

Distributed VNF Scaling in Large-scale Datacenters: An ADMM-based Approach

Farzad Tashtarian^{*}, Amir Varasteh[†], Ahmadreza Montazerolghaem[§], and Wolfgang Kellerer[†]

^{*} Department of Computer Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran
{Email: f.tashtarian@mshdiau.ac.ir}

[†] Chair of Communication Networks, Department of Electrical and Computer Engineering,
Technical University of Munich, Munich, Germany
{Email: amir.varasteh, wolfgang.kellerer@tum.de}

[§] Department of Computer Engineering, Ferdowsi University, Mashhad, Iran
{Email: ahmadreza.montazerolghaem@stu.um.ac.ir}

Abstract—Network Functions Virtualization (NFV) is a promising network architecture where network functions are virtualized and decoupled from proprietary hardware. In modern datacenters, user network traffic requires a set of Virtual Network Functions (VNFs) as a service chain to process traffic demands. Traffic fluctuations in Large-scale DataCenters (LDCs) could result in overload and underload phenomena in service chains. In this paper, we propose a distributed approach based on Alternating Direction Method of Multipliers (ADMM) to jointly load balance the traffic and horizontally scale up and down VNFs in LDCs with minimum deployment and forwarding costs. Initially we formulate the targeted optimization problem as a Mixed Integer Linear Programming (MILP) model, which is NP-complete. Secondly, we relax it into two Linear Programming (LP) models to cope with over and underloaded service chains. In the case of small or medium size datacenters, LP models could be run in a central fashion with a low time complexity. However, in LDCs, increasing the number of LP variables results in additional time consumption in the central algorithm. To mitigate this, our study proposes a distributed approach based on ADMM. The effectiveness of the proposed mechanism is validated in different scenarios.

Keywords—Network function virtualization, datacenters, VNF scaling, service chaining, distributed optimization, alternating direction method of multipliers (ADMM).

I. INTRODUCTION

Today, different technologies such as Software Defined Networking (SDN) and Network Function Virtualization (NFV) are being integrated to facilitate the modern datacenter service providers [1]. In these datacenters, the network services are frequently deployed as Virtual Network Function (VNF) chains to process the incoming traffic [2] (see Fig. 1). The NFV architecture relies on off-the-shelf Physical Machines (PMs) which provide computational resources to several Virtual Machines (VMs), each VM implementing a network function with special software programs. In addition, SDN technology could manage the communication network between these VNFs in the datacenter.

Due to traffic fluctuations in datacenters, resource over and underload can occur in service chains. The impact of an overload could be the degradation of overall performance and SLA violations, while underload could lead to significant resource wastage. One of the most popular methods to ensure

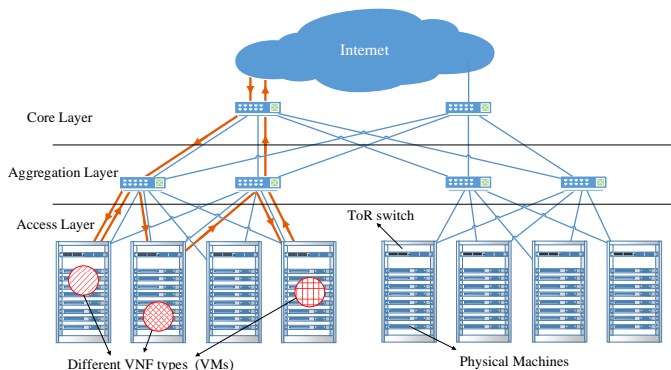


Fig. 1: A service chain with three VNFs in a conventional datacenter

efficient resource management, is the appropriate placement of VNFs on PMs, which is called the VNF placement problem [3]. This problem has been addressed in literature regarding various parameters, such as network delay and end-to-end latency [4], [5], and operational costs of VNFs (deployment cost, forwarding/processing cost, and energy consumption) [6], [7], and [8]. Although the majority of studies focus on the initial placement of the service chain, network-aware resource management of the service chains during their operational time has emerged as a critical issue. To overcome this issue, the traffic could be load balanced between VNF instances, or the over and underloaded VNF(s) could be scaled up and down. In this paper, we jointly load balance the traffic and horizontally scale up and down VNFs with minimum deployment and forwarding costs. Notably, we use thresholds to automatically detect over and underloaded VM(s), respectively.

Based on the above considerations, the problem is modeled using Mixed Integer Linear Programming (MILP) which is NP-complete [9]. Therefore, we relax our MILP model into two LP formulations. These models perform well in small and medium size datacenters. However, in Large-scale DataCenters (LDCs), the number of variables and constraints of the LP models escalate significantly resulting in an increase in the time complexity of the algorithm, which requires the use of a distributed approach. Hence, we extend our approach to distributed optimization, using a technique known as Alternating Direction Method of Multipliers (ADMM) [10]. In addition to simplicity and its powerfulness, this method has several advantages like fast convergence, low computational complexity, and low overhead of message passing [10]. Using

this method, the convex optimization problem is decomposed into different subproblems, each of which is easier to handle. These subproblems are solved iteratively and in parallel using defined agents [10].

Although the conventional ADMM is guaranteed to converge for two-block optimization problems, for n -block variable problems (i.e. multi-block variables) with $n > 2$, updating different variables sequentially has no convergence guarantee [11]. Therefore, in this paper, because our problem is in form of multi-block ADMM, we use a specific ADMM extension called Randomly Permuted ADMM (RP-ADMM) [12] to solve the problem in a distributed way.

Afterall, considering the aforementioned explanations, the main contributions of our proposed work are described as follows:

- 1) Introducing a joint traffic load balancing and VNF scaling mechanism to mitigate both over and underloaded situations in datacenters,
- 2) Formulating the above problem as a MILP optimization model, and proposing two LP-relaxations to address time complexity,
- 3) Extending our approach to a distributed method based on the RP-ADMM technique,
- 4) Evaluation of the proposed work in different scenarios via simulation.

The remainder of this paper is organized as follows, we start by presenting the related work in the literature (section II). Then, in sections III and IV we illustrate the system model and problem formulation, respectively. Next, the proposed approaches are described in section V. We validate our approach in section VI, and finally, section VII concludes the whole paper.

II. RELATED WORK

VM placement has been widely studied in the cloud computing environment [13], [14], [15], and [3]. However, specific characteristics and constraints of the NFV paradigm such as service chaining requirements, deployment and license costs, etc., has made the VNF placement problem more complex. Several studies exist in the literature that tackle this problem. For instance, StEERING [16] focused on VNF chain placement and traffic steering and uses a simple heuristic algorithm to solve the problem. Bari *et. al.* [6] proposed a model that can be used to determine the optimal number of VNFs and to place them at optimal locations (PMs). In [7], authors modeled this problem to minimize client-server communication distance and operational costs. Also, [17] presented a joint traffic steering and VNF placement using an Integer Linear Programming (ILP) model which sought to minimize the total delay of the traffic. While these approaches largely focus on initial placement of VNFs, they do not monitor VNF chains during their operational time. Hence, because of traffic fluctuations, a resource over or underload could happen in VNF chains, which is a significant challenge to address.

The work most similar to our approach is *Stratos* [18]. It uses a four-step solution attempting to solve the VNF chain bottlenecks during its operational time. When an overload happens, it initially uses a flow distribution method. When

this method fails to solve the bottleneck, *Stratos* attempts to migrate the VNFs. If the problem persists, *Stratos* scales up the bottlenecked VNFs. Finally, when all these procedures failed, it scales up all the VNFs in the chain with a fixed number of instances. The primary drawback of *Stratos* is that it works in a trial manner, which creates resource wastage and inefficiencies to solve the resource bottlenecks. Another close work to us is [19] which attempts to tackle dynamic provisioning of one or multiple service chains in cloud datacenters in a centralized and time-slotted manner. However, [19] does not address the forwarding cost in their model. This could bring additional growth of east-west traffic in the datacenter which increases the wastage of datacenter computation resources [20].

In contrast to afore-mentioned works, we introduce a joint traffic load balancing and VNF scaling mechanism to deal with over and underload phenomena during service chain operational time, while effectively considering forwarding and deployment costs in form of a mathematical optimization model. To the best of our knowledge, there is no literature that is applicable in LDCs. Therefore, we extend the model to a distributed form of optimization, using RP-ADMM method, to be able to use this approach in LDCs.

III. SYSTEM MODEL

We consider a datacenter as a graph $\mathcal{G} = \{\mathbb{V}, \mathbb{L}\}$, where the vertices set V consists of U physical machines $\mathcal{P}_i, i = 1 \dots U$ and S switches $\mathcal{S}_i, i = 1 \dots S$, and \mathbb{L} represents edges of \mathcal{G} where $l_{ij} = 1$ if a direct communication link exists between two entities $i, j \in \mathbb{V}$. For the ease of problem formulation, we assume a conventional network topology for the datacenter which consists of three layers: access layer, aggregation layer, and core layer (see Fig. 1). The total available resources of a PM is considered as \mathcal{P}_i^r where $r \in R$ and R is the set of limited types of PM resources such as CPU, memory, bandwidth, etc. Let C be a set of service chains that each of them is constituted by an ordered sequence of VNFs f_k , where $k \in K$ and K is the set of various network functions such as Firewall, IDS, NAT, etc.

We assume that each VM is providing only one network function. Also, we assume all configured VMs providing network function k , are using an identical computational resource. Each VM i that belongs to chain $c \in C$ and provides f_k , denoted by $v_k^{i,c}$, is configured with initial resource $u_{k,r}^{i,c}$ ($r \in R$).

