A Novel Approach for Service Function Chain (SFC) Mapping with Multiple SFC instances in a Fog-To-Cloud Computing System

Ali Zamani
Electrical Engineering department
Amirkabir university of Technology
Tehran, Iran
az_1995@aut.ac.ir

Saeed Sharifian
Electrical Engineering department
Amirkabir university of Technology
Tehran, Iran
Sharifian_S@aut.ac.ir

Abstract— Internet of Things (IoT) has been ever-growing over the last few years. The IoT devices generate a massive amount of data that should be transmitted to the cloud for computing. Cloud consolidation and centralization lead to many network hops between the IoT devices and its associated cloud which makes two critical problems: (i) high latencies (ii) high bandwidth consumption in the IoT domain. Network Function Virtualization (NFV), Software Defined Network (SDN) and fog computing have been emerged to address these problems. In the Fog-to-Cloud (F2C) architecture, Fog and cloud work together to provide computing, storage, and application services in the IoT domain. To build complex services a specific set of virtual network functions can be chained together in a specific order which is known as Service Function Chaining (SFC). The joint VNF placement and traffic routing are called SFC mapping. In this paper, we propose an Integer Linear Program (ILP) model to solve SFC mapping in the fog-to-cloud Computing System in order to minimize the overall end-to-end (e2e) latency of IoT devices. We observe that our approach reduces the overall e2e latency of IoT devices significantly. Moreover, our approach helps us to analyze the effect of a number of instances in the end-to-end latency of IoT devices.

***Keywords*— Internet of Things (IoT); Network Function Virtualization (NFV); Software Defined Networking (SDN); fog computing; Service Function Chain (SFC)**

I. INTRODUCTION

Internet of Things (IoT), which interconnects billions or even trillions of diverse devices such as sensors, vehicles, and smartphones, has been ever-growing over the last few years. The IoT devices, which are named "Things", generate a massive amount of data that should be transmitted to the cloud for computing. Although the cloud offers various benefits such as scalability and elasticity, its consolidation and centralization lead to many network hops between the Things and its associated cloud and results in high latencies and high bandwidth consumption in the IoT domain [1].

To address these problems, many technologies related to the expansion of the IoT have emerged, including network function virtualization (NFV) [2], Software Defined Network (SDN)[3] and fog computing [1]. The fog computing offers distributed edge cloud close to the Things, therefore, fog and cloud work together to provide computing, storage, and application services in the IoT domain. This architecture (Fig. 1) is named "Fog-to-Cloud (F2C)"[4]. Due to the complex management of such a network of distributed fogs and providing services mainly in the IoT domain, SDN and NFV have been proposed [1]. On the one hand, the SDN separates the control and data planes [3]. on the other hand, the NFV reshapes dedicated hardware functionality as software modules named virtual network functions (VNFs) for an agile and scalable service placement and reducing Capital Expenditure (CAPEX) and Operation Expense (OPEX). To build complex services such as e-healthcare, face recognition and augmented reality, a specific set of VNFs can be chained together in a specific order which is known as Service Function Chaining (SFC) [2]. The joint VNF placement and traffic routing are called SFC mapping. The NFV and SDN convergence in the edge cloud leads to fast new services and application delivery and deployment [1].

There are several surveys [2],[5],[6] on NFV that the various NFV challenges are explained. Most of the works in the literature report exact and heuristic mathematical formulations for the SFC mapping. Draxler et al. [7] solve VNFs placement and routing with heuristic and exact solutions. The authors consider the total data rate and total latency as their objective function. Huin et al. [8] propose an exact mathematical model for choosing the number and the location of the VNFs. Masri et al.[9] proposed an algorithm in order to select optimal fog or cloud for doing services. In the reference [10], the authors consider task scheduling in a fog-to-cloud computing system, and propose a heuristic-based algorithm, whose major objective is maximizing the profits of fog service provider while meeting the tasks' deadline constraint. A scalable approach base on Integer Linear Program (ILP) column-generation-based model, which provides quasi-optimal solutions, is proposed by Gupta et al. [11]. On account of scalability limit of ILP column-generation-based model to the quadratic constraints, a two-phase column-generation-based model is proposed by Gupta et al. [12]. [6], [7], [8], [9] do not consider SFC mapping in the F2C architecture which is critical in the IoT domain. [9] and [10] do not consider the concept of NFV in their system model. Our assumptions and terminology are based on[12], but there are serious differences that our method concentrates on minimizing overall end-to-end (e2e)

