Stochastic Performance Analysis of Network Function Virtualisation in Future Internet

Wang Miao, Geyong Min, Yulei Wu, Haojun Huang, Zhiwei Zhao, Haozhe Wang and Chunbo Luo

***Abstract*—Network Function Virtualisation (NFV) has been considered as a promising technology for future Internet to increase network flexibility, accelerate service innovation and reduce the Capital Expenditures (CAPEX) and Operational Expenditures (OPEX) costs, through migrating network functions from dedicated network devices to commodity hardware. Recent studies reveal that although this migration of network function brings the network operation unprecedented flexibility and controllability, NFV-based architecture suffers from serious performance degradation compared with traditional service provisioning on dedicated devices. In order to achieve a comprehensive understanding of the service provisioning capability of NFV, this paper proposes a novel analytical model based on Stochastic Network Calculus (SNC) to quantitatively investigate the end-to-end performance bound of NFV networks. To capture the dynamic and on-demand NFV features, both the non-bursty traffic, *e.g.* Poisson process, and the bursty traffic, *e.g.* Markov Modulated Poisson Process (MMPP), are jointly considered in the developed model to characterise the arriving traffic. To address the challenges of resource competition and end-to-end NFV chaining, the property of convolution associativity and leftover service technologies of SNC are exploited to calculate the available resources of Virtual Network Function (VNF) nodes in the presence of multiple competing traffic, and transfer the complex NFV chain into an equivalent system for performance derivation and analysis. Both the numerical analysis and extensive simulation experiments are conducted to validate the accuracy of the proposed analytical model. Results demonstrate that the analytical performance metrics match well with those obtained from the simulation experiments and numerical analysis. In addition, the developed model is used as a practical and cost-effective tool to investigate the strategies of the service chain design and resource allocations in NFV networks.**

***Index Terms*—NFV, Performance Analysis, Stochastic Network Calculus, Performance Bounds, SLA, Future Internet.**

I. Introduction

With the explosive increase in demand for wireless broadband services to deliver content-rich and resource hungry applications, traditional network architectures are struggling to provide the satisfactory network performance in terms of flexibility, scalability and reliability, due to its inherent features such as hardware-based service provision, multi-protocols coexistence and manual configuration and management. The significant improvements of network infrastructures are urgently needed to guarantee the continuous network evolution and innovation. To address this important challenge, Network Function Virtualisation (NFV) has been recently proposed as a promising networking technology for future Internet and attracted significant research efforts [1] [2] [3]. NFV is a network architecture [4] that uses IT virtualisation technology to decouple network functions, *e.g.* firewall, load balancer, router and gateway, from the dedicated hardware device, and deploy these functions on commodity hardware, *e.g.* x86 standard servers. The network function migration brings appealing benefits for network operation and maintenance. For instance, for network service provisioning with highly complex and dynamic demand for applications and services, NFV brings better flexibility, programmability and faster innovation. In addition, through deploying the network functions on commodity hardware devices rather than purchasing new dedicated devices, NFV provides a good solution for network service providers to reduce their CAPEX and OPEX, as the standard servers and network functions can be dynamically re-used in an on-demand manner.

NFV has been regarded as the key technology for the evolution of future Internet. However, one of the major obstacles to realise the ambitions of NFV architecture is the performance degradation caused by NFV compared with dedicated network devices. For instance, Ge *et al.* [5] showed that industry standard servers could not provide satisfactory performance for Quality-of-Service (QoS) constrained network functions, *e.g.* Deep Packet Inspection (DPI). The issue of performance degradation has attracted tremendous research efforts to enhance the performance of NFV networks; and interesting research results have been reported in the literature [6] [7] [8], *e.g.* hardware and software acceleration technologies. Most of the existing researches barely exploit software simulations or network measurements to investigate the performance for the NFV architecture, paying little attention to accurate and quantitative analysis of the performance bottleneck, especially for large-scale and complex network configurations. Compared with the simulation or measurement methods, analytical models could capture the inherent features of a complex system and yield significant insights into the system design and performance improvement. Therefore, analytical models have been considered as a cost-effective method to study system performance.

For the analytical modelling of NFV networks, it is of paramount importance to capture the unique mechanism of NFV network operation and management. In this area, Stochastic Network Calculus (SNC) [9] [10] [11] has attracted research interests for dynamic network analysis. Instead of giving the average performance metrics, SNC focuses on

W. Miao, G. Min, Y. Wu, H. Wang and C. Luo are with the College of Engineering, Mathematics and Physical Sciences, University of Exeter, UK
H. Huang is with China University of Geosciences, China.
Z. Zhao is with University of Electronic Science and Technology of China, China.

Manuscript received 10 October 2018. (Corresponding authors: Geyong Min and Yulei Wu.)

deriving the performance bounds within the constraint of Service Level Agreement (SLA) requirements, making it suitable for studying the performance of NFV service chains in dynamic network environment. There have been some research works that leverage SNC to investigate the performance of service provisioning [12] [13]. For instance, the authors in [12] explored SNC to estimate the available bandwidth of wireless communication systems. The work in [13] proposed an analytical model to investigate the performance of cross-domain networks, i.e. from access network, core network, to datacenter networks. Although these studies have achieved some promising progresses for the performance analysis of dynamic service provisioning, most of them overlooked the unique features of NFV service provisioning, such as underlying resource competition, on-demand NFV service chain deployment and SLA guarantee. For example, Virtual Network Functions (VNFs) in NFV service chains may be added and removed during service operation in order to continuously offer satisfied performance, exhibiting the on-demand feature [14]. Furthermore, multiple VNFs may be deployed in the same underlying physical server, consuming and competing the limited resources. Some resource isolation technologies [15] were developed for the resource provisioning in virtual network environment, but the practical deployment of resource isolation technologies highly depends on the degree of reliability requirement, and may not be fully implemented in the network [16]. To address this issue, analytical models are a promising alternative to quantify the effects of resource competition on service provisioning where resource isolation technologies are not deployed or fully functional due to the low reliability circumstance.

To the best of our knowledge, modelling and analysis of NFV networks with the features of dynamic service and scalable resource have not been reported in the existing literatures. In order to fully harvest the merits of NFV for network operators, there is an urgent need and opportunity to use the probabilistic features of SNC to investigate the performance of NFV networks. To bridge this gap, this paper proposes a novel analytical model for NFV networks, 1) to capture the on-demand and dynamic features of NFV services by characterising the traffic and service processes of NFV networks, and 2) to develop an efficient method to investigate the end-to-end performance of the service chain with the input of the SLAs requirements. The main contributions of this paper are summarised as follows:

- A novel analytical model is designed based on SNC to calculate the delay bound of the NFV networks, which can be used as a cost-effective tool for network operators to analyse the performance of NFV architecture.
- To investigate the features of resource competition and on-demand of NFV service chains, the property of convolution associativity and leftover service technologies are exploited to calculate the available resources for VNF nodes and the equivalent system for the entire NFV chain.
- To validate the accuracy of the developed analytical model, both the numerical analysis and extensive simulation experiments are conducted under various network configurations. The results show that the proposed analytical model can generate accurate performance prediction for NFV networks.
- To illustrate its applications, the proposed analytical model is used as an effective performance evaluation tool to investigate the effects of the number of VNF nodes, cross traffic and violation error probability on the performance of NFV networks. Based on these, further optimisations on the resource allocation are devised.

The organisation of this paper is summarised as follows. Related work is provided in Section II. Section III presents the working mechanism of NFV chain, fundamental information of SNC, and an abstracted NFV chain model. Then, Section IV describes the methodology to derive the end-to-end performance bound for NFV under different packet arrival patterns and service processes. Section V evaluates the accuracy of the developed analytical model with both numerical and simulation results and conducts thorough performance analysis based on the proposed model. Finally, Section VI concludes this study.

II. Related Work

Due to its profound impact on the future networks, NFV has gained global awareness [2] [5] [17] [18] [19], leading to a lot of research results reported in the literatures. For instance, to tackle the problem of the optimal VNF placement, Xia *et al.* [20] formulated this problem as a binary integer programming for minimising the overall optical/electronic conversions and designed a heuristic algorithm to achieve high computation efficiency. Moens *et al.* [21] addressed the similar problem through formulating the placement problem as an integer linear program with the aim of minimising the number of servers needed. To reduce the complexity of NFV placement in the practical environment, a Mixed-integer Linear Programming (MILP) optimisation model was formulated and solved in [22] to enable mobile service operator to customise the service deployment and meet the requirements of end users. Laghrissi *et al.* [23] presented a novel network slice planner tool for spatio-temporal simulation and proposed a predictive VNF placement strategy to increase the QoS and reduce the overload of virtual machines in 5G service usage. Miloud *et al.* [24] considered the types and requirements of services as key metrics in the creations of VNF instances, *e.g.* Packet Data Network Gateways (PDN-GW), and formed and solved the optimisation problem for the service-aware VNF deployment.

In order to improve the performance of NFV networks, Ge *et al.* [5] conducted a comprehensive research to show that industry standard servers may not provide satisfied performance for some network functions, such as DPI and Network Address Translation (NAT), and highlighted that hardware acceleration technology is an alternative method for improving the performance of VNFs. Raza *et al.* [2] discovered that the implementation of network functions for IP Multimedia Subsystem (IPMS) in the virtualised platform would occur high latency due to the feedback loop among different VNFs. To address the issue of performance degradation, Yamazaki *et al.* [8] deployed a virtual DPI on Application Specific

Instruction-set Processor (ASIP) and achieved better performance. Compared with the ASIP solution, Intel in [7] proposed a set of Data Plane Development Kit (DPDK) to accelerate the packet processing speed in the VNFs, which has been adopted by many vendors. In addition, China Mobile in [25] conducted extensive field trails to test the performance of NFV and Software Defined Networking (SDN) architectures in C-RAN. The results in [25] showed that common hardware enabled by SDN and NFV technologies can support multi-Radio Access Technology (RAT) and achieve improved performance. The authors in [26] investigated an optimisation solution of configuring SDN-enabled mobile networks and optimising resource allocation to guarantee the required end-to-end QoS. Furthermore, to realise the backward compatibility, a hybrid framework was proposed in [27], which combines both the programmable hardware infrastructure and traditional software infrastructure in NFV architecture. Existing research work has made a lot of progresses on the performance improvement of NFV networks. However, based on the simulation trials, existing studies can hardly be deployed to provide accurate and quantitative performance investigation on NFV networks, especially with large-scale and complex network configuration scenarios.