We use the notation $g_k^c = \{v_j^{i,z} \mid \forall j = k, z = c\}$ indicating a set of VMs that provide f_k on chain $c \in C$. In addition, *ingress* and *egress* points of g_k^c are defined as g_{k-1}^c and g_{k+1}^c , where $k+1$ and $k-1$ refer to the *previous* and *next* service function of f_k in chain c , respectively. We consider two main phases for each chain: *setup* phase and *operational* phase. In setup phase, the requested service chain is deployed in datacenter network with respect to SLA and the available resources of the datacenter. In the second phase, the chain begins to process the incoming traffic, which rate may vary from time to time. Therefore, by having an initial VNF placement for service chains, in this paper, we focus on the second phase of chain lifetime and address the problem of managing the over and underloaded chain with the minimum deployment and forwarding costs.

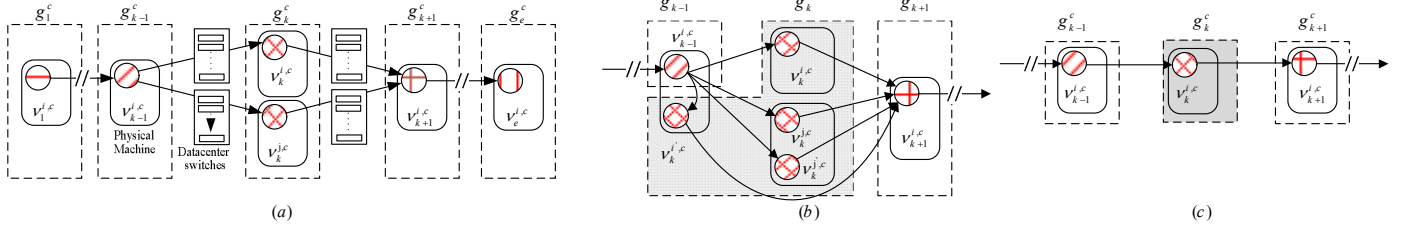


Fig. 2: VNF service chain states in a datacenter network, (a) Normal state: an operational VNF chain in a network; (b) Overload occurrence: VMs i' and j' are launched to manage the overload situation; and (c) Underload occurrence: VM j is turned off to optimize resource usage efficiency in the VNF chain.

Our proposed system architecture includes four main components: Monitoring, SDN Controller, VNF Placement and Scaling Module, and Resource Management (REM). Since the proposed approach should respond efficiently and quickly to traffic fluctuations, we use monitoring component to continuously collect the amount of CPU, memory, and bandwidth usage of VMs. In addition to the reported data by monitoring component, the SDN controller collaborates with REM through sharing the routing paths and configuration of network switches. It also updates and configures the routing policy using the REM output. We define V and V^* as the set of running VMs in g_k^c (*online VMs*), and the set of configured VMs to join g_k^c (*offline VM pool*), respectively. Also, let $\tilde{u}_{k,r}^{i,c}$ be the utilization of resource type r of $v_k^{i,c}$ in τ units of time that is measured by monitoring module. Further, we define $\hat{u}_{k,r}^{i,c}$, $\bar{u}_{k,r}^{i,c}$, and $\check{u}_{k,r}^{i,c}$ as *hot*, *warm*, and *cold* thresholds on resource type r for $v_k^{i,c}$ to enable us to specify the state of a chain, respectively, as follows:

(1) *overload* state:

$$\exists k : g_k^c \in c \mid \exists i : v_k^{i,c} \in g_k^c \text{ and } \tilde{u}_{k,r}^{i,c} \geq \hat{u}_{k,r}^{i,c}$$

(2) *underload* state:

$$\exists k : g_k^c \in c \mid \frac{1}{|V|} \sum_{i: v_k^{i,c} \in g_k^c} \tilde{u}_{k,r}^{i,c} \leq \check{u}_{k,r}^{i,c} \\ \text{and } \max\{\tilde{u}_{k,r}^{i,c} \mid \forall i : v_k^{i,c} \in g_k^c\} \leq \bar{u}_{k,r}^{i,c}$$

(3) *normal* state: If chain c is neither in overload nor underload state, it is considered being in the *normal* state.

For illustration, suppose chain c consists of three functions: f_1 , f_2 , and f_3 , where $v_1^{1,c}$, $\{v_2^{1,c}, v_2^{2,c}\}$, and $v_3^{1,c}$ are serving these functions, respectively. We assume 50%, 65% as CPU utilization of $v_1^{1,c}$ and $v_3^{1,c}$, respectively. In this example, we focus on f_2 to show how the state of the chain c can be determined. Therefore, we consider the following two scenarios for different CPU utilization of the VMs serving f_2 . Also, we assume *hot*, *warm*, and *cold* thresholds as 90%, 80% and 30%, respectively. Scenario (a) $\{\tilde{u}_{2,[cpu]}^{1,c} = 95\%, \tilde{u}_{2,[cpu]}^{2,c} = 75\%\}$: Since the CPU utilization of $v_2^{1,c}$ violates the *hot* threshold, chain c falls into the *overload* state. We note that the presence of only one violated VM is sufficient to put a chain in the *overload* state. Scenario (b) $\{\tilde{u}_{2,[cpu]}^{1,c} = 40\%, \tilde{u}_{2,[cpu]}^{2,c} = 15\%\}$: In this scenario, it can be seen that the average of CPU utilizations is less than the cold threshold. Consequently, this scenario makes chain c to be considered as *underload*. Notably, f_1 and f_3 cannot be in *underload* state, since there is only one VM assigned for each of these functions.

Remarkably, the values of *hot*, *warm* thresholds should be opted in such a way that the gap between their values provides a safety margin to cope with the temporary network traffic

fluctuations. Additionally, different types of resources can have different thresholds. For example, the hot thresholds for CPU and memory resources can be set to 90% and 80%, respectively [21]. In Fig. 2, we illustrate the three operational states for a sample chain.

IV. PROBLEM FORMULATION

In this section, we address the problem of overload and underload events that occur in network service chain. First, we present a central optimization model, and then extend our study by proposing a distributed model using the ADMM technique. Table I shows the main notations.

To deal with the over/underloaded chain c , a number of constraints must be satisfied. Suppose g_k^c is identified as an over/underloaded set in chain c . Let b_{ij} be a binary decision variable, where if $b_{ij} = 1$, then VM $j \in \{V \cup V^*\}$ must be running on \mathcal{P}_i . We denote α_{ij} as the percentage of collaboration of VM $j \in \{V \cup V^*\}$ on \mathcal{P}_i , with existing VMs $\in V$ to process the total traffic coming from g_{k-1}^c . The value of α_{ij} must be restricted as follows:

$$\varepsilon b_{ij} \leq \alpha_{ij} \leq \varphi_k b_{ij}, \quad \forall j \in \{V \cup V^*\}, \forall i \in \mathcal{P} \quad (1)$$

where ε shows the minimum amount of collaboration that causes a VM to be launched on a PM. Additionally, φ_k shows the maximum percentage of service capacity of VMs running network function k in such a way that $\forall j \in \{V \cup V^*\}, \forall r \in R$, $\varphi_k \omega_k^r T < u_{k,r}^{j,c}$, where ω_k^r is defined as the service impact factor of network function k on resource $r \in R$. Let n_{ij}^d and m_{ij}^d , ($i, j \in \{S \cup \mathcal{P}\}$ and $d \in \{V \cup V^*\}$), be the amount of unprocessed traffic rates destined to $v_k^{d,c}$, and processed traffic by $v_k^{d,c}$ transmitting from i to j , respectively. The following constraints state that all online VMs must cover the total traffic flow originated by ingress point g_{k-1}^c :

$$\sum_{j \in \{V \cup V^*\}} \sum_{i \in \mathcal{P}} \alpha_{ij} = 1 \quad (2)$$

$$\sum_{s \in \{S \cup \mathcal{P}\}} l_{si} n_{si}^j = T \alpha_{ij}, \quad \forall j \in \{V \cup V^*\}, \forall i \in \mathcal{P} \quad (3)$$

where T is the total traffic coming from g_{k-1}^c (i.e. *ingress* point) to the over/underloaded g_k^c . The following constraint indicates that the total required resources for deploying VM(s) on \mathcal{P}_i must not be more than its available resources, \mathcal{P}_i^r :

$$\sum_{j \in \{V \cup V^*\}} b_{ij} u_{k,r}^{j,c} \leq \mathcal{P}_i^r, \quad \forall i \in \mathcal{P}, r \in R \quad (4)$$

Additionally, $\forall i \in S$ and $\forall d \in \{V \cup V^*\}$, the next constraint is presented as follows:

$$\sum_{j \in \{S \cup \mathcal{P}\}} l_{ij} (n_{ij}^d - n_{ji}^d) = 0 \quad (5)$$

$$\sum_{j \in \{S \cup \mathcal{P}\}} l_{ij} (m_{ij}^d - m_{ji}^d) = 0 \quad (6)$$