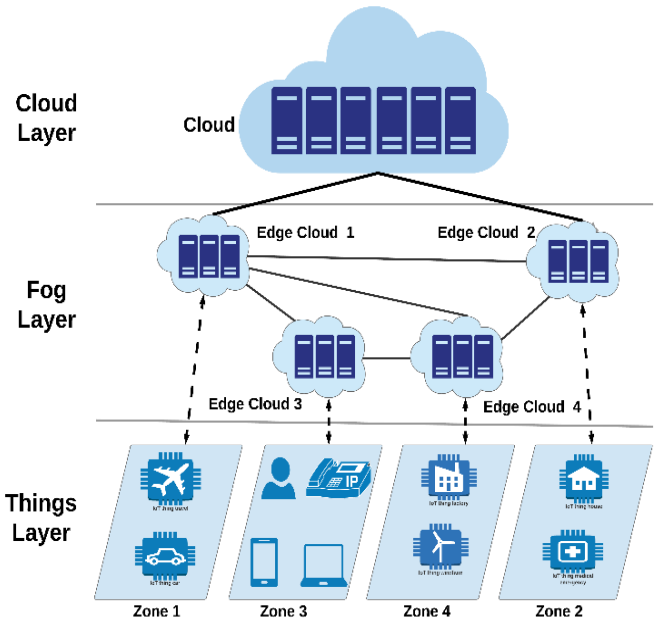

Fig. 1. Fog-to-cloud computing system.

latency of IoT devices, while [9] considers minimizing bandwidth consumption as an objective function, we consider F2C architecture in system model, while [12] did not consider. moreover, [12] do not consider the Service Level Agreement (SLA) as a constraint, while in our formulation it is considered.

In this paper, we propose an ILP solution that provides the exact solution for SFC mapping in the F2C architecture, moreover, in order to minimize total latency of network, the number of instances for each SFC and replicas for each VNF are considered. Numerical results show that our model significantly minimizes the total latency of network in comparison with cloud computing.

The composition of this paper is as follows: Section II introduces optimization model and the input parameter. Numerical results are presented in section III, eventually, section IV draws the conclusions.

II. System Model

IoT devices request multiple services and each service is realized by traversing an SFC. We consider F2C architecture (Fig. 1) and multiple SFCs which should map to the F2C architecture. Fig. 1 shows 4 zones in the Things layer. Each zone is connected to the one fog and finally two fogs are connected to the cloud layer. We develop ILP solution for SFC mapping.

A. Problem statement

According to a network topology, capacity and latency of links, computing resources at the fogs and cloud nodes, traffic flows between two pairs of fog or fogs and cloud requiring a specific SFC, users' SLA, instances' number, we determine the placement of VNFs, corresponding traffic routing and users' assignment to the SFC instances to minimize overall latency of network.

TABLE I. INPUT PARAMETERS

G	Physical topology of fog-to-cloud architecture $G=(V,L)$ with: V set of cloud (V_{CLOUD}) and fog (V_{FOG}) nodes and L : set of links
I_c	Maximum number of instances for SFC c
F	Set of VNFs, indexed by f
R_f	Maximum number of replicas of VNF
$n_v^{FOGCORE}$	Number of CPU cores present in a fog node $v \in V_{FOG}$
$n_v^{CLOUDCORE}$	Number of CPU cores present in a cloud node $v \in V_{CLOUD}$
n_f^{CORE}	Number of CPU cores per Gbps for function f
C	Set of chains, indexed by c
n_c	Number of VNFs in SFC c
SD	Set of source-destination (V_s,V_d) pairs
SD_c	Source-destination (V_s,V_d) pairs for SFC c
D_{sd}^c	Traffic demand between V_s , V_d and for SFC c
$\sigma_i(c)$	ID of i th VNF in SC c , where $f_{\sigma_i(c)} \in F$
CAP_ℓ	Capacity of link ℓ ,where $\ell \in L$
SLA_{sd}^c	SLA of user $sd \in SD_c$
$delay_\ell$	Latency of link $\ell \in L$

