar, *"Elastic network functions: opportunities and challenges"*, IEEE Network Magazine, vol. 29, No. 3, May/June 2015, pp. 15-29.

[14] W. ding, W. Qi, J. Wang, and B. Chen, *"OpenSCaaS: an open service chain as service platform toward the integration of SDN and NFV"*, IEEE Network Magazine, vol. 29, No. 3, May/June 2015, pp. 30-35.

[15] T. Wood, K.K. Ramakrishnan, J. Hwang, G. Liu, and W. Zhang, *"Toward a software-based network: integrating software defined networking and network function virtualization"*, IEEE Network Magazine, vol. 29, No. 3, May/June 2015, pp. 36-41.

[16] R. Yu, G. Xue, V. T. Kilari, and X. Zhang, *"Network function virtualization in the multi-tenant cloud"*, IEEE Network Magazine, vol. 29, No. 3, May/June 2015, pp. 42-47.

[17] P.J. Schweitzer, G. Serazzi, M. Broglia, *"A queue-shift approximation technique for product-form queuing networks"*, Computer performance evaluation, vol. 1469 of the series lecture notes in computer science, pp. 267-279.