TABLE I: Notations

Notation	Description
\mathcal{S}_i	The i^{th} switch in datacenter network
\mathcal{P}_i	The i^{th} physical machine in datacenter network
R	The set of resource types (CPU, Memory, etc.)
K	The set of network function types
f_k	The network function type $k \in K$
l_{ij}	$l_{ij} = 1$ if there is a link between two entities $i, j \in \{\mathcal{S} \cup \mathcal{P}\}$
\mathcal{P}_i^r	The total available amount of resource $r \in R$ in \mathcal{P}_i
$v_k^{i,c}$	VM i that belongs to chain c and hosts f_k
g_k^c	The set of VMs that provide f_k on chain c
g_{k-1}^c, g_{k+1}^c	The ingress/egress points of g_k^c
$\tilde{u}_{k,r}^{i,c}$	The utilization of r^{th} resource of $v_k^{i,c}$
$\hat{u}_{k,r}^{i,c}, \bar{u}_{k,r}^{i,c}, \check{u}_{k,r}^{i,c}$	Hot, warm, and cold thresholds for r^{th} resource of $v_k^{i,c}$
$u_{k,r}^{i,c}$	The initial resource type r for $v_k^{i,c}$
n_{ij}^d	The amount of unprocessed traffic destined to $v_k^{d,c}$ transmitting from i to j , where $i, j \in \{\mathcal{S} \cup \mathcal{P}\}$
m_{ij}^d	The amount of processed traffic by $v_k^{d,c}$ transmitting from i to j where $i, j \in \{\mathcal{S} \cup \mathcal{P}\}$
T	The total traffic received by g_k^c
γ_k	Traffic changing factor of f_k
α_{ij}	The percentage of collaboration of $v_k^{j,c}$, hosted on \mathcal{P}_i , with other VMs providing identical network function and belong to g_k^c
φ_k	The maximum percentage of service capacity of VMs running network function k
ω_k^r	The service impact factor of network function k on resource $r \in R$
b_{ij}	$b_{ij} = 1$ if VM j is hosted by \mathcal{P}_i
w_{ij}	Total available bandwidth of link (i, j) , $i, j \in \{\mathcal{S} \cup \mathcal{P}\}$
v^*	The number of required VNFs to process the received traffic by g_k^c
V, V^*	Online VMs belongs to g_k^c , offline VM pool for g_k^c
Ψ, Ψ^*	The set of PMs hosting V and candidate PMs for hosting V^*
e_{ij}	The amount of interest of each PM $\in \{\Psi \cup \Psi^*\}$ in processing the incoming traffic

These constraints state that the total incoming (processed and unprocessed) traffic to \mathcal{S}_i must be equal to the total outgoing traffic from it. A similar constraint $\forall i \in \mathcal{P}$ and $\forall d \in \{V \cup V^*\}$ could be written as follows:

$$\sum_{j \in \{\mathcal{S} \cup \mathcal{P}\}} l_{ij} (m_{ij}^d - \gamma_k n_{ji}^d) = 0 \quad (7)$$

where $\gamma_k > 0$ is data changing factor of f_k serviced by g_k^c [22]. Although a connection between any pairs of PMs in the datacenter must be established via a switch, the data flow between two different types of VMs, both hosted by an identical PM, is applicable. The following constraints guarantee that the total traffic $T\gamma_k$, processed by g_k^c , is received by g_{k+1}^c (i.e. egress point):

$$\sum_{d \in \{V \cup V^*\}} \sum_{i: \mathcal{P}_i \in g_{k-1}^c} \sum_{j \in \{\mathcal{S} \cup \mathcal{P}\}} l_{ij} n_{ij}^d = T \quad (8)$$

$$\sum_{d \in \{V \cup V^*\}} \sum_{i: \mathcal{P}_i \in g_{k+1}^c} \sum_{j \in \{\mathcal{S} \cup \mathcal{P}\}} l_{ji} m_{ji}^d = T\gamma_k \quad (9)$$

Moreover, $\forall i, j \in \{\mathcal{S} \cup \mathcal{P}\}$ a bandwidth constraint could be formulated as follows:

$$\sum_{d: v_k^{d,c} \in g_k^c} l_{ij} (n_{ij}^d + m_{ij}^d) \leq w_{ij} \quad (10)$$

where w_{ij} is the total available bandwidth of link (i, j) . To evaluate the performance of any solutions that satisfies the above constraints, we now propose the following two cost functions:

Deployment Cost :

$$\sum_{i \in \mathcal{P}} \sum_{j \in V^*} b_{ij} \mathbb{X} - \sum_{i \in \mathcal{P}} \sum_{j \in V} b_{ij} \hat{\mathbb{X}} \quad (11)$$

Forwarding Cost :

$$\sum_{d \in \{V \cup V^*\}} \sum_{i \in \{\mathcal{S} \cup \mathcal{P}\}} \sum_{j \in \{\mathcal{S} \cup \mathcal{P}\}} F(i, j) (n_{ij}^d + m_{ij}^d) \quad (12)$$

Also, \mathbb{X} and $\hat{\mathbb{X}}$ are considered as two penalty values in the deployment cost formulation. We use the deployment cost to control the power states, migration, and placement of VMs. In fact, it can be utilized for both over and underload situations by adjusting the proper values \mathbb{X} and $\hat{\mathbb{X}}$ (see Proposition. 1 for proof). Further, we use the forwarding cost to force the model to deliver traffic via efficient routes. To achieve this, we should prevent the traffic from being transferred through higher layers of network in the topology (i.e. aggregation and core layers, see Fig. 1). Hence, we defined $F(i, j)$ as the forwarding cost which applies on transmitting composite bit-stream per unit of time, and it is proportional to layers of datacenter topology in which i and j are located. Finally, the following MILP is proposed to jointly balance the traffic load and scale VNFs with the minimum costs:

$$\text{minimize} \quad \aleph_1 Eq.(11) + \aleph_2 Eq.(12) \quad (13)$$

s.t.

Constraints (1) – (10)

vars. $b_{ij} \in \{0, 1\}, \alpha_{ij}, n_{ij}^d, m_{ij}^d \geq 0$

where \aleph_1 and \aleph_2 are two constant weights for deployment and forwarding costs.

Proposition 1. *The proposed MILP model (13) could be used to mitigate both over and underloaded service chains with adjusting the appropriate penalty values of \mathbb{X} and $\hat{\mathbb{X}}$.*

Proof: Suppose g_k^c is an over/underloaded set which serves VNF type k in chain c . In Eq. (11), we presented the deployment cost that consists of two terms: the first term handles the online VMs, $v_k^{i,c} \in g_k^c$, and the second term

focuses on offline VM instances of k which are ready to be transferred to a PM and start servicing in g_k^c . In the presence of an overload, there is a solution that jointly balances the traffic rates among the online VMs, and launches one or more offline VMs. Considering overall cost, by choosing two positive values for \mathbb{X} and $\hat{\mathbb{X}}$, where $\hat{\mathbb{X}} \gg \mathbb{X}$, the MILP overcomes overload using flow distribution on online VMs, horizontal scaling, and migration, respectively. On the other hand, when an underload is detected, the best solution is to turn off one or more online VMs and load balance the traffic rates among the remained online VMs. To preserve the resource efficiency, there is no need to turn on new VM instances, unless migrating a VM dramatically decreases the forwarding cost. Hence, the value of $\hat{\mathbb{X}}$ must be set to a large negative number while \mathbb{X} should remain as a very large positive number. ■

Considering the aforementioned observations, by detecting over/underloaded chain, the REM adjusts the proper values for \mathbb{X} and $\hat{\mathbb{X}}$ and triggers the optimization algorithm.

V. THE PROPOSED APPROACH

Since the proposed MILP model (13) is NP-complete and suffers from high time complexity [9], we consider two linear relaxations of the MILP model to address over and

underloaded VNFs. These LP models could be centrally run in small or even medium size datacenters. However, using these models in LDCs is not applicable due to increasing the number of variables and constraints. Thus, we extend our study by proposing a distributed mechanism based on the RP-ADMM technique.

A. The Central Design

In case of an overloaded g_k^c , suppose that the number of required VNFs f_k , which is denoted by v^* , to process the received traffic equals to $\text{Max}\{[T\omega_k^r/u_{k,r}^{j,c}] \mid \forall r \in R\}$. Also, we define Ψ and Ψ^* sets, which are the set of PMs that are hosting V , and the set of candidate PMs that has adequate resources to host at least one VM $\in V^*$, respectively. The relaxed overload model of MILP (13) is presented as follows:

The overload LP model:

$$\text{minimize} \quad \text{Eq. (12)} \quad (14)$$

s.t.

$$\text{Constraints (5) - (9)} \quad (\text{I})\text{-(V)}$$

$$\alpha_{ij} - \alpha_{qj} = 0, \quad i = |V| + 1 : v^*, \forall j, q \in \{\Psi \cup \Psi^*\} \quad (\text{VI})$$

$$\sum_{q \in \{\Psi \cup \Psi^*\}} e_{qj} - \alpha_{ij} = 0, \quad i = 1 : v^*, \forall j \in \{\Psi \cup \Psi^*\} \quad (\text{VII})$$

$$\sum_{s \in \{\text{SU}\mathcal{P}\}} l_{si} n_{si}^j - T e_{ij} = 0, \quad i = |V| + 1 : v^*, \forall j \in \{\Psi \cup \Psi^*\} \quad (\text{VIII})$$

$$\sum_{s \in \{\text{SU}\mathcal{P}\}} l_{si} n_{si}^j - T \alpha_{ij} = 0, \quad i = 1 : |V|, \forall j \in \Psi \quad (\text{IX})$$

$$\text{vars.} \quad n_{ij}^d, m_{ij}^d \geq 0, 0 \leq \alpha_{ij} \leq \varphi_k, 0 \leq e_{ij} \leq 1$$

Since the number of desired VMs is known, the deployment cost in the objective function could be omitted. In fact, having v^* , the relaxed overload model (14) specifically focuses on the forwarding cost. In this model, the routing constraints could be used as they were before, except the bandwidth constraint (Eq. (10)). We ignored the bandwidth constraint in the proposed relaxed models. However, we will relax this assumption in our future work. Additionally, to relax the binary variable b_{ij} in the MILP model (13), we define $e_{ij} \in \mathbb{R}^+$ as a new variable in which $j = 1 : v^*$, and $i \in \{\Psi \cup \Psi^*\}$. This variable identifies the amount of interest of each PM in processing the incoming traffic. For each required instance j , the PM with the highest value of e_{ij} is selected to run the j^{th} VM instance and handle $T\alpha_{ij}$ of traffic (constraints (VI)-(VIII)). Finally, constraint (IX) determines the amount of received traffic that must be processed by each PM $i \in \Psi$.