Modelling and analysis of the network functions and service have been very challenging and attracted tremendous research efforts [12] [13]. Because analytical models can capture the inherent features of a complex system and yield significant insights into the system design and performance in a cost-effective way. For performance modelling and analytics, most of the existing studies resort to the queueing theory [28] [29] [30] [31] [32], which is a fundamental mathematical approach to capture the behaviour of the network system. For instance, Fahmin *et al.* [32] designed an analytical model for NFV-enabled SDN architecture, where VNFs are deployed aside to SDN controller; the network devices in the proposed architecture are modelled as M/M/1 queueing system and the average delay for NFV packets is achieved and validated by the simulation results. Miao *et al.* [33] designed an analytical model for SDN network with priority queues in data plane and the queues are modelled as MMPP/M/1 system. Alfio *et al.* [34] developed an analytical model based on queueing theory to calculate the average latency. The proposed analytical model takes into account both the functions requested by the traffic flows and the QoS that the network could provide. The classical queueing theory focuses on the average quantities in the equilibrium and derives the average system performances such as latency, throughput and packet loss probability. However, due to the dynamic and on-demand features of the service provisioning, it is hardly to use classical queueing theory to investigate the performance of the service-guaranteed system. In this area, SNC [9] [10] [11] has attracted many research interests. Instead of giving the average performance metrics, it derives the worst-case of system performance. For instance, the authors [35] presented a hybrid scheduling model for SDN architecture and applied SNC to derive an analytical model to evaluate the QoS metrics of the proposed scheme. In order to achieve the available bandwidth, Lubben *et al.* [12] explored the properties of stochastic min-plus linear system theory, which expresses bandwidth availability in terms of bounding functions with defined violation probability, and accurately estimated the available bandwidth with random servers in wireless communication systems. While, this work only focuses on modelling the service capability of the access network, lacking the support for modelling the end-to-end service provisioning. As a result, a new model was developed in [13] for characterising the service capability of the converged networks, from access network, core network, to datacenter networks. The analysis technique was developed with the aim of obtaining the delay metric of the SDN-enabled NFV infrastructure and being agnostic to the specific network implementation. However, this work did not consider the on-demand feature and resource competition of the VNF service provision; especially, when the service capabilities of VNFs are different due to their different network functions and also may change over the whole life-cycle in an on-demand manner. Furthermore, the network functions in NFV chains may be added and removed in order to continuously offer satisfied performance and form new NFV chains. To the best of our knowledge, modelling and analysis of NFV networks with the features of dynamic service and scalable resource have not been reported in the existing literature.

III. Preliminaries

This section firstly presents the working mechanism of the ETSI NFV architecture with a focus on the management and implementation of NFV service chains. Then we describe the basic ideas of SNC and present how to leverage its inherent features to abstract and model NFV networks. Finally, a subsection is provided to show the closeness of the SNC to VNF performance analysis and deployment. Details of the performance derivation will be discussed in Section V.

A. NFV Service Chain

ETSI network architecture [36] gives a guide on how to design and implement the NFV architecture. The key concept in ETSI NFV architecture is the NFV service function chain, which defines a list of network functions and connects them to provide network services. To demonstrate how VNFs are deployed in the practical network, we exemplify a cloud-based web service visited by mobile devices, as shown in Fig. 1. We follow the traffic sent by a mobile device and check what kinds of VNF services the traffic may visit and be processed. In this example, the overall transmission network covers the radio access network, the Enhanced Packet Core (vEPC) network, the backbone network, and the datacenter network. As shown in Fig. 1, the mobile device sends a message to the base station through the access network. Based on service type, this message will be processed by various VNFs within vEPC network, including Mobility Management Entity (MME), Home Subscriber Server (HSS), Policy and Charging Rules Function (PCRF), Service Gateway (S-Gateway) and Packet Data Network Gateway (P-Gateway). When the packet leaves the access network and enters the backbone network, the packets may be transmitted or processed by virtual Provider Edge (vPE) and virtual Router Recovery. In backbone network, the

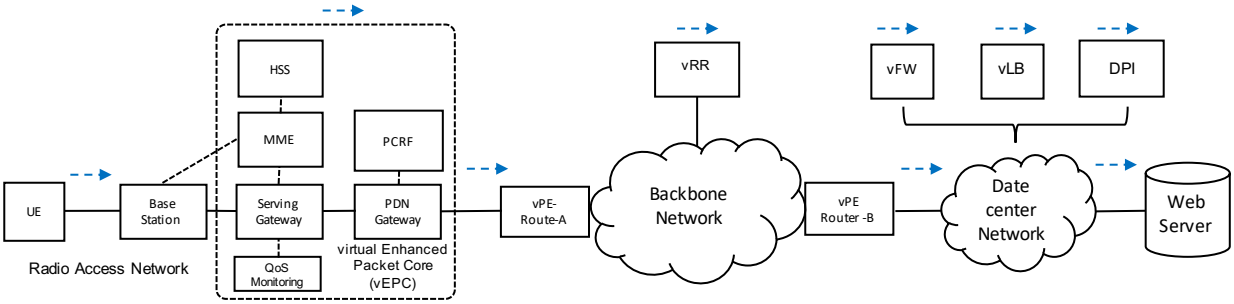

Fig. 1. Example of VNF deployment from access network to datacenter network (User Equipment (UE), Home Subscriber Server (HSS), Mobility Management Entity (MME), Policy and Charging Rules Function (PCRF), Packet Data Network Gateway (PDN GW), virtual Provider Edge (vPE), virtual Router Recovery (vRR), virtual Firewall (vFirewall), virtual Load Balancer (vLB), Deep Packet Inspection (DPI))

packets will be transmitted by Multiprotocol Label Switching (MPLS), which uses path labels instead of IP address for packet routing, avoiding complex table lookups and improving the overall transmission performance. Once the packet leaves the vPE Router-B and arrives at the datacenter network, VNFs, *e.g.* virtual FireWall (vFW), virtual Load Balance (vLB) and Deep Packet Inspection (DPI), would be offered by the service providers to meet certain requirements in terms of SLA, reliability and security.

Through implementing the network function in the form of software, VNFs can be easily added and removed in the service chain at any time during the life-cycle of a service provision. For instance, a DPI service can be installed in the datacenter network when higher security and reliability requirements are needed. Another important merit of NFV is the scalability and flexibility of resource provisioning. For instance, the underlying virtual resources, such as storage spaces and computing resource, can be upgraded to continuously and quickly satisfy the upper-level service requirement. In the Section III-C, we will abstract the working mechanism of NFV service chain into a mathematical model for the performance analysis and investigation.

B. Stochastic Network Calculus

This part introduces the basic principles of SNC, which provides system performance bound with the constraint of violation probability. Let the cumulative traffic generated by the end user as $A(t)$, the cumulative service of the system as $S(t)$, and the cumulative departures from the system as $D(t)$. Assume that the server is lossless and work-conserving, it is readily seen that $A(t)$, $S(t)$ and $D(t)$ are non-negative and non-decreasing functions. Denote $A(\tau,t)$ as the cumulative traffic arrival at time interval $[\tau,t]$, it is equal to $A(t)-A(\tau)$. $S(t)$ and $D(t)$ have the similar notations as $S(\tau,t)$ and $D(\tau,t)$, respectively. The relationship between $A(t)$, $S(t)$ and $D(t)$ is given by

$$D(t) \geq \min_{\tau\in[0,t]} \{A(\tau)+S(\tau,t)\} \tag{1}$$

The right side of the above inequality achieves its minimum when τ is equal to the beginning time of the last busy period before t, denoted as τ^*. At time τ^*, the queueing buffer is empty and $A(\tau^*) = D(\tau^*)$. The backlog at time t can be defined as $B(t) = A(t) - D(t)$, which comprises the packets in the buffer and server. From Eq. (1), $B(t)$ can be described as follows,

$$B(t) \leq \max_{\tau\in[0,t]} \{A(\tau,t) - S(\tau,t)\} \tag{2}$$

The system delay at time t is defined as

$$W(t) = \min\{\upsilon \geq 0, A(t) \leq D(t+\upsilon)\} \tag{3}$$

where υ denotes the time slot between the packet arrival and departure from the server. In addition, SNC uses min-plus convolution, denoted as operator $\otimes$, to simplify the complex equation derivation.

$$(A \otimes S)(t) = \min_{\tau\in[0,t]} \{A(\tau)+S(\tau,t)\} \tag{4}$$

With the definition of min-plus convolution, Eq. (1) can be described as $D(t) \geq (A \otimes S)(t)$. Having some useful properties, *e.g.* associativity and distributivity, min-plus convolution is very useful for deriving the end-to-end network performance.

C. NFV Model Abstraction

To quantitatively investigate the performance of the NFV network, this section abstracts the complex service chain in Fig. 1 into a mathematical model as shown in Fig. 2. In the abstracted NFV model, the arrival traffic for each server consists of two parts of traffic: the traffic of the interesting NFV service chain (denoted as the through traffic), and the competing traffic of other NFV service chains (denoted as cross traffic). As multiple NFV service chains may share the physical server, the cross traffic is used to capture the effect of resource competition.

There are n servers in the abstracted model. The first server in the abstracted model is set to be the memory-less on-off server to represent the wireless communication channel [37]. The other $n-1$ servers model the packet processing of VNFs located in the access, core, backbone and datacenter networks. For the ith server, let $A_{th_i}(t)$ and $A_{cr_i}(t)$ denote the through traffic and the cross traffic, respectively. As the SLA requirements for service provisioning should be guaranteed by the network operators, this work assumes the transmission bandwidth between two VNFs in NFV chain can be guaranteed by the underlying NFV Infrastructure (NFVI), *e.g.* QoS differentiation and tag technologies.