B. Modeling

Table I illustrates the input parameters used. Each SFC indexed by c, is defined by an ordered set of n_c functions. The following notation represents SFC c :

$$[SFC\ c] \qquad f_{\sigma_1(c)} \rightarrow f_{\sigma_2(c)} \rightarrow \cdots \rightarrow f_{\sigma_{n_c}(c)} \qquad (1)$$

we generate all configurations of each SFC c $(\hat{\Gamma}_c)$ and select I_c number of them that result in minimizing overall latency. Let us consider $\hat{\Gamma}$ the aggregation of all configuration of SFCs. Each configuration $(\hat{\gamma})$ of SFC c is characterized by the following parameters:

- Location of VNFs: $a_{vi}^{\hat{\gamma}}=1$ if i th function $f_i \in c$ is located in node v in configuration $\hat{\gamma}$; 0 otherwise.
- Connectivity of located VNFs: path from location of current VNF to next VNF in SFC c . If link ℓ is used in the path from location of $f_{\sigma_i}(c)$ to the location of $f_{\sigma_i+1}(c)$, then $b_{i\ell}^{\hat{\gamma}}=1$; 0 otherwise.

- User assignment: $\delta_{sd}^{\hat{\gamma}}$ =1 if (v_s, v_d) uses configuration $\hat{\gamma}$; 0 otherwise.

To clarify the concept of configuration, we consider SFC c_1 as follows:

$$f_{\sigma_1(c_1)} \rightarrow f_{\sigma_2(c_1)} \rightarrow f_{\sigma_{n_c}(c_1)} \quad \left(f_{\sigma_1(c_1)} = f_1 \,;\, f_{\sigma_2(c_1)} = f_5 \,;\, f_{\sigma_3(c_1)} = f_8\right)$$

Three configurations of SFC c_1 are displayed in the network topology of Fig. 2. In the first topology, f_1 is located in the fog node 3, f_5 , f_8 are located in the fog node 1, blue dash line is used for traversing data and user 1, 2 are using this topology. The Second topology is similar to the first topology but green dash line is used for traversing data. In the third topology, f_1 , f_5 , f_8 are located in the fog node 4 and 2 and cloud node 5, respectively. Purple dash line is used for traversing data and user 2 is using this topology. Potential set of configurations for an SFC *c* can be computed by following equation:

$$\hat{\Gamma}_c = \underbrace{\{N_V\}^{n_c}}_{\text{Location of VNFs}} \times \underbrace{P_{paths}{}^{n_c-1}}_{\text{Connectivity of located VNFs}} \times \underbrace{\sum_{sd=1}^{N_{SD_c}} \binom{N_{SD_c}}{sd}}_{\text{User assignment}} \tag{2}$$

Where N_V gives the number of nodes(cloud and fogs), P_{paths} gives the number of paths from the location of $f_{\sigma_i}(c)$ to the location of $f_{\sigma_i+1}(c)$, sd is the number of users (source-destination (V_s, V_d) pairs) using a configuration, N_{SD_c} is the number of users that request SFC *c.*

C. Objective function

The SFC mapping can be divided into two steps. First, I_c number of configurations is selected while satisfying various constraints on nodes and links of networks. Second, the route from the source (v_s) to the first VNF and from the last VNF to the destination (v_d) are selected while satisfying various constraints on links of networks. For the first step, we consider following variables:

- $z_{\hat{\gamma}}$: 1 if configuration $\hat{\gamma}$ is selected; 0 otherwise
- x_{vf} : 1 if function f is located in v ; 0 otherwise.