On the other hand, when an underload occurs, the following LP-relaxation model, which can be obtained easily from the above model, runs on the first v^* online VMs to appropriately assign them to PMs $\in \Psi$ and distribute received load from g_{k-1}^c on g_k^c :

The underload LP model:

$$\text{minimize} \quad \text{Eq. (12)} \quad (15)$$

s.t.

$$\text{Constraints (5) - (9)} \quad (\text{I})\text{-(V)}$$

$$\alpha_{ij} - \alpha_{qj} = 0, \quad j = 1 : v^*, \forall i, q \in \Psi \quad (\text{VI})$$

$$\sum_{q \in \Psi} e_{qj} - \alpha_{ij} = 0, \quad j = 1 : v^*, \forall i \in \Psi \quad (\text{VII})$$

$$\sum_{s \in \{\text{SU}\mathcal{P}\}} l_{si} n_{si}^j - T e_{ij} = 0, \quad i = 1 : v^*, \forall j \in \Psi \quad (\text{VIII})$$

$$\text{vars.} \quad n_{ij}^d, m_{ij}^d \geq 0, 0 \leq \alpha_{ij} \leq \varphi_k, 0 \leq e_{ij} \leq 1$$

For clarification on details of the proposed models, an example for overloaded and underloaded chains is shown in Fig. 3. In Fig. 3(a), the overloaded g_k^c has two VMs $(1, 2) \in V$, which are running on $(\mathcal{P}_1, \mathcal{P}_2) \in \Psi$, respectively. By assuming overloaded VM 2 and setting $v^* = 4$, two extra VMs $(3, 4) \in V^*$ with the same VNF type must be added to g_k^c regarding $(\mathcal{P}_2, \mathcal{P}_3, \mathcal{P}_4) \in \Psi^*$. Obtained α_{ij} and e_{ij} , VMs 3 and 4 are selected to be launched on \mathcal{P}_3 and \mathcal{P}_2 , respectively. Additionally, in Fig. 3(b), a final decision has been made based on the calculated values of e_{ij} and α_{ij} for the underloaded chain. In this case, the relaxed underload LP model turns off VM 3 running on \mathcal{P}_2 .

B. Introducing ADMM

After many years without too much attention, the ADMM technique has recently witnessed a renaissance in many application areas [11]. Generally, there are two types of distributed methods: Sub-gradient and ADMM. The best known rate of convergence for sub-gradient methods is $O(\frac{1}{\sqrt{\ell}})$, while the ADMM algorithm converges at the rate $O(\frac{1}{\ell})$ for the general case, where ℓ is the number of iterations [23]. Therefore, since both approaches are iteration-based and the speed of convergence in the ADMM is more than the sub-gradient method, we develop our distributed approach based on the ADMM algorithm.

We first introduce the background of the ADMM and then apply it to design our distributed algorithm. Consider the following convex minimization problem with a separable objective function and linear constraints:

$$\begin{aligned} \text{minimize} \quad & f_1(x_1) + f_2(x_2) + \dots + f_n(x_n), \\ \text{s.t.} \quad & A_1 x_1 + A_2 x_2 + \dots + A_n x_n = b, \\ \text{vars.} \quad & x_i \in \mathcal{X}_i, i = 1, \dots, n, \end{aligned} \quad (16)$$

where $A_i \in \mathbb{R}^{N \times d_i}$, $b \in \mathbb{R}^{N \times 1}$, $\mathcal{X}_i \subseteq \mathbb{R}^{d_i}$ is a closed convex set, and $f_i : \mathbb{R}^{d_i} \rightarrow \mathbb{R}$ is a closed convex function, $i = 1, \dots, n$ [12]. In case of $n = 2$ (i.e. two blocks of variables), the augmented Lagrangian function for model (16) can be written as follows:

$$\begin{aligned} L_\beta(x_1, x_2, \mu) = & f_1(x_1) + f_2(x_2) + \mu^\top (A_1 x_1 + A_2 x_2 - b) \\ & + \frac{\beta}{2} \|A_1 x_1 + A_2 x_2 - b\|_2^2 \end{aligned} \quad (17)$$

where μ is the Lagrangian multiplier and β is the positive penalty scalar. The two-block ADMM updates the primal variables x_1 and x_2 , followed by a dual variable update μ in an iterative manner. By having (x_1^0, x_2^0, μ^0) as an initial vector, the updated variables at iteration $\ell > 0$ are computed in an alternating fashion as follows:

$$x_1^\ell = \underset{x_1}{\text{argmin}} L_\beta(x_1, x_2^{\ell-1}, \mu^{\ell-1}) \quad (18)$$

$$x_2^\ell = \underset{x_2}{\text{argmin}} L_\beta(x_1^\ell, x_2, \mu^{\ell-1}) \quad (19)$$

$$\mu^\ell = \mu^{\ell-1} - \beta(A_1 x_1^\ell + A_2 x_2^\ell - b) \quad (20)$$

Separable structure of convex model (16) allows ADMM to decompose it over its primal variables and achieve the optimal solution in a distributed way. To formulate the ADMM into a scaled form that is more convenient to be used in designing algorithm, we define $u = (\frac{1}{\beta})\mu$ as the scaled dual variable. Thus, we have:

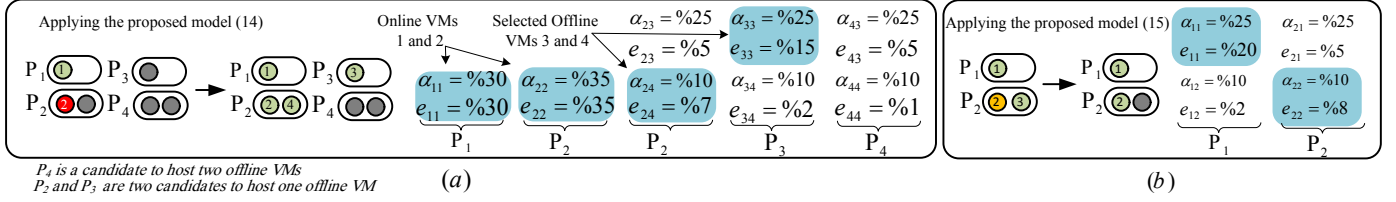


Fig. 3: An example of over and underloaded chains and their solution, (a) The overload case (b) The underload case

$$\begin{aligned} & \mu^\top (A_1 x_1 + A_2 x_2 - b) + \frac{\beta}{2} \|A_1 x_1 + A_2 x_2 - b\|_2^2 \\ & = \frac{\beta}{2} \|A_1 x_1 + A_2 x_2 - b + u\|_2^2 - \frac{\beta}{2} \|u\|_2^2 \end{aligned} \quad (21)$$

By disregarding independent terms of the minimization variables, the ADMM method could be expressed as:

$$x_1^\ell = \underset{x_1}{\operatorname{argmin}} (f_1(x_1) + \left(\frac{\beta}{2}\|A_1 x_1 + A_2 x_2^{\ell-1} - b + u^{\ell-1}\|_2^2\right)) \quad (22)$$

$$x_2^\ell = \underset{x_2}{\operatorname{argmin}} (f_2(x_2) + \left(\frac{\beta}{2}\|A_1 x_1^\ell + A_2 x_2 - b + u^{\ell-1}\|_2^2\right)) \quad (23)$$

$$u^\ell = u^{\ell-1} + A_1 x_1^\ell + A_2 x_2^\ell - b \quad (24)$$

The optimality and convergence of the two-block ADMM for model (16) is proved in [10] under mild technical assumptions.

On the other hand, many applications need more than two blocks of variables to be considered (i.e. $n > 2$). By direct extending of ADMM to multi-block case, the convergence cannot be established well [11]. The authors in [12] proposed Randomly Permuted ADMM (RP-ADMM), as a simple modification in updating process of primal variables to make the convergence possible in a desirable duration. In fact, RP-ADMM draw a permutation Ω of $\{1, \dots, n\}$ uniformly at random and update the primal variables in the order of the permutation, followed by updating the dual variables in a usual way [12]. Since our proposed approach has more than two blocks of variables, we apply RP-ADMM technique to solve the problem in a distributed way.

C. Applying RP-ADMM

In this section, we develop our distributed algorithm based on the proposed model (14). We firstly reformulate the proposed model (14) to an applicable model for applying RP-ADMM, and then introduce our RP-ADMM-based distributed algorithm.

Since $\forall j \in \{\mathcal{S} \cup \Psi \cup \Psi^*\}$, each node $i \in \{\mathcal{S} \cup \Psi \cup \Psi^*\}$ is responsible to determine its local flow variables n_{ij}^d and m_{ij}^d in a distributed manner, we have to reformulate flow conservation constraints (I)-(V) in (14). So, by introducing auxiliary variables A_{ij}^d , B_{ij}^d , and C_{ij}^d , we reform constraints (I)-(III), $j \in \{\mathcal{S} \cup \Psi \cup \Psi^*\}$, $d = 1 : v^*$, where $l_{ij} = 1$, as follows:

$$n_{ij}^d - n_{ji}^d - A_{ij}^d = 0, \quad \forall i \in \mathcal{S} \quad (25)$$

$$\sum_{j \in \{\mathcal{S} \cup \Psi \cup \Psi^*\}} A_{ij}^d = 0 \quad (26)$$

$$m_{ij}^d - m_{ji}^d - B_{ij}^d = 0, \quad \forall i \in \mathcal{S} \quad (27)$$

$$\sum_{j \in \{\mathcal{S} \cup \Psi \cup \Psi^*\}} B_{ij}^d = 0 \quad (28)$$

$$m_{ij}^d - \gamma_k n_{ji}^d - C_{ij}^d = 0, \quad \forall i \in \mathcal{P} \quad (29)$$

$$\sum_{j \in \{\mathcal{S} \cup \Psi \cup \Psi^*\}} C_{ij}^d = 0 \quad (30)$$

Without loss of generality, assume that ingress and egress points of g_k^c have one PM indexed by 0 and 1, respectively. To reform constraints (IV) and (V) in (14), we similarly define two more auxiliary variables D_d and E_d , ($d = 1 : v^*$):

$$\sum_{j \in \{\mathcal{S} \cup \mathcal{P}_0\}} l_{0j} n_{0j}^d - D_d = 0 \quad (31)$$

$$\sum_{d=1:v^*} D_d - T = 0 \quad (32)$$

$$\sum_{j \in \{\mathcal{S} \cup \mathcal{P}_1\}} l_{j1} m_{j1}^d - E_d = 0 \quad (33)$$

$$\sum_{d=1:v^*} E_d - T \gamma_k = 0 \quad (34)$$

Now, the RP-ADMM can be easily applied on the following model:

$$\begin{aligned} & \text{minimize} && Eq.(12) \\ & \text{s.t.} && Eq.(25) - Eq.(34), \\ & && \text{constraints (VI) - (IX) in model (14)} \\ & \text{vars.} && n_{ij}^d, m_{ij}^d, A_{ij}^d, B_{ij}^d, C_{ij}^d, D_d, E_d \geq 0, \\ & && 0 \leq \alpha_{ij} \leq \varphi_k, 0 \leq e_{ij} \leq 1 \end{aligned} \quad (35)$$