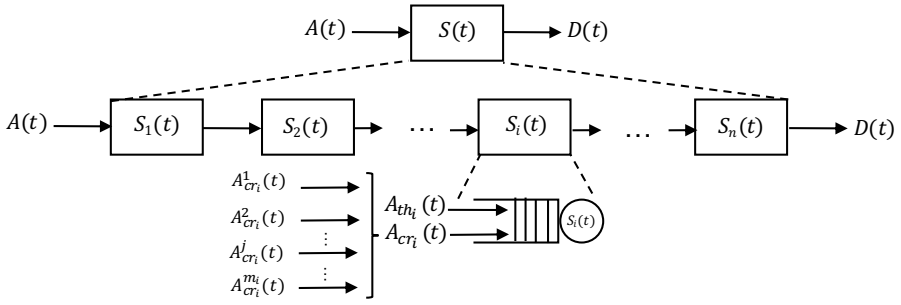

Fig. 2. End-to-End NFV Chain

D. Closeness of the SNC to VNF Deployment and Performance Analysis

By leveraging SNC method, the proposed analytical model is capable of capturing three key features of NFV deployment, including the on-demand service provisioning, the resource competition among multiple NFV service chain and the SLA-oriented service guarantee.

- Resource Competition

For practical NFV deployments, multiple VNFs may share the same underlying physical or virtual servers in order to maximise the resource utilisation. However, this co-existence deployment strategy poses a serious issue for service quality guarantee. Therefore, the negative impact of the cross traffic on the performance of NFV chain should be considered in the analytical model. Within the abstracted model, the cross traffic is modelled as $A_{cr_i}(t)$, and the number of the cross traffic flow for the ith server is m_i. Let $A_{cr_i^j}(t)$ denote the jth cross traffic flow in the ith server. The traffic rate of $A_{cr_i^j}(t)$ is λ_i^j. By using the stochastic multiplexing and leftover service technologies of SNC, we can combine m_i cross traffic into an equivalent aggregation cross traffic and calculate the available resource of the through traffic in the present of the cross traffic. In Section V-B, we will investigate the impacts of m_i and λ_i^j on the end-to-end performance bound of NFV service chain.

- Dynamic and On-demand Deployment

The second issue for modelling NFV networks is how to capture the dynamics of the NFV chain, including the processing ability of each VNF and and the operations of adding or removing specific VNFs during the service provisioning. For SNC, the VNF node in Fig. 2 is modelled as a random server and the service process is denoted as $S_i(t)$. The capacity of each server can be modified in an on-demand manner. In addition, by exploiting the associativity property of the min-plus convolution, the multiple nodes within an NFV chain can be transferred to a single equivalent system. Let $S(t)$ denote the service process of the whole NFV chain. By leveraging the associativity property, $S(t)$ can be expressed as the convolution of the individual service process, shown as $S(t) = S_1 \otimes S_2 \otimes \cdots \otimes S_n$. The end-to-end system performance is a function of n, which is a free parameter during the whole performance derivation. In case that a VNF is removed from or added into the service chain, the analytical model can immediately capture this effect and quickly calculate the new performance bounds for the modified service chain by updating the value of n. It is noted that the operation of min-plus convolution does not pose the requirement for the independence among service processes. The details of the performance derivation will be studied in Section IV.

- SLA-oriented Service Guarantee

For service operators, the NFV services are designed and offered in the manner of SLA constraints. Strict requirements will be applied to the service availability and violation. Therefore, analytical models to be designed are expected to derive the system performance regarding SLA guarantee, instead of the basic average QoS metrics only. Therefore, SNC is suitable to investigate the performance of NFV service chains. SNC exploits probability theory and mathematically derives the relationship between the SLA requirement and the probability that this demand cannot be met under the network resource constraints. According to [10], this relationship is defined as follows,

$$P(W(t) > w) \leq \epsilon \tag{5}$$

where w denotes the latency bound for the end-to-end service chain and ϵ is the violation probability that the required latency bound cannot be guaranteed. Therefore, the aim of analytical modelling is to calculate w if ϵ is given, and vice versa.

IV. Performance Derivation

In this section, we will present the methodology to derive the end-to-end latency performance bound for the abstracted NFV model in Fig. 2. To capture the dynamic feature of various traffic arrival in NFV chain, a unified representation method is presented in Subsection IV-B, which could transfer different kinds of arrival traffic into an unified representation in the form of Stochastic Arrival Curves (SACs). Subsection IV-C presents a stochastic multiplexing approach to calculate the aggregation traffic for multiple traffic processes. Subsection IV-D derives the Stochastic Service Curves (SSCs) for different service processes and exploits the leftover technology to calculate the effective resources for the interesting traffic in the presence of multiple competing traffic. After achieving the effective SSCs, a single equivalent system is envisioned and the end-to-end performance bound is calculated in subsection IV-E.

During the performance derivation, different kinds of traffic and service processes are adopted in this study. For traffic

TABLE I
KEY NOTATIONS IN THE PERFORMANCE DERIVATION

Notations	Definitions
$A(t)$, $A(\tau, t)$	The cumulative traffic arriving at the server during the time interval $[0, t]$ and $[\tau, t]$
$S(t)$, $S(\tau, t)$	The cumulative service offered by the server during the time interval $[0, t]$ and $[\tau, t]$
$D(t)$, $D(\tau, t)$	The cumulative departures from the server during the time interval $[0, t]$ and $[\tau, t]$
$S_i(t)$	The cumulative service offered by the ith server during the time interval $[0, t]$
$S_{th_i}(t)$	The leftover cumulative service offered by the ith server for the through traffic $A_{th_i}(t)$
$A^j_{cr_i}(t)$, $A_{cr_i}(t)$, $A_{th_i}(t)$	The jth cumulative cross traffic, the total cumulative cross traffic, and the cumulative through traffic arriving at the ith server during the time interval $[0, t]$
υ	The time slot between the packet arrival and departure from the server
$W(t)$	The delay process
m_i	The number of the cross traffic at the ith server
n	The number of the VNF nodes
λ^j_i	The rate of the jth cross traffic at the ith server
w	The latency bound for the end-to-end service chain
ϵ	The violation probability that the required latency bound cannot be guaranteed
θ	The free parameter
ρ	The slope of an affine envelope
b	The burst parameter of an affine envelope
$N^j_{cr_i}(\tau, t)$	The numbers of the bits at the ith server from the jth cross traffic during the time interval $[\tau, t]$
$M_A(\theta, t-\tau)$	The Moment Generating Function of $A(\tau, t)$
$M_{N^j_{cr_i}}(\theta, t-\tau)$	The Moment Generating Function of $N^j_{cr_i}(\tau, t)$
$M_{S_i}(-\theta, t-\tau)$	The Moment Generating Function of $S_i(\tau, t)$
φ_1	The transmission rate from state 1 to state 2 in MMPP process
φ_2	The transmission rate from state 2 to state 1 in MMPP process
λ	The arrival rate when Markov chain is in state 1
r	The rate when the channel is in the state "on"
P_{on}	The probability that the state of the wireless channel is in the state "on"
τ^*	The last busy time for the server

characterisation, as the existing studies have shown the bursty and correlated nature of network traffic, the analytical model is developed with both bursty and non-bursty network traffic in NFV networks. Firstly, Poisson process is exploited to model the non-busty network traffic in NFV network, *e.g.* FTP connection inter-arrivals. Then an analytical model is developed with the bursty and correlated network traffic modelled by Markov Modulated Poisson Process (MMPP). For the service process, to be consistent with the existing work on NFV analytical modelling [32], we used exponential distribution process to model the normal NFV service process. However, to obtain an end-to-end NFV deployment, from access, core, backbone to datacenter networks, we need to consider the impact of wireless channel on the end-to-end delay bound and Markov-modulated off-on process is used to model wireless channels.

A. Moment Generating Function and Exponentially Bounded Burstiness

In order to achieve the latency bound with the constraint of SLA requirement, two functions named Moment Generating Function (MGF) and Exponentially Bounded Burstiness (EBB) are used in this subsection to establish the link between the latency bound and SLA requirement. Based on the statistic theory of SCN [38], the MGF of an arrival process, $A(t)$, is defined as $M_A(\theta, t-\tau) = E[e^{\theta A(\tau,t)}]$, where θ is a free parameter with the constraint of time τ. The MGF of the arrival process is bounded by

$$E[e^{\theta A(\tau,t)}] \leq e^{\theta((\rho(t-\tau)+\sigma)} \tag{6}$$

where ρ and σ are the functions of the free parameter θ.

EBB is a stochastic envelop for the arrival process, $A(t)$. Given the violation probability, $\epsilon(b)$, EBB calculates the probability that cumulative traffic process, $A(\tau, t)$, is larger than an affine envelop $\rho(t-\tau)+b$. ρ and b are the slope and bursty of the affine envelop respectively. EBB is defined as,

$$P(A(t,\tau) > \rho(t-\tau)+b) \leq \epsilon(b) \tag{7}$$

where $\epsilon(b)$ is defined as an exponential function of b, given by $\epsilon(b) = \alpha e^{-\theta b}$. The SNC exploits Chernoff bound to link the EBB and MGF and provides the end-to-end stochastic performance bounds with given SLA requirements.

In the statistics [39], the generic Chernoff bound for a random variable, X, is defined as

$$P(X > x) = P(e^{\theta X} > e^{\theta x}) \leq \frac{E[e^{\theta X}]}{e^{\theta x}} \tag{8}$$

By applying the Chernoff bound in Eqs. (6) and (7), the EBB can be rewritten as

$$\begin{aligned} P(A(t,\tau) > \rho(t-\tau)+b) \leq \epsilon(b) &\leq \frac{E[e^{\theta A(t,\tau)}]}{e^{\theta[\rho(t-\tau)+b]}} \\ &\leq \frac{e^{\theta[\rho(t-\tau)+\sigma]}}{e^{\theta[\rho(t-\tau)+b]}} = e^{\theta\sigma}e^{-\theta b} \end{aligned} \tag{9}$$

Combining Eq. (7) and Eq. (9), $\epsilon(b)$ is equal to $e^{\theta\sigma}e^{-\theta b}$. This subsection presents the method to derive the statistical performance bound for the arrival process. The service statistical performance bounds can be obtained following the methods shown in [9].