And for the second step, we consider following variables:

- $y_\ell^{f_1(c),sd}$: 1 if ℓ is on path from v_s to location of first VNF in c ; 0 otherwise.
- $y_\ell^{f_{n_c}(c),sd}$: 1 if ℓ is on path from location for last VNF in c to v_d ; 0 otherwise.

We model our problem as ILP which selects the configuration(first term) and also selects links from v_s to location of first VNF in c and from last VNF in c to v_d (second term) that result in minimizing e2e latency. Equation (3) shows our objective function. It is obvious that objective and constraint functions are linear for all $z_{\hat{\gamma}}$, x_{vf} , $y_\ell^{f_1(c),sd}$, $y_\ell^{f_{n_c}(c),sd}$,hence, our problem can be solved by ILP solvers.

$$\min. \underbrace{\sum_{c\in C}\sum_{\hat{\gamma}\in\hat{\Gamma}_c}\left(\sum_{sd\in SD}\sum_{\ell\in L}\sum_{i=1}^{n_c-1} b_{i\ell}^{\hat{\gamma}}\,\delta_{sd}^{\hat{\gamma}}\,delay_\ell\right) z_{\hat{\gamma}}}_{First\ Term} + \underbrace{\sum_{c\in C}\sum_{\ell\in L}\sum_{sd\in SD} delay_\ell \left(y_\ell^{f_1(c),sd} + y_\ell^{f_{n_c}(c),sd}\right)}_{Second\ Term} \tag{3}$$

D. Constrants

$$\sum_{\hat{\gamma}\in\hat{\Gamma}_c} z_{\hat{\gamma}} \le I_c \qquad c \in C \tag{4}$$

$$\sum_{c\in C}\sum_{\hat{\gamma}\in\hat{\Gamma}_c}\sum_{i=1}^{n_c} T_{fi}^c\, a_{vi}^{\hat{\gamma}}\, z_{\hat{\gamma}} \le M\, x_{vf} \qquad f \in F, v \in V \tag{5}$$

$$\sum_{c\in C}\sum_{\hat{\gamma}\in\hat{\Gamma}_c}\sum_{i=1}^{n_c} T_{fi}^c\, a_{vi}^{\hat{\gamma}}\, z_{\hat{\gamma}} \ge x_{vf} \qquad f \in F, v \in V \tag{6}$$

$$\sum_{v\in V} x_{vf} \le R_f \qquad f \in F \tag{7}$$

$$\sum_{c\in C}\sum_{\hat{\gamma}\in\hat{\Gamma}_c}\sum_{sd\in SD} D_{sd}^c\,\delta_{sd}^{\hat{\gamma}} * \left(\sum_{f\in F}\sum_{i=1}^{n_c} T_{fi}^c\, n_f^{CORE}\, a_{vi}^{\hat{\gamma}}\right) z_{\hat{\gamma}} \le n_v^{FOGCORE} \qquad v \in V_{FOG} \tag{8}$$

$$\sum_{c\in C}\sum_{\hat{\gamma}\in\hat{\Gamma}_c}\sum_{sd\in SD} D_{sd}^c\,\delta_{sd}^{\hat{\gamma}} * \left(\sum_{f\in F}\sum_{i=1}^{n_c} T_{fi}^c\, n_f^{CORE}\, a_{vi}^{\hat{\gamma}}\right) z_{\hat{\gamma}} \le n_v^{CLOUDCORE} \qquad v \in V_{CLOUD} \tag{9}$$

$$\sum_{c\in C}\sum_{sd\in SD} D_{sd}^c * \left(y_\ell^{f_1(c),sd} + y_\ell^{f_{n_c}(c),sd} + \sum_{\hat{\gamma}\in\hat{\Gamma}_c}\delta_{sd}^{\hat{\gamma}}\, z_{\hat{\gamma}} \sum_{i=1}^{n_c-1} b_{i\ell}^{\hat{\gamma}}\right) \le CAP_\ell \quad \ell \in L \tag{10}$$