The augmented Lagrangian function for the model (35) can be presented as follows:

$$\begin{aligned} & L_\beta(n, m, A, B, C, D, E, \alpha, e; \\ & \delta, \bar{\delta}, \eta, \bar{\eta}, \zeta, \bar{\zeta}, \lambda_0, \bar{\lambda}_0, \lambda_1, \bar{\lambda}_1, \mu, \xi, \sigma, \rho) = Eq.(12) \\ & + \delta^\top Eq.(25) + \frac{\beta}{2} \|Eq.(25)\|_2^2 + \bar{\delta}^\top Eq.(26) + \frac{\beta}{2} \|Eq.(26)\|_2^2 \\ & + \eta^\top Eq.(27) + \frac{\beta}{2} \|Eq.(27)\|_2^2 + \bar{\eta}^\top Eq.(28) + \frac{\beta}{2} \|Eq.(28)\|_2^2 \\ & + \zeta^\top Eq.(29) + \frac{\beta}{2} \|Eq.(29)\|_2^2 + \bar{\zeta}^\top Eq.(30) + \frac{\beta}{2} \|Eq.(30)\|_2^2 \\ & + \lambda_0^\top Eq.(31) + \frac{\beta}{2} \|Eq.(31)\|_2^2 + \bar{\lambda}_0^\top Eq.(32) + \frac{\beta}{2} \|Eq.(32)\|_2^2 \\ & + \lambda_1^\top Eq.(33) + \frac{\beta}{2} \|Eq.(33)\|_2^2 + \bar{\lambda}_1^\top Eq.(34) + \frac{\beta}{2} \|Eq.(34)\|_2^2 \\ & + \mu^\top \text{Cons. (VI, model (14))} + \frac{\beta}{2} \|\text{Cons. (VI model (14))}\|_2^2 \\ & + \xi^\top \text{Cons. (VII, model (14))} + \frac{\beta}{2} \|\text{Cons. (VII model (14))}\|_2^2 \\ & + \sigma^\top \text{Cons. (VIII, model (14))} + \frac{\beta}{2} \|\text{Cons. (VIII model (14))}\|_2^2 \\ & + \rho^\top \text{Cons. (IX, model (14))} + \frac{\beta}{2} \|\text{Cons. (IX model (14))}\|_2^2 \end{aligned} \quad (36)$$

where $\delta, \bar{\delta}, \eta, \bar{\eta}, \zeta, \bar{\zeta}, \lambda_0, \bar{\lambda}_0, \lambda_1, \bar{\lambda}_1, \mu, \xi, \sigma$, and ρ are the Lagrangian multiplier and β is the positive penalty scalar.

Considering the above explanations, we are able to utilize the n -block RP-ADMM algorithm to distributedly solve the model (35) (see Algorithm 1). As illustrated in Algorithm 1, it starts by initializing primal and dual variables in the first line. In line 2, we defined an array to store the last values of the variables (primal and dual). The iterative process begins in line 3 where ℓ indicates the round number. In each round, RP-ADMM algorithm picks a uniformly random permutation Ω to specify the order of updating primal variables. In fact, the inner *for* loop (lines 6 to 10), selects the $\Omega(i)^{th}$ variable to update its value in ℓ^{th} round by considering the last value of variables stored in the *LastValueArray*. Then, in line 9 the value of $\Omega(i)^{th}$ primal variable is overwritten. For illustration, suppose that $\Omega = \{3, 4, 2, 1\}$ ($n = 4$) and $\ell = 12$. Therefore, according to the order of Ω , the third primal variable (i.e. $\Omega(1)^{th}$ variable) must be updated first, followed by fourth, second, and first primal variable. Also, in this round, for example, the second variable updating process uses the current values (computed in 12^{th} round) of third and fourth variables, and also the obtained value of the first variable in previous round (i.e. 11^{th} round). Using the scaled form of ADMM algorithm, the primal variables are updated as follows:

Updating Switches Primal Variables: If the selected variable belongs to the set of network switches, each switch i , updates its traffic flow rate variable to its neighbor $j \in \{S \cup \Psi \cup \Psi^*\}$ using the following equations:

$$\begin{aligned} n_{ij}^{d(\ell)} &= \underset{n_{ij}^d}{\operatorname{argmin}}(Eq.(36)) \\ &= \underset{n_{ij}^d}{\operatorname{argmin}}(Eq.(12) + \frac{\beta}{2} \| n_{ij}^d - n_{ji}^{d(\ell^*)} - A_{ij}^{d(\ell^*)} \\ &+ \frac{1}{\beta} \delta_{ij}^{d(\ell^*)} \|_2^2 + \frac{\beta}{2} \| n_{ji}^{d(\ell^*)} - n_{ij}^d - A_{ji}^{d(\ell^*)} + \frac{1}{\beta} \delta_{ji}^{d(\ell^*)} \|_2^2 \\ &+ \frac{\beta}{2} \| m_{ji}^{d(\ell^*)} - \gamma_k n_{ij}^d - C_{ji}^{d(\ell^*)} + \frac{1}{\beta} \zeta_{ji}^{d(\ell^*)} \|_2^2 + \frac{\beta}{2} \| n_{ij}^d \\ &+ \sum_{s \in \{S \cup \Psi \cup \Psi^*\}, s \neq i} l_{sj} n_{sj}^{d(\ell^*)} - T e_{jd}^* + \frac{1}{\beta} \sigma_j^{d(\ell^*)} \|_2^2 \\ &+ \Theta \frac{\beta}{2} \| n_{ij}^d + \sum_{s \in \{S \cup \Psi \cup \Psi^*\}, s \neq i} l_{sj} n_{sj}^{d(\ell^*)} - T \alpha_{jd}^{(\ell^*)} \\ &+ \frac{1}{\beta} \rho_j^{d(\ell^*)} \|_2^2) \end{aligned} \quad (37)$$

where Θ plays as an if statement; i.e. it returns 1 if condition "X" is true. Also, $a^{d(\ell^*)}$ or $a^{(\ell^*)}$ shows the most recent value of the primal variable a , stored in *LastValueArray*. Similarly, $m_{ij}^{d(\ell)}$, $A_{ij}^{d(\ell)}$, and $B_{ij}^{d(\ell)}$ are obtained from the following equations:

$$\begin{aligned} m_{ij}^{d(\ell)} &= \underset{m_{ij}^d}{\operatorname{argmin}}(Eq.(12) + \frac{\beta}{2} \| m_{ij}^d - m_{ji}^{d(\ell^*)} - B_{ij}^{d(\ell^*)} \\ &+ \frac{1}{\beta} \eta_{ij}^{d(\ell^*)} \|_2^2 + \frac{\beta}{2} \| m_{ji}^{d(\ell^*)} - m_{ij}^d - B_{ij}^{d(\ell^*)} + \frac{1}{\beta} \eta_{ji}^{d(\ell^*)} \|_2^2 \\ &+ \Theta \frac{\beta}{2} \| m_{ij}^d + \sum_{s \in \{S \cup \mathcal{P}_1\}, s \neq i} l_{sj} m_{sj}^{d(\ell^*)} - E_d^{(\ell^*)} \\ &+ \frac{1}{\beta} \lambda_j^{d(\ell^*)} \|_2^2) \end{aligned} \quad (38)$$

$$\begin{aligned} A_{ij}^{d(\ell)} &= \underset{A_{ij}^d}{\operatorname{argmin}}(\frac{\beta}{2} \| n_{ij}^{d(\ell^*)} - n_{ji}^{d(\ell^*)} - A_{ij}^d + \frac{1}{\beta} \delta_{ij}^{d(\ell^*)} \|_2^2 \\ &+ \frac{\beta}{2} \| A_{ij}^d + \sum_{s \in \{S \cup \Psi \cup \Psi^*\}, s \neq j} A_{is}^{d(\ell^*)} + \frac{1}{\beta} \bar{\delta}_i^{d(\ell^*)} \|_2^2) \end{aligned} \quad (39)$$

Algorithm 1: n -block RP-ADMM

```

1 Initialize:  $n_{ij}^{d(0)}, m_{ij}^{d(0)}, \alpha_{ij}^{d(0)}, e_{ij}^{d(0)}, A_{ij}^{d(0)}, B_{ij}^{d(0)},$ 
 $C_{ij}^{d(0)}, D_d^{(0)}, E_d^{(0)}$  and dual variables
2 LastValueArray[] = initialized variables;
3 for  $\ell = 1, 2, \dots$  do
4    $n = \text{length}(\text{primal variables})$ ;
5   Pick a permutation  $\Omega$  of  $\{1, \dots, n\}$  uniformly at
   random;
6   for  $i = 1 : n$  do
7     //Updating primal variables:
8     Update the  $\Omega(i)^{th}$  primal variable;
9     Update LastValueArray[ $\Omega(i)$ ];
10  end
11  Update dual variables;
12 end