B. Stochastic Arrival Curves

To capture the dynamic and on-demand NFV features, both the non-bursty traffic and the bursty traffic are simultaneously considered in this study. Poisson process is adopted to model the non-bursty traffic [40] and MMPP is used to model the bursty traffic [41]. This section will derive the SACs for Poisson process and MMPP process.

- Poisson Traffic

Let $A^j_{cr_i}$ and $N^j_{cr_i}$ denote the numbers of the packets and bits arriving at the ith server from the jth cross source during the time interval $[0, t]$. According to [10], the packet size is defined as $1/v$, which brings convenience for mathematics derivation. Then $A^j_{cr_i} = N^j_{cr_i}/v$. Let $M_{N^j_{cr_i}}(\theta, t-\tau)$ denote the MGF of $N^j_{cr_i}$. Then the MGF of $A^j_{cr_i}$, can be calculated by:

$$\begin{aligned} M_{A^j_{cr_i}}(\theta, t-\tau) = E[e^{\theta A^j_{cr_i}}] &= E[e^{\theta N^j_{cr_i}/v}] \\ &= M_{N^j_{cr_i}}(\theta/v, t-\tau) \end{aligned} \tag{10}$$

In [38], the distribution of a Poisson process is $P[N(t) = k] = e^{-\lambda t}(\lambda t)^k/k!$, where λ is the average arriving rate. Inserting the Poisson distribution into Eq. (10), the MGF of $N^j_{cr_i}$ can be obtained as

$$M_{N^j_{cr_i}}(\theta, t-\tau) = E[e^{\theta N^j_{cr_i}}] = \sum_{k=0}^{+\infty}[(\lambda^j_i t)^k/k!]e^{-\lambda^j_i t}e^{\theta k} = e^{-\lambda^j_i t(e^\theta - 1)} \quad (11)$$

where λ^j_i is the arrival rate for the jth NFV cross traffic source at the ith server. Then the MGF and affine envelop model of $A^j_{cr_i}(t)$ are $M_{A^j_{cr_i}}(\theta, t-\tau) = e^{-\lambda^j_i t(e^\theta-1)}$ and $E[e^{\theta A^j_{cr_i}(t-\tau)}] \leq e^{\theta[\rho^j_i(t-\tau)+\sigma^j_i]}$, with the parameters $\rho^j_i = \frac{\lambda^j_i(e^{\theta/\nu}-1)}{\theta}$ and $\sigma^j_i = 0$.

- Markov Modulated Poisson Process (MMPP) Traffic

MMPP process has two transmission rates, the transmission rate from state 1 to state 2, φ_1 and the transmission rate from state 2 to state 1, φ_2; when Markov chain is in state 1, the arrival rate for through traffic is λ; when Markov chain is in state 2, there is no arrival for through traffic.

The MGF of $A_{th}(t)$ is defined as $M_{A_{th}}(\theta, t-\tau) = E[e^{\theta A_{th}}]$ According to [11], the $M_{A_{th}}(\theta, t-\tau)$ can be calculated by

$$E[e^{\theta A_{th}}] = \left[\frac{\varphi_2}{\varphi_1+\varphi_2} \ \frac{\varphi_1}{\varphi_1+\varphi_2}\right] exp\left(\begin{bmatrix}-\varphi_1+\theta\lambda & \varphi_1\\ \varphi_2 & -\varphi_2\end{bmatrix}t\right)\begin{bmatrix}1\\1\end{bmatrix} \quad (12)$$

From [42], Eq. (12) can be simplified as

$$E[e^{\theta A_{th}}] \leq e^{\theta\lambda-\varphi_1-\varphi_2+\sqrt{(\theta\lambda+\varphi_2-\varphi_1)^2+4\varphi_1\varphi_2}} \quad (13)$$

Furthermore, the affine envelop of the arrival process, $A_{th}(t)$, is defined as,

$$E[e^{\theta A_{th}(\tau,t)}] \leq e^{\theta(\rho_{a_{th}}(t-\tau)+\sigma_{a_{th}})} \quad (14)$$

After inserting Eq.14 in Eq.13, $\rho_{a_{th}}$ and $\sigma_{a_{th}}$ can be calculated as $\rho_{a_{th}} = \frac{1}{2\theta}(\theta\lambda-\varphi_1-\varphi_2+\sqrt{(\theta\lambda+\varphi_2-\varphi_1)^2+4\varphi_1\varphi_2})$ and $\sigma_{a_{th}} = 0$.

C. Stochastic Multiplexing

For the ith server, there are m_i cross traffic competing the resources with the through traffic. To analyse the impact of the cross traffic on the performance of the through traffic, a stochastic multiplexing approach [9] is tailored in this subsection to combine m_i cross traffic into one equivalent aggregation traffic. This simplifies the complex model derivation and provides an efficient solution for calculating the leftover service in Subsection IV-D. For a lossless system, let $A_{Agg_i}(t)$ denote the sum of $A^j_{cr_i}(t)$, $A_{Agg_i}(t) = \sum_{j=1}^{m_i} A^j_{cr_i}(t)$. Based on this equation, it can readily obtain the relationship between the MGFs of the individual cross traffic process and the aggregated traffic process.

The MGF of the aggregated traffic, $M_{A_{Agg_i}}(\theta, t-\tau)$, is given by

$$M_{A_{Agg_i}}(\theta, t-\tau) = E[e^{\theta A_{Agg_i}(t-\tau)}] = E[e^{\theta\sum_{j=1}^{m_i}A^j_{cr_i}(t-\tau)}] = \prod_{j=1}^{m_i} e^{\theta A^j_{cr_i}(t-\tau)} \quad (15)$$

Given the affine envelop model of $A^j_{cr_i}(t)$, $E[e^{\theta A^j_{cr_i}(t-\tau)}] \leq e^{\theta[\rho^j_i(t-\tau)+\sigma^j_i]}$, $M_{A_{Agg_i}}(\theta, t-\tau)$ can be rewritten in the following equation,

$$M_{A_{Agg_i}}(\theta, t-\tau) \leq e^{\theta[\rho_{Agg_i}(t-\tau)+\sigma_{Agg_i}]} \quad (16)$$

where $\rho_{Agg_i} = \prod_{j=1}^{m_i}\rho^j_i$ and $\sigma_{Agg_i} = \prod_{j=1}^{m_i}\sigma^j_i$;

From the perspective of the ith server, the m_i cross traffic can be regarded as an aggregation traffic, $A^j_{cr_i}(t)$, with parameters ρ_{Agg_i} and σ_{Agg_i}, which are calculated from individual cross traffic, $A^j_{cr_i}(t)$.

D. Stochastic Service Curves

The SSC gives the bound of the least service available in the system for the arrival traffic. Two kinds of service models are envisioned in the abstracted NFV model, memoryless on-off service process and exponential service process. Memoryless on-off process is used to model the dynamic wireless channel [42] and exponential process is for normal NFV server [43]. This section investigates how to achieve the SSCs for both the on-off service process and exponential process. To investigate the impact of the cross traffic on the performance of through traffic, an efficient approach to calculate the leftover service is exploited in this subsection to obtain the effective resource for the through traffic.

- On-off Wireless Channel Model

As mentioned in Subsection III-C, the wireless channel is modelled as an on-off server. The on-off wireless channel has two states ("on" and "off") during service provisioning. When the channel is in the "on" state, it provides a constant transmission at rate r. When the channel is in the "off" state, the wireless channel does not provide any service. Let P_{on} denote the probability that the wireless channel is in the state of "on"; and $1-P_{on}$ denote the probability that wireless channel is in the state of "off". Let $X(t)$ be the service available at time interval t. For an on-off wireless channel, $X(t)$ is a two states Bernoulli variable, $P(X(t) = r) = P_{on}$ and $P(X(t) = 0) = 1-P_{on}$. MGF of $X(t)$, $M_X(-\theta, t)$, can be expressed as,

$$M_X(-\theta, t) = E[e^{-\theta X(t)}] = P_{on} * e^{-\theta r} + 1 - P_{on} \quad (17)$$

Let $S_1(\tau, t)$ denote the cumulative service in time interval $[\tau, t]$. For a lossless server, the cumulative service should be equal to the sum of the $X(t)$ over time interval $[\tau, t]$, expressed as $S_1(\tau, t) = \sum_{\upsilon=\tau}^{t} X(\upsilon)$. Then the MGF of $S_1(\tau, t)$, $M_S(-\theta, t)$, can be achieved by,

$$M_S(-\theta, t) = E[e^{-\theta S(t)}] = E[e^{-\theta\sum_{\upsilon=1}^{t}X_\upsilon}] \quad (18)$$

As $A(t)$ follows iid Bernoulli distribution, Eq. (18) could be rewritten as

$$M_S(-\theta, t) = \{E[e^{-\theta X(t)}]\}^t = \{M_X(-\theta, t)\}^t = (P_{on} * e^{-\theta r+1-P_{on}})^t = e^{-\theta t\frac{ln(P_{on}*e^{-\theta r}+1-P_{on})}{-\theta}} \quad (19)$$

Then, $S_1(\tau, t)$ has the MGF of $E[e^{-\theta S_1(t-\tau)}] \leq e^{-\theta[\rho_{S1}(t-\tau)-\sigma_{S1}]}$ with the parameter $\rho_{S1} = \frac{ln(P_{on}*e^{-\theta r}+1-P_{on})}{-\theta}$ and $\sigma_{S1} = 0$.