$$\sum_{\hat{\gamma}\in\hat{\Gamma}_c}\delta_{sd}^{\hat{\gamma}} z_{\hat{\gamma}} = 1 \qquad c \in C,\ sd \in SD: D_{sd}^c > 0 \tag{11}$$

$$\sum_{\hat{\gamma}\in\hat{\Gamma}_c}\sum_{\ell\in L} delay_\ell * (y_\ell^{f_1(c),sd} + y_\ell^{f_{n_c}(c),sd} + \sum_{\hat{\gamma}\in\hat{\Gamma}_c}\delta_{sd}^{\hat{\gamma}}\, z_{\hat{\gamma}} \sum_{i=1}^{n_c-1} b_{i\ell}^{\hat{\gamma}}) \le SLA_{sd}^c \qquad c \in C, sd \in SD_c \tag{12}$$

$$\sum_{\hat{\gamma}\in\hat{\Gamma}_c} \delta_{sd}^{\hat{\gamma}}\, a_{v_s 1}^{\hat{\gamma}}\, z_{\hat{\gamma}} + \sum_{\ell\in\omega^+(v_s)} y_{\ell}^{f_1(c),sd} = 1 \tag{13}$$

$$c\in C, sd\in SD: D_{sd}^{c} > 0$$

$$\sum_{\hat{\gamma}\in\hat{\Gamma}_c} \delta_{sd}^{\hat{\gamma}}\, a_{v1}^{\hat{\gamma}}\, z_{\hat{\gamma}} - \sum_{\ell\in\omega^-(v)} y_{\ell}^{f_1(c),sd} \le 0 \tag{14}$$

$$c\in C, sd\in SD: D_{sd}^{c} > 0, v\in V\setminus\{v_s\}$$

$$\sum_{\hat{\gamma}\in\hat{\Gamma}_c} \delta_{sd}^{\hat{\gamma}}\, a_{v1}^{\hat{\gamma}}\, z_{\hat{\gamma}} + \sum_{\ell\in\omega^+(v)} y_{\ell}^{f_1(c),sd} - \sum_{\ell\in\omega^-(v)} y_{\ell}^{f_1(c),sd} = 0 \tag{15}$$

$$c\in C, sd\in SD: D_{sd}^{c} > 0, v\in V\setminus\{v_s\}$$

$$\sum_{\hat{\gamma}\in\hat{\Gamma}_c} \delta_{sd}^{\hat{\gamma}}\, a_{v_s n_c}^{\hat{\gamma}}\, z_{\hat{\gamma}} + \sum_{\ell\in\omega^-(v_s)} y_{\ell}^{f_{n_c}(c),sd} = 1 \tag{16}$$

$$c\in C, sd\in SD: D_{sd}^{c} > 0$$

$$\sum_{\hat{\gamma}\in\hat{\Gamma}_c} \delta_{sd}^{\hat{\gamma}}\, a_{v n_c}^{\hat{\gamma}}\, z_{\hat{\gamma}} - \sum_{\ell\in\omega^+(v)} y_{\ell}^{f_{n_c}(c),sd} \le 0 \tag{17}$$

$$c\in C, sd\in SD: D_{sd}^{c} > 0, v\in V\setminus\{v_d\}$$

$$\sum_{\hat{\gamma}\in\hat{\Gamma}_c} \delta_{sd}^{\hat{\gamma}}\, a_{v n_c}^{\hat{\gamma}}\, z_{\hat{\gamma}} - \sum_{\ell\in\omega^+(v)} y_{\ell}^{f_{n_c}(c),sd} + \sum_{\ell\in\omega^-(v)} y_{\ell}^{f_{n_c}(c),sd} = 0 \tag{18}$$