```

$$\begin{aligned} B_{ij}^{d(\ell)} &= \underset{B_{ij}^d}{\operatorname{argmin}}(\frac{\beta}{2} \| m_{ij}^{d(\ell^*)} - m_{ji}^{d(\ell^*)} - B_{ij}^d + \frac{1}{\beta} \eta_{ij}^{d(\ell^*)} \|_2^2 \\ &+ \frac{\beta}{2} \| B_{ij}^d + \sum_{s \in \{S \cup \Psi \cup \Psi^*\}, s \neq j} B_{is}^{d(\ell^*)} + \frac{1}{\beta} \bar{\eta}_i^{d(\ell^*)} \|_2^2) \end{aligned} \quad (40)$$

Updating PMs Primal Variables: If the selected variable belongs to the set of PMs, each PM i updates the primal dual variables n_{ij}^d , m_{ij}^d , α_{id} , e_{id} , $C_{ij}^{d(\ell)}$, $D_d^{(\ell)}$, and $E_d^{(\ell)}$ using the following equations:

$$\begin{aligned} n_{ij}^{d(\ell)} &= \underset{n_{ij}^d}{\operatorname{argmin}}(Eq.(12) + \frac{\beta}{2} \| n_{ij}^{d(\ell^*)} - n_{ij}^d - A_{ji}^{d(\ell^*)} \\ &+ \frac{1}{\beta} \delta_{ji}^{d(\ell^*)} \|_2^2 + \Theta \frac{\beta}{2} \| n_{ij}^d + \sum_{s \in \{S \cup \mathcal{P}_0\}, s \neq j} l_{is} n_{is}^{d(\ell^*)} \\ &- D_d^{(\ell^*)} + \frac{1}{\beta} \lambda_0^{d(\ell^*)} \|_2^2 + \frac{\beta}{2} \| n_{ij}^d \\ &+ \sum_{s \in \{S \cup \Psi \cup \Psi^*\}, s \neq i} l_{sj} n_{sj}^{d(\ell^*)} - T e_{jd}^{d(\ell^*)} + \frac{1}{\beta} \sigma_j^{d(\ell^*)} \|_2^2 \\ &+ \frac{\beta}{2} \| n_{ij}^d + \sum_{s \in \{S \cup \Psi \cup \Psi^*\}, s \neq i} l_{sj} n_{sj}^{d(\ell^*)} - T \alpha_{jd}^{d(\ell^*)} \\ &+ \frac{1}{\beta} \rho_j^{d(\ell^*)} \|_2^2) \end{aligned} \quad (41)$$

$$\begin{aligned} m_{ij}^{d(\ell)} &= \underset{m_{ij}^d}{\operatorname{argmin}}(Eq.(12) + \frac{\beta}{2} \| m_{ji}^{d(\ell^*)} - m_{ij}^d - B_{ji}^{d(\ell^*)} \\ &+ \frac{1}{\beta} \eta_{ji}^{d(\ell^*)} \|_2^2 + \frac{\beta}{2} \| m_{ij}^d - \gamma_k n_{ij}^{d(\ell^*)} - C_{ij}^{d(\ell^*)} + \frac{1}{\beta} \zeta_{ij}^{d(\ell^*)} \|_2^2 \\ &+ \Theta \frac{\beta}{2} \| m_{ij}^d + \sum_{s \in \{S \cup \mathcal{P}_1\}, s \neq i} l_{sj} m_{sj}^{d(\ell^*)} - E_d^{(\ell^*)} \\ &+ \frac{1}{\beta} \lambda_i^{d(\ell^*)} \|_2^2) \end{aligned} \quad (42)$$

$$\begin{aligned} \alpha_{id}^{(\ell)} &= \underset{\alpha_{id}}{\operatorname{argmin}}(\Theta \frac{\beta}{2} \sum_{q \in \{\Psi \cup \Psi^*\}, q \neq i} \| \alpha_{id} - \alpha_{qd}^{(\ell^*)} \\ &+ \frac{1}{\beta} \mu_{iq}^{d(\ell^*)} \|_2^2 + \frac{\beta}{2} \| \sum_{q \in \{\Psi \cup \Psi^*\}} e_{qd}^{(\ell^*)} - \alpha_{id} + \frac{1}{\beta} \xi_i^{d(\ell^*)} \|_2^2 \\ &+ \Theta \frac{\beta}{2} \| \sum_{s \in \{S \cup \Psi \cup \Psi^*\}} l_{si} n_{si}^{d(\ell^*)} - T \alpha_{id} + \frac{1}{\beta} \rho_i^{d(\ell^*)} \|_2^2) \end{aligned} \quad (43)$$

$$\begin{aligned}
e_{id}^{(\ell)} &= \underset{e_{id}}{\operatorname{argmin}} \left(\frac{\beta}{2} \| e_{id} + \sum_{q \in \{\Psi \cup \Psi^*\}, q \neq i} e_{id}^{(\ell^*)} - \alpha_{qd}^{(\ell^*)} \right. \\
&+ \frac{1}{\beta} \xi_i^{d(\ell^*)} \|_2^2 + \Theta_{d>|V} \frac{\beta}{2} \| \sum_{s \in \{\mathcal{S} \cup \Psi \cup \Psi^*\}} l_{si} n_{si}^{d(\ell^*)} - T e_{id} \\
&\left. + \frac{1}{\beta} \sigma_i^{d(\ell^*)} \|_2^2 \right) \quad (44)
\end{aligned}$$

$$\begin{aligned}
C_{ij}^{d(\ell)} &= \underset{C_{ij}^d}{\operatorname{argmin}} \left(\frac{\beta}{2} \| m_{ij}^{d(\ell^*)} - \gamma_k n_{ji}^{d(\ell^*)} - C_{ij}^d + \frac{1}{\beta} \zeta_{ij}^{d(\ell^*)} \|_2^2 \right. \\
&\left. + \frac{\beta}{2} \| C_{ij}^d + \sum_{s \in \{\mathcal{S} \cup \Psi \cup \Psi^*\}, s \neq j} C_{is}^{d(\ell^*)} + \frac{1}{\beta} \bar{\zeta}_i^{d(\ell^*)} \|_2^2 \right) \quad (45)
\end{aligned}$$

In case of $i \in \text{ingress}$, we should update $D_d^{(\ell)}$ as below:

$$\begin{aligned}
D_d^{(\ell)} &= \underset{D_d}{\operatorname{argmin}} \left(\frac{\beta}{2} \left(\| \sum_{s \in \{\mathcal{S} \cup \mathcal{P}_0\}} l_{is} n_{is}^{d(\ell^*)} - D_d \right. \right. \\
&+ \frac{1}{\beta} \lambda_i^{d(\ell^*)} \|_2^2 + \| D_d + \sum_{s=1:v^*, s \neq d} D_s^{(\ell^*)} - T \\
&\left. \left. + \frac{1}{\beta} \bar{\lambda}_i^{(\ell^*)} \|_2^2 \right) \right) \quad (46)
\end{aligned}$$

However, if $i \in \text{egress}$, $E_d^{(\ell)}$ should be updated as follows:

$$\begin{aligned}
E_d^{(\ell)} &= \underset{E_d}{\operatorname{argmin}} \left(\frac{\beta}{2} \left(\| \sum_{s \in \{\mathcal{S} \cup \mathcal{P}_1\}} l_{si} m_{si}^{d(\ell^*)} - E_d \right. \right. \\
&+ \frac{1}{\beta} \lambda_i^{d(\ell^*)} \|_2^2 + \| E_d + \sum_{s=1:v^*, s \neq d} E_s^{(\ell^*)} - T \\
&\left. \left. + \frac{1}{\beta} \bar{\lambda}_i^{(\ell^*)} \|_2^2 \right) \right) \quad (47)
\end{aligned}$$

Finally, in line 11 of Algorithm 1, the dual variables can be updated as follows:

- $\forall i \in \mathcal{S}$ and $\forall j \in \{\mathcal{S} \cup \Psi \cup \Psi^*\}$:

$$\delta_{ij}^{d(\ell)} = \delta_{ij}^{d(\ell-1)} + \beta(n_{ij}^{d(\ell)} - n_{ji}^{d(\ell)} - A_{ij}^{d(\ell)}) \quad (48)$$

$$\bar{\delta}_i^{d(\ell)} = \bar{\delta}_i^{d(\ell-1)} + \beta \left(\sum_{s \in \{\mathcal{S} \cup \mathcal{P}\}} A_{is}^{d(\ell)} \right) \quad (49)$$

$$\eta_{ij}^{d(\ell)} = \eta_{ij}^{d(\ell-1)} + \beta(m_{ij}^{d(\ell)} - m_{ji}^{d(\ell)} - B_{ij}^{d(\ell)}) \quad (50)$$

$$\bar{\eta}_i^{d(\ell)} = \bar{\eta}_i^{d(\ell-1)} + \beta \left(\sum_{s \in \{\mathcal{S} \cup \mathcal{P}\}} B_{is}^{d(\ell)} \right) \quad (51)$$

- $\forall i \in \{\Psi \cup \Psi^*\}$ and $\forall j \in \{\mathcal{S} \cup \Psi \cup \Psi^*\}$:

$$\zeta_{ij}^{d(\ell)} = \zeta_{ij}^{d(\ell-1)} + \beta(m_{ij}^{d(\ell)} - \gamma_k n_{ji}^{d(\ell)} - C_{ij}^{d(\ell)}) \quad (52)$$

$$\bar{\zeta}_i^{d(\ell)} = \bar{\zeta}_i^{d(\ell-1)} + \beta \left(\sum_{s \in \{\mathcal{S} \cup \mathcal{P}\}} C_{il}^{d(\ell)} \right) \quad (53)$$

$$\sigma_i^{d(\ell)} = \sigma_i^{d(\ell-1)} + \beta \left(\sum_{s \in \{\mathcal{S} \cup \mathcal{P}\}} l_{si} n_{si}^{d(\ell)} - T e_{id}^{(\ell)} \right) \quad (54)$$

$$\mu_{ij}^{d(\ell)} = \mu_{ij}^{d(\ell-1)} + \beta(\alpha_{id}^{(\ell)} - \alpha_{jd}^{(\ell)}) \quad (55)$$

$$\xi_i^{d(\ell)} = \xi_i^{d(\ell-1)} + \beta \left(\sum_{l \in \{\Psi \cup \Psi^*\}} e_{id}^{d(\ell)} - \alpha_{id}^{(\ell)} \right) \quad (56)$$