- Exponential Service Model

According to [11], the MGF of service process, $S_i(\tau,t)$, is defined as $M_{S_i}(-\theta,t-\tau)=E[e^{-\theta S_i(\tau,t)}]$ with parameter $-\theta$ for $\theta>0$. Considering the affine service envelop, $S_i(\tau,t)\geq\rho(t-\tau)-b_s$, the MGF of service process, $M_{S_i}(-\theta,t-\tau)$, can be calculated by,

$$E[e^{-\theta S_i(\tau,t)}]\leq e^{-\theta(\rho_{s_i}(t-\tau)-\sigma_{s_i})} \tag{20}$$

Similar to the derivation process of the SACs, the MGF of service process can be achieved by,

$$M_{N_{s_i}}(-\theta/v,t-\tau)=e^{-\mu_i t(e^{-\theta}-1)} \tag{21}$$

where μ_i is the service rate for the ith server. Then $S_i(t)$ has the MGF, $M_{S_i}(-\theta,t-\tau)$, to be $e^{-\mu_i t(e^{-\theta}-1)}$, and the affine envelop model is $E[e^{-\theta A_{S_i}(t-\tau)}]\leq e^{-\theta[\rho_{s_i}(t-\tau)+\sigma_{s_i}]}$, with the parameters $\rho_{s_i}=\frac{\mu_i(e^{-\theta/v}-1)}{-\theta}$ and $\sigma_{s_i}=0$.

- Leftover Service

For each server, we are interested in the service available for the through traffic, not for the cross traffic. Therefore, this subsection will calculate the amount of the service capacity offered to the through traffic in the presence of the cross traffic, named leftover service. Let $A_{t_i}(t)$ be the total cumulative arrival at the ith server, which consists of the through traffic, $A_{th_i}(t)$, and the total cross traffic, $A_{Agg_i}(t)$. Let $D_{t_i}(t)$ be the total cumulative departure from the ith server, which consists of the departures from through traffic, $D_{th_i}(t)$, and the total departures from the cross traffic, $D_{Agg_i}(t)$. Let $S_i(\tau,t)$ denote the service available at the ith server during the time interval $[\tau,t]$. Based on the definition of Eq. (1), the following inequality holds for $S_i(\tau,t)$, $A_{t_i}(\tau,t)$, and $D_{t_i}(\tau,t)$,

$$D_{t_i}(t)\geq\min_{\tau\in[0,t]}\left(S_i(\tau,t)+A_{t_i}(t)\right) \tag{22}$$

Assume that the last busy time is τ^*. When $\tau=\tau^*$, the right side of the inequality achieves the minimum. Inserting $A_{th_i}(t)$, $A_{Agg_i}(t)$, $D_{th_i}(t)$, and $D_{Agg_i}(t)$ in Eq. (22), it can be written as

$$D_{th_i}(t)\geq A_{th_i}(\tau^*)+\{S_i(\tau^*,t)-[D_{Agg_i}(t)-A_{Agg_i}(\tau^*)]\}_+ \tag{23}$$

where the symbol "+" means nonnegative value. For the cross traffic, the departure can not be larger than the arrival, denoted as $D_{Agg_i}(t)\leq A_{Agg_i}(t)$. Eq. (23) can be further rewritten as

$$D_{th_i}(t)\geq A_{th_i}(\tau^*)+\{S_i(\tau^*,t)-[A_{Agg_i}(t)-A_{Agg_i}(\tau^*)]\}_+ \tag{24}$$

In Eq. (24), $A_{Agg_i}(t)-A_{Agg_i}(\tau^*)=A_{Agg_i}(\tau^*,t)$; Since τ^* denotes the last busy time that can be determined in advance, we use minimum operation to generalise τ^* as τ. Eq. (24) is transferred to

$$D_{th_i}(t)\geq\min_{\tau\in[0,t]}\{[S_i(\tau,t)-A_{Agg_i}(\tau,t)]_+ + A_{th_i}(\tau)\} \tag{25}$$

For the through traffic, the following inequality holds for $A_{th_i}(t)$, $S_{th_i}(t)$, and $D_{th_i}(t)$,

$$D_{th_i}(t)\geq\min_{\tau\in[0,t]}\{S_{th_i}(t)+A_{th_i}(\tau)\} \tag{26}$$

where $S_{th_i}(t)$ is the leftover service for the through traffic in the presence of the cross traffic. From Eq. (25) and Eq. (26), the leftover service, $S_{th_i}(t)$, is calculated as $S_{th_i}(t)=[S_i(\tau,t)-A_{Agg_i}(\tau,t)]_+$.

Under the assumption that $S_{th_i}(t)$ and $A_{Agg_i}(\tau,t)$ are stochastically independent, the MGF of the leftover service, $M_{S_{th_i}}(-\theta,t-\tau)$, can be calculated by

$$\begin{aligned}M_{S_{th_i}}(-\theta,t-\tau)&=E[e^{-\theta S_{th_i}(t-\tau)}]\\&=E[e^{-\theta[S_i(\tau,t)-A_{Agg_i}(\tau,t)]_+}]\\&\leq e^{-\theta[\rho_{S_i}(t-\tau)-\sigma_{S_i}]}e^{\theta[\rho_{Agg_i}(t-\tau)+\sigma_{Agg_i}]}\\&=e^{-\theta[(\rho_{S_i}-\rho_{Agg_i})(t-\tau)-(\sigma_{S_i}+\sigma_{Agg_i})]}\end{aligned} \tag{27}$$

Then, the MGF of $S_{th_i}(t)$ is calculated by $E[e^{-\theta S_{th_i}(t-\tau)}]\leq e^{-\theta[\rho_{th_i}(t-\tau)-\sigma_{th_i}]}$, with the parameters $\rho_{th_i}=\rho_{S_i}-\rho_{Agg_i}$ and $\sigma_{th_i}=\sigma_{S_i}+\sigma_{Agg_i}$.

E. End-to-end Latency Bounds

From Subsections IV-B and IV-D, we have achieved the envelop models for the arrival process and service process. For the through traffic, the envelop is shown as,

$$A_{th}(\tau,t)\leq(\rho_{A_{th}}+\sigma)(t-\tau)+b_{A_{th}} \tag{28}$$

The probability that $A_{th}(\tau,t)$ cannot satisfy the above inequality is equal to $\epsilon(b_{A_{th}})=e^{\theta\sigma_{A_{th}}}e^{-\theta b_{A_{th}}}/(1-e^{-\theta\delta})$. For the service process of the entire NFV chain, the envelop is shown as,

$$S(\tau,t)\geq(\rho_S-\sigma)(t-\tau)-b_S \tag{29}$$

The probability that $A_{th}(\tau,t)$ cannot meet the above inequality is equal to $\epsilon(b_S)=e^{n\theta\sigma_S}/(1-e^{-\theta\sigma})^n$.

For the NFV system stability, it is required that the service envelop should be always larger than the arrival envelop, $\rho_{A_{th}}+\delta<\rho_S-\delta$, and $\delta<(\rho_S-\rho_{A_{th}})/2$. This means that the slope of the service envelop should be larger than that of the arrival envelop.

For SNC, the end-to-end upper bound latency, w, is defined as

$$W(t)\leq\min\{w\geq0:\max\{A_{th}(\tau,t)<D(\tau,t+w)\}\} \tag{30}$$

With the MGFs of the traffic process and overall service process, the upper bound latency in Eq. (30) can be calculated as:

$$w=\inf_{\theta>0}\left[\tau:\frac{1}{\theta}ln\left(\sum_{s=\tau}^{\infty}M_{A_{th}}(\theta,s-\tau)M_S(-\theta,s)\right)-ln\epsilon\leq0\right] \tag{31}$$

Under the FIFO scheduling strategy, the work in [44] gave the approach to solve the above inequality.

The end-to-end delay bound, w, can be achieved when τ satisfies the following condition:

$$\frac{1}{\theta}ln\left(\sum_{s=\tau}^{\infty}M_{A_{th}}(\theta,s-\tau)M_S(-\theta,t-\tau)\right)-ln\epsilon\leq0 \tag{32}$$

Under the stability condition, $\rho_{A_{th}}(\theta)\leq\rho_S(-\theta)$, τ has the following latency bound,

$$\tau\geq\frac{\sigma(\theta)}{\rho_S(-\theta)}+\frac{n*ln\gamma-ln\epsilon}{\theta\rho_S(-\theta)} \tag{33}$$

where γ is calculated by $\gamma=1+\frac{1}{1-e^{-\theta(\rho_S(-\theta)-\rho_{A_{th}}(\theta))}}$.

Finally, the end-to-end upper latency bound is achieved as $w = inf_{\theta>0}[inf[\tau]]$; and the backlog bound is $b = inf_{\theta>0}[\sigma(\theta) + (ln\gamma - ln\epsilon)/\theta]$.

V. Performance Validation and Analysis

In this section, we first evaluate the accuracy of the developed analytical model; that is then used to study the effects of the number of VNF nodes, the arrival rates of cross traffic and violation probability on the end-to-end performance of NFV networks, with the aim of obtaining the fundamental understanding of the performance of NFV networks.

A. Performance Evaluation

Following the approach to validate the analytical model in [45], the accuracy of the proposed performance model is firstly evaluated through the comparison with the exact queueing results subject to the non-bursty traffic [10][42][46]. Furthermore, we conducted comprehensive simulation experiments to evaluate the performance of the proposed algorithm with the bursty traffic. Additionally, the baseline parameters in the analysis and simulation are set to be consistent with the work in [32]. For numerically evaluating the accuracy of the developed model, we apply the proposed analytical model to the analysis of two common queueing systems, *i.e.* single and multiple nodes systems. For the single queueing system, Bolch *et al.* [39] provided the relation between the latency bound and violation error rate in the following equation,

$$w = log(\epsilon)/(-\mu_s(1-\rho_s)) \tag{34}$$

where μ_s and ρ_s are the service rate and utilisation of the server. The Eq. (34) is derived based on the assumption that the arrival packets are served by the FIFO scheduling algorithm. The through traffic and cross traffic for the server, $A_{th}(t)$ and $A_{cr_j}(t)$, follow Poisson processes with the arrival rates λ_{th} and λ_{cr_j} respectively. The server provides the independent service for arriving packets, where the service time, X, follows the exponential distribution ($X \sim exp(\mu_s)$). Then, the utilisation rate of the server can be computed by $\rho_s = \left[\sum_{j=1}^{m_j} \lambda_{cr_j} + \lambda_{th}\right]/\mu_s$. In addition, for the multiple node case, Florin in [47] provided the relation between the upper latency bound and violation error as follows,

$$P(W > w) = \left[\sum_{i=0}^{n-1} \frac{\mu_s((1-\rho_s)w)^i}{i!}\right] e^{-\mu_s(1-\rho_s)w} \tag{35}$$

where n is the server number and the latency, w, follows the Gamma distribution $\Gamma(\mu_s(1-\rho_s), n)$.