$$c\in C, sd\in SD: D_{sd}^{c} > 0, v\in V\setminus\{v_d\}$$

Constraint (3) guarantees that I_c configurations are selected for SFC c. Each $\hat{\gamma}$ is associated with a set of $a_{vi}^{\hat{\gamma}}$ required to be consistent with x_{vf}, which is resolved by constraints (5), (6) where T_{fi}^{c} is to find the VNF f at sequence i in SFC c. Constraint (7) limits the number of VNF replicas. Constraints (8), (9) ensure that each fog or cloud node has a sufficient number of CPU cores for hosting f. Constraint (10) ensure each link has sufficient capacity. Constraint (11) enforce that, for each source-destination pair (v_s, v_d) requesting SFC c, there is exactly one configuration $\hat{\gamma}$. Constraint (12) enforce that, SLA for all source-destinations pair (v_s, v_d) are satisfied. We assume that a unique route exists from v_s to the first VNF location and from v_d to the last VNF location. Equation (13) and (16) guarantee that exactly one outgoing link from v_s (unless first VNF is located at v_s) and one incoming link to v_d (unless last VNF is located at v_d) are selected. Equation (14), (15), (17) and (18) enforce flow-conservation constraints at intermediate nodes. $\omega^+(v)$ and $\omega^-(v)$ represent outcome and income links to node v, respectively.

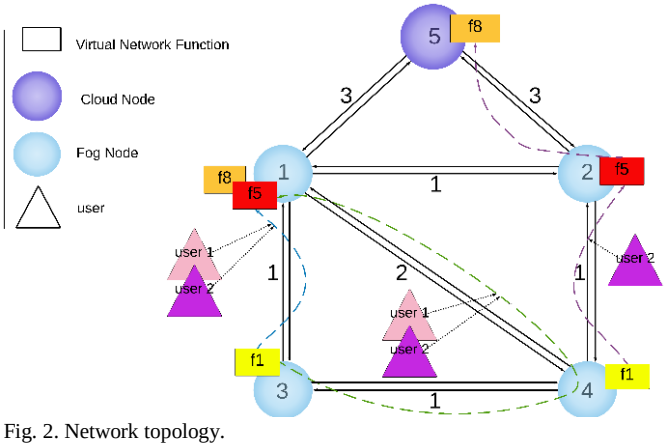

Fig. 2. Network topology.

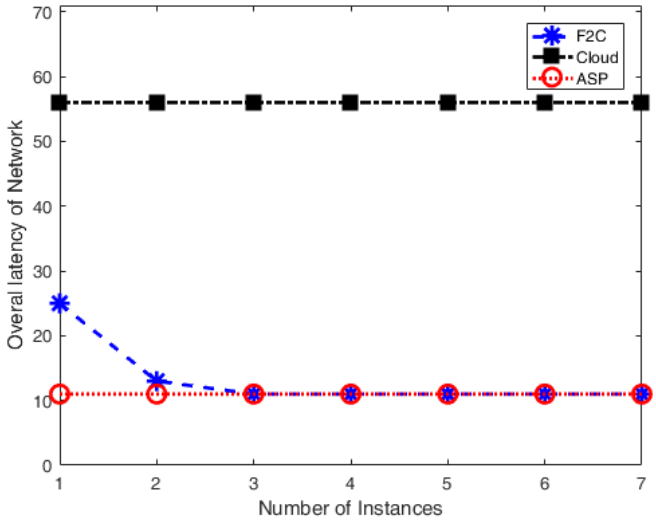

Fig. 3. The overall end-to-end latency of network vs. number of instances.

III. Numerical Results

We tested our ILP optimization process on a network topology Fig. 2 (four fog nodes and one cloud node) with traffic flows between 10 node pairs. Since the distance between cloud and fog is greater than fog and fog, we consider latency of links between fogs and cloud greater than fog and fog. The latency of links between fog and cloud are considered 3 and between fogs are considered 1 except link between fog 1 and 4 which is considered 2. We assume that each VNF uses one core to run. In order to compare F2C and cloud scenario, the links capacities, computing resource, and users' SLA are sufficient to support all flows, traffic demand, and all paths. Each traffic flow is 1Gbps and demand the same 5 VNF service function chain (SFC) for online gaming, as shown in Table II. All Shorter Path (ASP) calculation which assumes all traffic flows requiring an SFC c will have an SFC instance deployed on their shortest path is used to find the least e2e latency. We compare our result with ASP and cloud scenario results. Fig. 3 shows comparing of overall latency of cloud and F2C scenario as the number of SFC instances increases. We find that F2C architecture reduces the e2e latency of users, significantly. Given comparing F2C and ASP result, we find that, as the number of instances increases, F2C results become near to the ASP results. The reason is that when the number of instances increases, ILP can select more instances close to the users, therefore the overall latency decrease.