- If ($i \in \text{ingress}$) :

$$\lambda_0^{d(\ell)} = \lambda_0^{d(\ell-1)} + \beta \left(\sum_{s \in \{\mathcal{S} \cup \mathcal{P}_0\}} l_{0s} n_{0s}^{d(\ell)} - D_d^{(\ell)} \right) \quad (57)$$

$$\bar{\lambda}_0^{(\ell)} = \bar{\lambda}_0^{(\ell-1)} + \beta \left(\sum_{d=1:v^*} D_d^{(\ell)} \right) \quad (58)$$

- If ($i \in \text{egress}$) :

$$\lambda_1^{d(\ell)} = \lambda_1^{d(\ell-1)} + \beta \left(\sum_{s \in \{\mathcal{S} \cup \mathcal{P}_1\}} l_{s1} m_{s1}^{d(\ell)} - E_d^{(\ell)} \right) \quad (59)$$

$$\bar{\lambda}_1^{(\ell)} = \bar{\lambda}_1^{(\ell-1)} + \beta \left(\sum_{d=1:v^*} E_d^{(\ell)} \right) \quad (60)$$

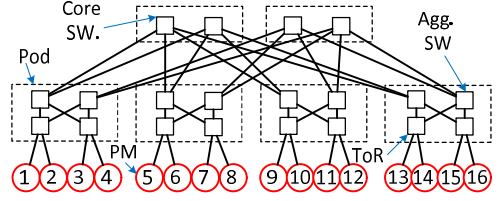


Fig. 4: 4-fat-tree topology

- If ($i \in \Psi$) :

$$\rho_i^{d(\ell)} = \rho_i^{d(\ell-1)} + \beta \left(\sum_{s \in \{\mathcal{S} \cup \mathcal{P}\}} l_{si} n_{si}^{d(\ell)} - T \alpha_{id}^{(\ell)} \right) \quad (61)$$

Clearly, the primal values of the RP-ADMM model can be updated according to the selected permutation. In case of overload, when $v_k^{i,c}$ violates the defined threshold, it determines v^* and then triggers PMs $\in \{\Psi \cup \Psi^*\}$ and some adjacent switches via sending a simple packet. Then, the RP-ADMM update process can be launched considering the selected PMs and switches.

VI. PERFORMANCE EVALUATION

To assess the performance of the proposed approach, we simulate the proposed MILP, LP, and ADMM models utilizing MATLAB software. We conducted our simulation on a hardware composing of an INTEL Core i7 1.73GHz CPU, 8GB of RAM, and Windows 10 x64 OS. We considered k -fat-tree topologies as the datacenter network structure. In general, a k -fat-tree has $k/2$ ToR and aggregate switches in each pod, and $k^2/4$ core switches [24]. We mainly use 4-ary fat-tree topology for performance evaluation (see Fig. 4). However, we run some of our experiments based on different topology sizes (i.e. 8, 16, 32, and 64-ary fat-trees). Also τ is considered as 5 seconds for all runs.

In our first experiment, we investigate the performances of MILP (13), LP (14), and (15) models in terms of the number of variables, constraints, and execution time on various k -fat-trees. We consider g_k^c with two VMs (i.e. $|V| = 2$), $|\Psi| = 2$, $|\Psi^*| = 1$, and each rack equipped with two PMs. As it can be seen in Table II, number of switches, PMs, variables, and constraints are proportional to the size of datacenter (i.e. k -fat-tree). Obviously, for larger datacenters, the execution time of the proposed models also increased specifically for MILP model. Table II clearly implies that in LDCs, we need to apply an efficient distributed algorithm to achieve a solution in a reasonable time.

As another experiment, we consider a 4-fat-tree topology to evaluate the performance of the proposed LP models for over and underloaded chains in different scenarios. Each of these scenarios is designed based on the number of PMs in g_k^c , ingress and egress points, Ψ^* (i.e. the candidate PMs that are able to launch new VMs), normalized (0 ~ 1) forwarding costs (N. FwdCost), and v^* (the number of required VNFs to process the received traffic by g_k^c) for overload and underload cases (see column 1-4 in Table III). We also considered the forwarding costs between different layers of datacenter as: PM-ToR=10, ToR-Aggregation=20, and Aggregation-Core=40. Further, to simulate overload and underload cases, we increased the traffic rate by 50% (for each extra VM) and decreased it by 50%, respectively.

TABLE II: Comparing time complexities of proposed models in different topologies

k -fat-tree	No. of Switches	No. of PMs	No. of Variables	MILP (13)		LP (Overload) (14)		LP (Underload) (15)	
				No. of Constraints	Run time (s)	No. of Constraints	Run time (s)	No. of Constraints	Run Time (s)
2	5	4	120	25	0.555	68	0.075	62	0.055
4	20	16	672	315	5.064	254	0.614	188	0.318
8	80	64	4224	1251	323.523	998	14.977	692	6.635
16	320	256	29184	4995	N/A	3974	974.235	2708	456.674
32	1280	1024	129024	19971	N/A	15878	N/A	10772	N/A
64	5120	4096	466944	79875	N/A	63494	N/A	43028	N/A

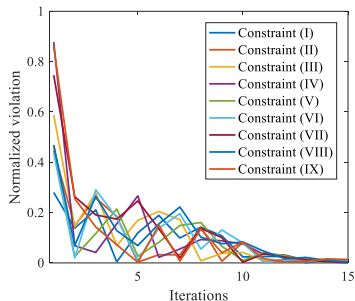


Fig. 5: The normalized violation of the constraints in model (14)

The normalized forwarding cost and its increase/decrease rates are shown in Table III. We also configured the initial parameters in such a way that LP is forced to turn on/off VMs to mitigate the over or underload g_k^c . Overall, Table III shows how LP models react to over/underloaded chains by turning on/off offline/online VMs. In case of overload, the proposed LP model (14) prefers to launch a VM on ingress and egress PMs in order to minimize the traffic forwarding cost. In fact, the traffic between two different VNFs on an identical PM (i.e. ingress or egress) is negligible due to an internal traffic handling. For instance, in the first scenario in Table III ($\mathcal{P}_1 \rightarrow (\mathcal{P}_2, \mathcal{P}_5) \rightarrow \mathcal{P}_4$), when we need to turn on two more VMs for g_k^c (i.e. $v^* = 4$), the LP model (14) selects \mathcal{P}_1 and \mathcal{P}_4 (i.e. ingress and egress points, respectively) to launch the new VMs ($\mathcal{P}_1 \rightarrow (\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_5, \mathcal{P}_4) \rightarrow \mathcal{P}_4$). Also, it can be seen that by using this new configuration, the forwarding cost increases by 20%, which is caused by the increased amount of network traffic. In addition, it can be seen that in the third scenario ($\mathcal{P}_1 \rightarrow (\mathcal{P}_1, \mathcal{P}_2) \rightarrow \mathcal{P}_2$), the maximum increase in forwarding cost is resulted among all scenarios (116%). This is because, the utilized PMs before chain overload (i.e. \mathcal{P}_1 and \mathcal{P}_2) were on a same rack, while after running LP model (14), two more PMs (i.e. \mathcal{P}_3 and \mathcal{P}_4) are used to handle the traffic, which are stored on another rack. It is noticeable that we assumed that only $\mathcal{P}_3 - \mathcal{P}_{16}$ PMs can afford enough resources to launch new VMs (i.e. $\Psi^* = \{\mathcal{P}_3 - \mathcal{P}_{16}\}$). On the other hand, when an underload occurs, the LP model (15) prefers to turn off a VM that has the most forwarding cost. For example, in the first scenario, \mathcal{P}_5 is turned off. Because \mathcal{P}_5 is in pod 2, so the traffic should pass the higher layers of network (i.e. aggregate and core layers) to reach the next hop in the chain (i.e. \mathcal{P}_4). Also, as it can be seen in the last column of Table III, the forwarding cost decreases by 14%. Additionally, it can be seen that maximum forwarding cost reduction took place in the last scenario. The reason is that the LP model (15) turns off the VMs on \mathcal{P}_3 and \mathcal{P}_4 , which are stored in another rack. Therefore, the traffic is not passing the higher network layers and as a result, the forwarding cost in the chain reduces significantly.

One of the well-known features of ADMM is its fast convergence which is clearly illustrated in Fig. 6. For declaration,

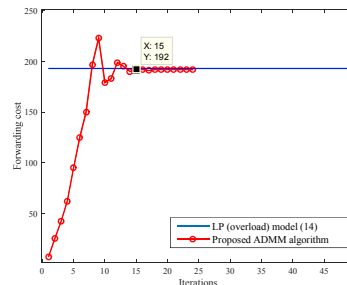


Fig. 6: The convergence rate of the proposed distributed RP-ADMM algorithm

we focused on the last scenario in Table III and compared the optimal cost value obtained by LP model (14) with the solution of distributed ADMM algorithm (we set $\beta = 5$). According to Fig. 6, the distributed ADMM converges to the optimal value in 15 iterations. This low number of iterations indicates that the proposed distributed ADMM technique can perform well with low overhead (e.g. message passing) in a LDC. Notably, we set the stop criterion to 25 iterations. In addition, in Fig. 5 we measure the normalized violation of each constraint in model (14) to show how fast violations decrease.