Next, we use the method presented in Subsection IV-E to calculate the end-to-end latency bound for NFV networks and compare the results with those obtained from the exact bounds in Eq. (34) and Eq. (35). Without loss of generality, according to the baseline values in [32], the system configuration is shown as follows: service rate is set to be 0.095M packets/second; the arrival rates are set from 0 to 0.095M packets/second with an interval of a 10% of service rate; two violation error settings are adopted to reflect different SLA requirements: 10^{-6} and 10^{-4}; the packet size is set to be 1024 bits. As shown in Fig. 3, the upper latency bounds are achieved from both the queueing theory and SNC models by varying the server utilisation rates. From Fig. 3, the results of the analytical model match well with those of the exact theory results. Furthermore, the smaller value of the violation error brings higher upper latency bound; this means that strict SLA requirement in terms of the violation error would push up the bound of the latency with given network resources.

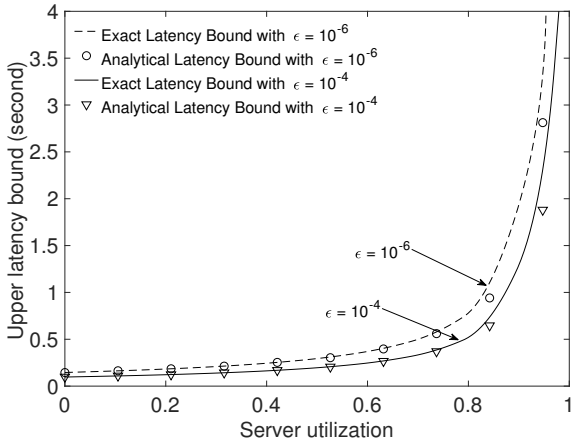

Fig. 3. Comparison of the Exact Theory Results with Those Obtained from the Analytical Model by Varying the Server Utilisation Rates: ϵ = [10^{-4}, 10^{-6}], μ_s=0.095M packets/second, λ_{th}=0-0.095M packets/second, and n=1.

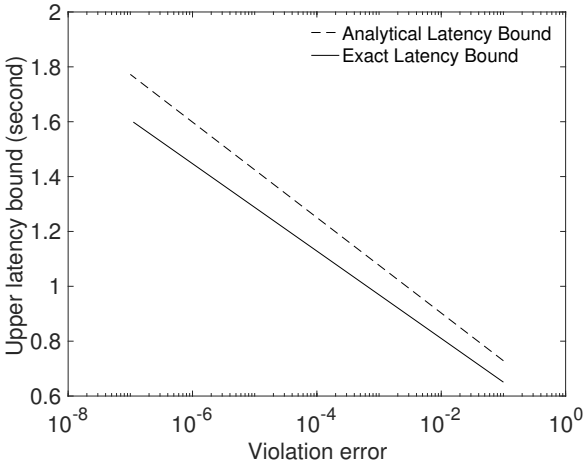

Fig. 4. Comparison of the Exact Theory Results with These Obtained from the Analytical Model by Varying the Violation Error: λ_{th}=0.075M packets/second, μ_s=0.095M packets/second, n=3, and ϵ=10^{-1}-10^{-7}.

In order to evaluate the accuracy of the proposed analytical model under different violation error rates, Fig. 4 shows the upper latency bounds obtained from both Eq. (35) and the analytical model. In Fig. 4, the arrival rate is fixed to 0.075M packets/second; the service rate and packets sizes are set to be the same as those in Fig. 3; the solid curve depicts the results obtained from the analytical model and the dotted curve depicts the results of queue theory in Eq. (35). It can be seen that the proposed analytical model has reasonable accuracy against the theory results. However, there is a relative error of

the proposed analytical model against the exact performance bound as shown in Fig. 5, which is caused by the inherent feature of SNC. For SNC analytical modelling, the service process should be concise and provide reasonable tight bound in Eq. 3. However, it is difficult to construct optimal service process, and approximation operation is hardly avoided in the performance derivation.

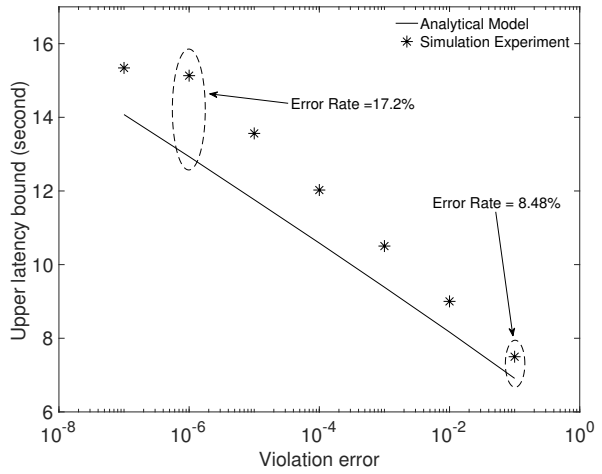

Fig. 5. Comparison of the Simulation Experiment Results with These Obtained from the Analytical Model by Varying the Violation Error: φ_1=0.8, φ_2=0.2, λ_1 = 0.075M packets/second, λ_2=0, n=4, m_i=3, μ_s=0.095M packets/second, λ_{cr_j}=0.001M packets/second, and ϵ=10^{-1}-10^{-7}.

For the multiple nodes case, the queuing theory method could provide the accurate performance bound when the arriving traffic follows the non-bursty Poisson process. However, with the bursty arriving traffic, *e.g.* following MMPP process, it is very complex and difficult for queueing theory to achieve a closed-form performance bound. In order to comprehensively evaluate the accuracy of the developed model, we conducted simulation experiments in the Objective Modular Network Testbed in C++ (OMNeT++) simulation environment [48]. The 95% confidence is used in the simulator to collect the results when the simulation experiments reach the steady state. The NFV network in the simulator consists of four kinds of network components, such as the through traffic generator, the cross traffic generator, VNF nodes, and the destination node for data collection and analysis. VNF nodes are connected in the form of tandem networks. The first node in the network is connected to the through traffic generator, and the last is connected to the data collection node. Multiple cross traffic generators are linked to each of the VNF nodes to form the resource competition for the through traffic. For the sake of a specific illustration, the network configuration is set as follows: the arriving traffic follows the bursty MMPP process, which has two states, "1" and "2" respectively. The traffic rates and service rates are set based on the work reported in [32]. The transmission rates in MMPP process are set based on the work in [49]. The arrival rate of state 1 is 0.075M packets/second; the arrival rate of state 2 is 0. The transition rate from state 1 to state 2 is set to be 0.8 and from state 2 to state 1 is 0.2. There are four VNF nodes, each of which serves one through traffic flow and three cross traffic flows. The arrival rate for the cross traffic is 0.001M packets/second and the service rate is 0.095 packets/s. The simulator was developed based on event-driven method, and the results are collected by averaging 50 simulation runs, each of which lasts for 60 seconds.

Fig. 5 presents the upper latency bound predicted by the analytical model and those generated by simulation experiment. It can be seen that the developed model provides a good degree of matching with the simulation experiment results. The relative error of the developed model against the simulation results is in the range of 8%-17%. The explanation of the relative error is as follows: to simplify the derivation process, the approach in Eq. (33) adopts approximation operation to derive the upper latency bound; otherwise it is difficult to achieve a conservative and closed-form upper latency bound. Compared with the results reported in the existing literatures [42]. 15%-20% is acceptable for the least upper latency bound in SNC.

B. Performance Analysis

In this subsection, the analytical model developed in Section IV will be leveraged to investigate the performance of NFV service chain in different service configurations, by varying the number of VNF nodes, the number and arrival rates of cross traffic and the violation error probability.

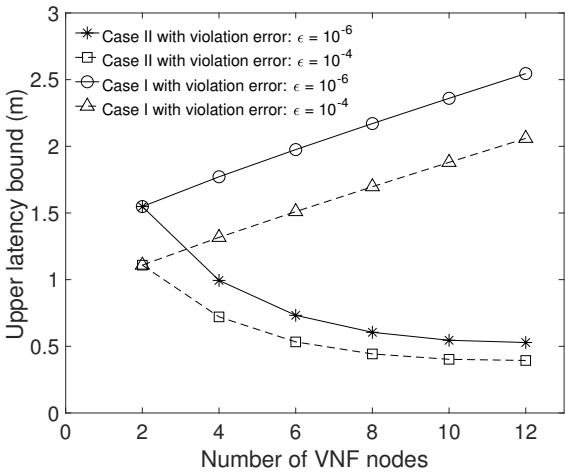

Fig. 6. Impacts of the Number of the VNF nodes on the Upper Latency Bound with Case I and II. Case I. μ_s=0.095M packets/second. Case II: μ_s = [0.095, 0.1045, 0.1140, 0.1235, 0.1335, 0.1425] packets/second. n = [2, 4, 6, 8, 10, 12], ϵ = [10^{-6}, 10^{-4}] , φ_1=0.8, φ_2=0.2, λ_1 = 0.075M packets/second, m_i = 3, and λ_{cr_j}=0.001M packets/second.