TABLE II. REQUIREMENTS FOR THE DEPLOYED SERVICE CHAINS[13]

Service Chain	Chained VNFs
Online Gaming	NAT-FW-VOC-WOC-IDPS

NAT: NETWORK ADDRESS TRANSLATOR, FW: FIREWALL, TM: TRAFFIC MONITOR, WOC: WAN OPTIMIZATION CONTROLLER, IDPS: INTRUSION DETECTION PREVENTION SYSTEM, VOC: VIDEO OPTIMIZATION CONTROLLER

IV. CONCLUTION

In this paper, we introduced a novel approach for multi Service Function Chain (SFC) mapping with multiple SFC instances in the fog-to-cloud architecture which is appropriate for the IoT domain. We formulate our problem as an Integer Linear Program to minimize the overall end-to-end latency of IoT devices. we demonstrate that fog-to-cloud architecture reduces overall end-to-end latency in Comparison with cloud scenario and our model can achieve the least overall end-to-end latency when the number of instances increases.

REFERENCES

[1] Pan, J. and J. McElhannon, *Future Edge Cloud and Edge Computing for Internet of Things Applications.* IEEE Internet of Things Journal, 2018. **5**(1): p. 439-449..

[2] Yi, B., et al., *A comprehensive survey of Network Function Virtualization.* Computer Networks, 2018. **133**: p. 212-262.

[3] Bera, S., S. Misra, and A.V. Vasilakos, *Software-Defined Networking for Internet of Things: A Survey.* IEEE Internet of Things Journal, 2017. **4**(6): p. 1994-2008.

[4] Masip-Bruin, X., et al., *Foggy clouds and cloudy fogs: a real need for coordinated management of fog-to-cloud computing systems.* IEEE Wireless Communications, 2016. **23**(5): p. 120-128.

[5] Mijumbi, R., et al., *Network Function Virtualization: State-of-the-Art and Research Challenges.* IEEE Communications Surveys & Tutorials, 2016. **18**(1): p. 236-262.

[6] Mijumbi, R., et al., Management and orchestration challenges in network functions virtualization. IEEE Communications Magazine, 2016. **54**(1): p. 98-105.

[7] Drãxler, S., H. Karl, and Z.Á. Mann. Joint Optimization of Scaling and Placement of Virtual Network Services. in 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID). 2017.

[8] Huin, N., B. Jaumard, and F. Giroire, *Optimal Network Service Chain Provisioning.* IEEE/ACM Transactions on Networking, 2018. **26**(3): p. 1320-1333.

[9] Masri, W., et al. Minimizing delay in IoT systems through collaborative fog-to-fog (F2F) communication. in 2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN). 2017.

[10] Fan, J., et al. *Deadline-Aware Task Scheduling in a Tiered IoT Infrastructure*. in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*. 2017.

[11] Gupta, A., et al. Service Chain (SC) Mapping with Multiple SC Instances in a Wide Area Network. in GLOBECOM 2017 - 2017 IEEE Global Communications Conference. 2017.

[12] Gupta, A., et al., A Scalable Approach for Service Chain Mapping With Multiple SC Instances in a Wide-Area Network. IEEE Journal on Selected Areas in Communications, 2018. **36**(3): p. 529-541.

[13] Savi, M., M. Tornatore, and G. Verticale. Impact of processing costs on service chain placement in network functions virtualization. in 2015 IEEE Conference on Network Function Virtualization and Software Defined Network (NFV-SDN). 2015.