- Impacts of the number of VNF nodes on the upper latency bound

The number of VNF nodes is an important aspect for the deployment of NFV chains; adding a VNF node to an existing NFV chain can introduce another network function for the end user. However, it could result in additional processing latency and performance degradation. This subsection aims to study the impact of the number of VNF nodes on the upper latency bound with different violation error requirements. The network setting is described as follows: the arrival

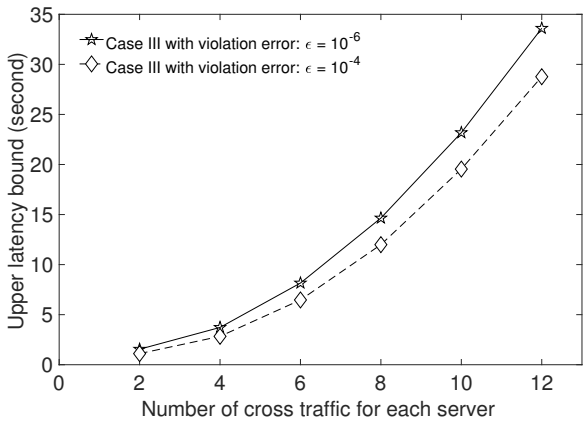

Fig. 7. Impacts of the Number of the VNF nodes on the Upper Latency Bound with Case III. μ_s = [0.095, 0.0855, 0.0760, 0.0665, 0.0570, 0.0475] packets/second. n = [2, 4, 6, 8, 10, 12], ϵ = [10^{-6}, 10^{-4}] , φ_1=0.8, φ_2=0.2, λ_1 = 0.075M packets/second, m_i = 3, and λ_{cr_j}=0.001M packets/second.

traffic is modelled as an on-off Poisson process, which has two states (0 and 1); the transmission rate from state 0 to state 1 is set to be 0.8; from state 1 to state 0 is 0.2. When the Markov chain is in state 0, there is no packet for through traffic. When Markov chain is in state 1, the arrival rate is set to be 0.075M packets/second. Both the homogeneous and non-homogeneous service processes are studied in this subsection. For homogeneous scenario (Case I), the service process follows the exponential distribution with the fixed service rate (0.095M packets/second). For the non-homogeneous scenario, the service rates change as the number of VNF nodes increases. Furthermore, we consider two scenarios in the non-homogeneous scenario, where the service rates are increasing and decreasing with the growth of VNF nodes. In the increasing scenario (Case II), the service rates are set to be [0.095, 0.1045, 0.1140, 0.1235, 0.1335, 0.1425] packets/second. In the decreasing scenario (Case III), the service rates are set to be [0.095, 0.0855, 0.0760, 0.0665, 0.0570, 0.0475] packets/second. The number of cross traffic flows is set to be zero, avoiding the impact of cross traffic on the investigation of VNF nodes and the upper latency bound. Fig. 6 shows the relation between the number of the VNF nodes, n, and the least upper latency bound, w for Case I and Case II. Fig. 7 demonstrates the results for Case III.

From Figs. 7 and 8, it can be seen that the upper latency bound has a linear relationship with the number of the VNF nodes for the homogeneous scenario in Case I. However, this linear relationship does not hold for Case II and Case III in the non-homogeneous scenario. Because, after transferring multiple VNF nodes into a single system through Min-plus convolution, the end-to-end latency bound is given by Eq. (33), which is a function of n, μ_{s_i}, λ^i_{th}, $\lambda^i_{cr_j}$, ϵ and θ. Under the non-homogenous scenario, the latency bound is not only determined by the number of VNF nodes, but also determined by the service process. Therefore, when n and μ_{s_i} change simultaneously in Case II and Case III, the linear relationship does not hold between the upper latency bound and the number of VNF nodes. In addition, through the comparison of the solid line and dotted line, it can be seen that the increase of the violation error rate brings the drop of the upper latency bound for all three cases. For instance, let us fix n as 6, when $\epsilon = 10^{-6}$, the latency bound is equal to 1.98 seconds in Fig. 6 for Case I; when ϵ increases from 10^{-6} to 10^{-4}, the upper latency bound drops from 1.98 seconds to 1.4 seconds. The reason is that the upper latency bound is an increasing function of the violation error probability in Eq. (33).

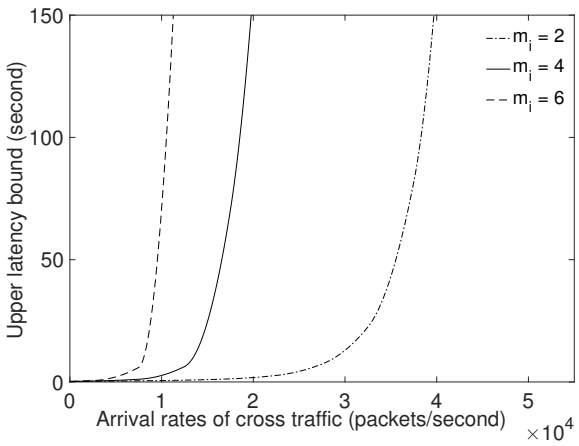

Fig. 8. Impacts of the Arrival Rates of the Cross Traffic on the Upper Latency Bound: μ_s = 0.095M packets/second. n = 2, m_i = [2, 4, 6], $\epsilon = 10^{-6}$, φ_1=0.8, φ_2=0.1, and λ_{cr_j}=0.001M packets/second.

- Impacts of cross traffic on the latency performance of through traffic

NFV brings network operators the benefits of flexible service deployment, scalable network architecture, and lower OPEX and CAPEX, while at the same time it struggles to provide end-to-end SLA-guaranteed services due to the co-existence NFV chain deployment. The developed model in this work provides a cost-efficient approach to investigate the effects of the cross traffic on the performance of the through traffic. Network configuration is described as follows: the through traffic and server are set the same as those in the above subsection. The number of NFV chains is set to be 2; and the violation error rate is fixed as 10^{-6}. The number of cross traffic flow on the ith server, m_i is set to be 2, 4, and 6. Fig. 8 shows the relationship between the arrival rates of the cross traffic and the upper latency bound when m_i is equal to 2, 4, and 6. Fig. 8 reveals that the cross traffic significantly affects the performance of the NFV network. A large number or higher volume of cross traffic brings higher upper latency bound for the through traffic. Because under the FIFO scheduling algorithm, the large number of cross traffic or high volume of cross traffic consumes a significant amount of server resources, leading to less resource available for the through traffic and negative impacts on the service provisioning for the through traffic.

- Impacts of violation error probability on the latency performance of through traffic

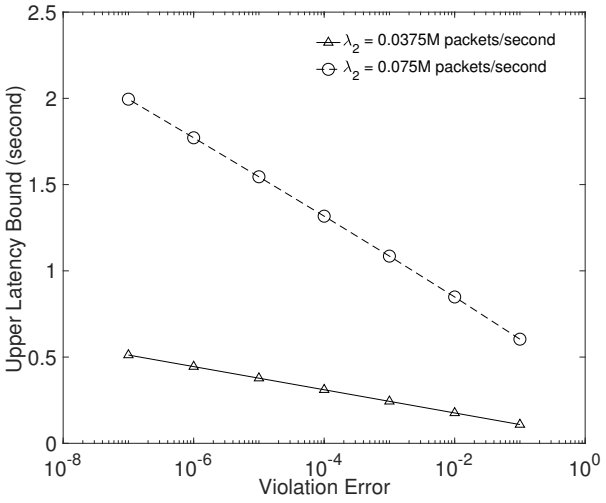

Fig. 9. Impacts of the Violation Error on the Latency Performance of the Through Traffic (μ_s = 0.095M packets/second. n = 2, m_i = 2, ϵ = 10^{-1}-10^{-6} , φ_1=0.8, φ_2=0.1, λ= [0.075, 0.0375] packets/second, and λ_{cr_j}=0.001M packets/second.

In NFV networks, network service is offered based on the SLA agreements between the network operator and the service provider. The analytical model developed in this study is SLA-aware and is capable of providing the relationship between end-to-end latency bound and violation error probability. In this subsection, the service rate is set to be 0.095M packets/second; the number of the VNF nodes is set to be 2; the number and arrival rate for cross traffic are 2 and 0.001M packets/second; the arrival rate of the through traffic is set to be 0.0375M and 0.075M packets/second. The transmission rates are set the same as those in previous sections. Fig. 9 gives the least upper latency bound by varying the violation error rates, ϵ, from 10^{-7} to 10^{-1}. It can be seen that the smaller value of ϵ significantly pushes up the upper latency bound; this means that with the strict violation *e.g.* 10^{-7}, the higher latency bound is needed to ensure that $1 - 10^{-7}$ of the packets are delivered within the defined latency bound. Furthermore, heavy volume of through traffic would also result in the higher latency bound. Because with the given server resource, more arrivals would increase the server burden and result in the longer waiting time and the higher upper latency bound. Therefore, in order to reduce the upper latency bound, additional computing, storage and network resources should be deployed once the upper latency bound cannot meet the requirements of transmission latency.

As shown in the above analysis, the developed analytical model could be used as a practical tool in the NFV service deployment and management. Especially in the phase of NFV service chain design, the developed analytical model could assist service providers to improve the performance of NFV service chaining and placement. For instance, most of the NFV chaining and placement strategies are designed with the aim of maximising the resource utilisation and pay little attention to SLA guarantee. In this area, the proposed analytical model provides an efficient method for service providers to determine which service chaining and placement method is more beneficial to meet the SLA requirements in terms of violation error and latency performance. During model design, we use a parameterisation method to develop the analytical model. Therefore, the service providers and network operators could customise the values of parameters according to their usage scenarios and conduct performance evaluation.

VI. Conclusion

NFV is regarded as a disruptive technology for telecommunication service providers to reduce the CAPEX and OPEX through decoupling individual network services from the underlying hardware devices. In this work, a novel analytical model has been developed based on SNC to investigate the upper latency bound of the NFV service chain. Instead of giving the average performance metrics, the developed model has derived the worst-case of system performance with the aim of quantitatively achieving the network performance in terms of SLA guarantee. During derivation of the upper latency bound, MGF and EBB have been used to achieve the SACs and SSCs for both non-bursty and bursty traffic. In order to consider the cross traffic in the analytical model, leftover service technology has been exploited to calculate the available service for through traffic. The end-to-end upper latency bound has been calculated based on the achieved SACs, SSCs, violation error requirements and network topology settings. Both the exact theory and simulation results have been used to validate the accuracy of the analytical models.

The proposed analytical model could be a useful tool for network service providers to optimise their service operation and management. For instance, instead of passively testing the performance when the NFV service chain is deployed and used in the practical environment, the proposed model enables service providers to actively predict the performance for NFV service chain deployment, determine whether a certain NFV deployment can meet the SLA requirement, and more importantly analyse the impact of the new service chain on performance of the existing services that have been deployed, which is a very important issue in the multiple-tenant environment. For the future research work, we plan to utilise the SNC to analyse the performance of NFV service chain with more complex deployments, where multiple Priority Queues (PQ) and complex schedulers, *e.g.* PQ-based scheduling or Early Deadline First (EDF) may be used to meet the QoS bounds.

References

[1] R. Gouareb, V. Friderikos, and A. H. Aghvami. Virtual network functions routing and placement for edge cloud latency minimization. *IEEE Journal on Selected Areas in Communications*, pages 1–1, 2018.

[2] M. T. Raza, S. Lu, M. Gerla, and X. Li. Refactoring network functions modules to reduce latencies and improve fault tolerance in nfv. *IEEE Journal on Selected Areas in Communications*, pages 1–1, 2018.

[3] F. Z. Yousaf and T. Taleb. Fine-grained resource-aware virtual network function management for 5g carrier cloud. *IEEE Network*, 30(2):110–115, March 2016.

[4] F. Yousaf, M. Bredel, S. Schaller, and F. Schneider. Nfv and sdn key technology enablers for 5g networks. *IEEE Journal on Selected Areas in Communications*, 35(11):2468–2478, Nov 2017.

[5] X. Ge, Y. Liu, D. Du, L. Zhang, H. Guan, J. Chen, Y. Zhao, and X. Hu. Openanfv: Accelerating network function virtualisation with a consolidated framework in openstack. *ACM SIGCOMM Computer Communication Review*, 44(4):353–354, 2015.

[6] A. Aissioui, A. Ksentini, A. M. Gueroui, and T. Taleb. Toward elastic distributed sdn/nfv controller for 5g mobile cloud management systems. *IEEE Access*, 3:2055–2064, 2015.

[7] Intel. Data plane development kit. *URL http://dpdk. org*, 2014.

[8] K. Yamazaki, T. Osaka, S. Yasuda, S. Ohteru, and A. Miyazaki. Accelerating sdn/nfv with transparent offloading architecture. In *Presented as part of the Open Networking Summit 2014 (ONS 2014)*, 2014.

[9] M. Fidler. Survey of deterministic and stochastic service curve models in the network calculus. *IEEE Communications Surveys & Tutorials*, 12(1):59–86, 2010.

[10] M. Fidler and A. Rizk. A guide to the stochastic network calculus. *IEEE Communications Surveys & Tutorials*, 17(1):92–105, 2015.

[11] Y. Jiang. A note on applying stochastic network calculus. In *Proceedings of SIGCOMM*, volume 10, pages 16–20. Citeseer, 2010.

[12] R. Lübben, M. Fidler, and J. Liebeherr. Stochastic bandwidth estimation in networks with random service. *IEEE/ACM Transactions on Networking*, 22(2):484–497, 2014.

[13] Q. Duan. Modeling and performance analysis for service function chaining in the sdn/nfv architecture. In *2018 4th IEEE Conference on Network Softwarization and Workshops (NetSoft)*, pages 476–481, June 2018.

[14] I. Benkacem, T. Taleb, M. Bagaa, and H. Flinck. Performance benchmark of transcoding as a virtual network function in cdn as a service slicing. In *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6, April 2018.

[15] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck. Network slicing & softwarization: A survey on principles, enabling technologies & solutions. *IEEE Communications Surveys & Tutorials*, 2018.

[16] D. Carlos, C. Amanda, and M. Germn. Container-based virtual elastic clusters. *Journal of Systems and Software*, 127:1 – 11, 2017.

[17] Alexej Grigorjew, Stanislav Lange, Thomas Zinner, and Phuoc Tran-Gia. *Performance Benchmarking of Network Function Chain Placement Algorithms*, pages 83–98. 01 2018.

[18] X. Cheng, Y. Wu, G. Min, and A. Y. Zomaya. Network function virtualization in dynamic networks: A stochastic perspective. *IEEE Journal on Selected Areas in Communications*, 36(10):2218–2232, Oct 2018.

[19] G. Wang, Y. Zhao, J. Huang, and Y. Wu. An effective approach to controller placement in software defined wide area networks. *IEEE Transactions on Network and Service Management*, 15(1):344–355, March 2018.

[20] M. Xia, M. Shirazipour, Y. Zhang, H. Green, and A. Takacs. Network function placement for nfv chaining in packet/optical datacenters. *Journal of Lightwave Technology*, 33(8):1565–1570, 2015.

[21] H. Moens and F. Turck. Vnf-p: A model for efficient placement of virtualised network functions. In *10th International Conference on Network and Service Management (CNSM) and Workshop*, pages 418–423. IEEE, 2014.

[22] M. Bagaa D.L.C. Dutra R. A. Addad, T. Taleb and H. Flinck. Towards modeling cross-domain network slices for 5g. In *GLOBECOM 2018 - 2018 IEEE Global Communications Conference*, Dec 2018.

[23] A. Laghrissi, T. Taleb, M. Bagaa, and H. Flinck. Towards edge slicing: Vnf placement algorithms for a dynamic amp;amp; realistic edge cloud environment. In *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, pages 1–6, Dec 2017.

[24] M. Bagaa, T. Taleb, and A. Ksentini. Service-aware network function placement for efficient traffic handling in carrier cloud. In *2014 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 2402–2407, April 2014.

[25] I. Chih-Lin, J. Huang, R. Duan, C. Cui, and L. Li. Recent progress on c-ran centralisation and cloudification. *IEEE Access*, 2:1030–1039, 2014.

[26] D. L. C. Dutra, M. Bagaa, T. Taleb, and K. Samdanis. Ensuring end-to-end qos based on multi-paths routing using sdn technology. In *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, pages 1–6, Dec 2017.

[27] C. Sun, J. Bi, Z. Zheng, and H. Hu. Hyper: A hybrid high-performance framework for network function virtualisation. *IEEE Journal on Selected Areas in Communications*, 35(11):2490–2500, Nov 2017.

[28] D. Gross. *Fundamentals of Queueing Theory*. John Wiley & Sons, 2008.

[29] G. Min and M. Ould-Khaoua. A performance model for wormhole-switched interconnection networks under self-similar traffic. *IEEE Transactions on Computers*, 53(5):601–613, May 2004.

[30] G. Min, J. Hu, and M. E. Woodward. Performance modelling and analysis of the txop scheme in wireless multimedia networks with heterogeneous stations. *IEEE Transactions on Wireless Communications*, 10(12):4130–4139, December 2011.

[31] X. Jin and G. Min. Modelling and analysis of priority queueing systems with multi-class self-similar network traffic: A novel and efficient queue-decomposition approach. *IEEE Transactions on Communications*, 57(5):1444–1452, May 2009.

[32] A. Fahmin, Y. Lai, M. S. Hossain, Y. Lin, and D. Saha. Performance modeling of sdn with nfv under or aside the controller. In *2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, pages 211–216, Aug 2017.

[33] Wang Miao, Geyong Min, Yulei Wu, Haozhe Wang, and Jia Hu. Performance modelling and analysis of software defined networking under bursty multimedia traffic. *ACM Transaction on Multimedia Computing, Communications, and Applications*, 12(5s):77:1–77:19, September 2016.

[34] A. Lombardo, A. Manzalini, V. Riccobene, and G. Schembra. An analytical tool for performance evaluation of software defined networking services. In *2014 IEEE Network Operations and Management Symposium (NOMS)*, pages 1–7, May 2014.

[35] Jun H, X. Liqian, D. Qiang, X. Cong-cong, L. Jiangtao, and Y. Shui. Modelling and performance analysis for multimedia data flows scheduling in software defined networks. *Journal of Network and Computer Applications*, 83:89 – 100, 2017.

[36] European Telecommunication Standards Institute Foundation. Network functions virtualisation (nfv): Management and orchestration. *ETSI NFV Group Report*, 2017.

[37] R. Eletreby and O. Yagan. Connectivity of wireless sensor networks secured by heterogeneous key predistribution under an on/off channel model. *IEEE Transactions on Control of Network Systems*, 2018.

[38] Michael George Bulmer. *Principles of statistics*. Courier Corporation, 2012.

[39] G. Bolch, S. Greiner, H. Meer, and K. Trivedi. *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*. John Wiley & Sons, 2006.

[40] Y. Xu, E. Altman, R. El-Azouzi, M. Haddad, S. Elayoubi, and T. Jimenez. Analysis of buffer starvation with application to objective qoe optimisation of streaming services. *IEEE Transactions on Multimedia*, 16(3):813–827, April 2014.

[41] B. Leng, X. Guo, X. Zheng, B. Krishnamachari, and Z. Niu. A wait-and-see two-threshold optimal sleeping policy for a single server with bursty traffic. *IEEE Transactions on Green Communications and Networking*, 1(4):528–540, Dec 2017.

[42] M. Fidler. A network calculus approach to probabilistic quality of service analysis of fading channels. In *IEEE Globecom 2006*, pages 1–6. IEEE, 2006.

[43] G. Faraci, A. Lombardo, and G. Schembra. A processor-sharing scheduling strategy for nfv nodes. *Journal of Electrical and Computer Engineering*, 2016:1, 2016.

[44] C. Chang. *Performance Guarantees in Communication Metworks*. Springer Science & Business Media, 2012.

[45] M. Fidler. An end-to-end probabilistic network calculus with moment generating functions. In *200614th IEEE International Workshop on Quality of Service*, pages 261–270. IEEE, 2006.

[46] J. Liebeherr, M. Fidler, and S. Valaee. A system theoretic approach to bandwidth estimation. *IEEE/ACM Transactions on Networking*, 18(4):1040–1053, 2010.

[47] Florin Ciucu. Network calculus delay bounds in queueing networks with exact solutions. In Lorne Mason, Tadeusz Drwiega, and James Yan, editors, *Managing Traffic Performance in Converged Networks*, pages 495–506, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[48] Andrs Varga and OpenSim Ltd. Omnet++ user guide. In *Version 5.4.1*, 2016.

[49] K. Wang, M. Tao, W. Chen, and Q. Guan. Delay-aware energy-efficient communications over nakagami-$m$fading channel with mmpp traffic. *IEEE Transactions on Communications*, 63(8):3008–3020, Aug 2015.