VII. CONCLUSION

Today, the integration of NFV and SDN introduces a new distinguished, yet challenging networking architecture. One of these challenges is VNF over and underloads in the network during their operational time. This problem could increase resource wastage, SLA violations, and also decrease efficiency. In this paper, we conducted a theoretical study of the VNF scaling problem with the aim of minimizing deployment and forwarding costs in a datacenter. In fact, we addressed the aforementioned issue by presenting a jointly traffic load balancing and VNF scaling mechanism to overcome both over and underload phenomena in datacenters. In this regard, we formulated the problem as a MILP optimization model. Also, to cope with the time complexity of MILP, we relaxed it into two LP models. However, in LDCs, there are huge number of network switches and servers, that increase the number of constraints and variables in LP models significantly. Hence, the proposed central LP model is practically useless for LDCs. Therefore, we extend our approach to a distributed method based on the ADMM technique. However, the conventional ADMM is guaranteed to converge for two-block variables problem, while for multi-block variable problems, updating different variables sequentially has no convergence guarantee. Hence, because our problem is formulated in a multi-block variable form, we used an extension of the ADMM technique, called Randomly Permuted ADMM (RP-ADMM) to solve the problem in a distributed manner. To the best of our knowledge, our work is the first one to tackle this problem

TABLE III: Comparing the results of LP models in different scenarios

$g_{k-1}^c \rightarrow (g_k^c) \rightarrow g_{k+1}^c$	N. FwdCost	LP overload model (14)				LP underload model (15)		
		Ψ^*	v^*	New Configuration	FwdCost	v^*	New Configuration	FwdCost
$\mathcal{P}_1 \rightarrow (\mathcal{P}_2, \mathcal{P}_5) \rightarrow \mathcal{P}_4$	0.68	ALL PMs	3	$\mathcal{P}_1 \rightarrow (\mathcal{P}_2, \mathcal{P}_5, \mathbf{\mathcal{P}}_4) \rightarrow \mathcal{P}_4$	+14%	1	$\mathcal{P}_1 \rightarrow (\mathcal{P}_2) \rightarrow \mathcal{P}_4$	-14%
			4	$\mathcal{P}_1 \rightarrow (\mathbf{\mathcal{P}}_1, \mathcal{P}_2, \mathcal{P}_5, \mathbf{\mathcal{P}}_4) \rightarrow \mathcal{P}_4$	+20%			
$\mathcal{P}_1 \rightarrow (\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3) \rightarrow \mathcal{P}_{16}$	1	$\mathcal{P}_3 - \mathcal{P}_{16}$	4	$\mathcal{P}_1 \rightarrow (\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \mathbf{\mathcal{P}}_{16}) \rightarrow \mathcal{P}_{16}$	+11%	2	$\mathcal{P}_1 \rightarrow (\mathcal{P}_1, \mathcal{P}_2) \rightarrow \mathcal{P}_4$	-6%
			5	$\mathcal{P}_1 \rightarrow (\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \mathbf{\mathcal{P}}_3, \mathbf{\mathcal{P}}_{16}) \rightarrow \mathcal{P}_{16}$	+17%			
$\mathcal{P}_1 \rightarrow (\mathcal{P}_1, \mathcal{P}_2) \rightarrow \mathcal{P}_2$	0.23	$\mathcal{P}_3 - \mathcal{P}_{16}$	3	$\mathcal{P}_1 \rightarrow (\mathcal{P}_1, \mathbf{\mathcal{P}}_1, \mathcal{P}_2) \rightarrow \mathcal{P}_2$	+25%	1	$\mathcal{P}_1 \rightarrow (\mathcal{P}_1) \rightarrow \mathcal{P}_2$	-33%
			4	$\mathcal{P}_1 \rightarrow (\mathcal{P}_1, \mathcal{P}_1, \mathbf{\mathcal{P}}_3, \mathbf{\mathcal{P}}_4) \rightarrow \mathcal{P}_2$	+116%			
$\mathcal{P}_1 \rightarrow (\mathcal{P}_3, \mathcal{P}_5, \mathcal{P}_6) \rightarrow \mathcal{P}_2$	0.77	$\mathcal{P}_3 - \mathcal{P}_{16}$	4	$\mathcal{P}_1 \rightarrow (\mathbf{\mathcal{P}}_2, \mathcal{P}_3, \mathcal{P}_5, \mathcal{P}_6) \rightarrow \mathcal{P}_2$	+13%	2	$\mathcal{P}_1 \rightarrow (\mathcal{P}_3, \mathcal{P}_5) \rightarrow \mathcal{P}_2$	-9%
			5	$\mathcal{P}_1 \rightarrow (\mathcal{P}_3, \mathbf{\mathcal{P}}_3, \mathbf{\mathcal{P}}_4, \mathcal{P}_5, \mathcal{P}_6) \rightarrow \mathcal{P}_2$	+30%			

using a distributed optimization approach. We evaluated the performance of the presented models numerically in different scenarios based on k-fat-tree topology. The results validated the performance of our proposed algorithms. The most important part of our future work is to implement the introduced distributed approach in real-world test beds. On the other hand, considering a workload prediction module in our architecture is worth to be explored in future. In this way, we would be able to predict the traffic load and adjust VNF placement in advance. Additionally, the performance (e.g., end-to-end delay) is also a critical factor in forwarding path selection of service chains. In fact, it would influence the selection of candidate PMs and VMs. Therefore, it could be considered as an extension to the mathematical formulation as another future direction.

VIII. ACKNOWLEDGEMENT

We are grateful to Islamic Azad University, Mashhad branch authorities, for their useful collaboration. This work has been partly supported by the German Federal Ministry of Education and Research (BMBF) under the project Secure Networking for a DATA center cloud in Europe (SENDATE) (Project ID 16KIS0261).

REFERENCES

- [1] Rashid Mijumbi, Joan Serrat, Juan-Luis Gorricho, Niels Bouten, Filip De Turck, and Raouf Boutaba. Network function virtualization: State-of-the-art and research challenges. *IEEE Communications Surveys & Tutorials*, 18(1):236–262, 2015.
- [2] Dilip A Joseph, Arsalan Tavakoli, and Ion Stoica. A policy-aware switching layer for data centers. In *ACM SIGCOMM Computer Communication Review*, volume 38, pages 51–62. ACM, 2008.
- [3] Arsany Basta, Wolfgang Kellerer, Marco Hoffmann, Hans Jochen Morper, and Klaus Hoffmann. Applying nfV and sdn to lte mobile core gateways, the functions placement problem. In *Proceedings of the 4th workshop on All things cellular: operations, applications, & challenges*, pages 33–38. ACM, 2014.
- [4] Dilip Krishnaswamy, Ravi Kothari, and Vijay Gabale. Latency and policy aware hierarchical partitioning for nfV systems. In *Network Function Virtualization and Software Defined Network (NFV-SDN), 2015 IEEE Conference on*, pages 205–211. IEEE, 2015.
- [5] Minh-Tuan Thai, Ying-Dar Lin, and Yuan-Cheng Lai. A joint network and server load balancing algorithm for chaining virtualized network functions. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2016.
- [6] Md Bari, Shihabur Rahman Chowdhury, Reaz Ahmed, Raouf Boutaba, et al. On orchestrating virtual network functions in nfV. *IEEE Transactions on Network and Service Management*, 2015.
- [7] Rami Cohen, Liane Lewin-Eytan, Joseph Seffi Naor, and Danny Raz. Near optimal placement of virtual network functions. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pages 1346–1354. IEEE, 2015.
- [8] Milad Ghaznavi, Aimal Khan, Nashid Shahriar, Khalid Alsubhi, Reaz Ahmed, and Raouf Boutaba. Elastic virtual network function placement. In *Cloud Networking (CloudNet), 2015 IEEE 4th International Conference on*, pages 255–260. IEEE, 2015.

- [9] Stephen A Vavasis. *Nonlinear optimization: complexity issues*. Oxford University Press, Inc., 1991.
- [10] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.
- [11] Caihua Chen, Bingsheng He, Yinyu Ye, and Xiaoming Yuan. The direct extension of admm for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming*, 155(1-2):57–79, 2016.
- [12] Ruoyu Sun, Zhi-Quan Luo, and Yinyu Ye. On the expected convergence of randomly permuted admm. *arXiv preprint arXiv:1503.06387*, 2015.
- [13] Joe Wenjie Jiang, Tian Lan, Sangtae Ha, Minghua Chen, and Mung Chiang. Joint vm placement and routing for data center traffic engineering. In *INFOCOM, 2012 Proceedings IEEE*, pages 2876–2880. IEEE, 2012.
- [14] Xiaoqiao Meng, Vasileios Pappas, and Li Zhang. Improving the scalability of data center networks with traffic-aware virtual machine placement. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9. IEEE, 2010.
- [15] Amir Varasteh and Maziar Goudarzi. Server consolidation techniques in virtualized data centers: A survey. *IEEE Systems Journal*, 2015.
- [16] Ying Zhang, Neda Beheshti, Ludovic Beliveau, Geoffrey Lefebvre, Ravi Manghirmalani, Ramesh Mishra, Riton Patney, Meral Shirazipour, Ramesh Subrahmaniam, Catherine Truchan, et al. Steering: A software-defined networking for inline service chaining. In *2013 21st IEEE International Conference on Network Protocols (ICNP)*, pages 1–10. IEEE, 2013.
- [17] Marcelo Caggiani Luizelli, Leonardo Richter Bays, Luciana Salete Buriol, Marinho Pilla Barcellos, and Luciano Paschoal Gaspary. Piecing together the nfV provisioning puzzle: Efficient placement and chaining of virtual network functions. In *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, pages 98–106. IEEE, 2015.
- [18] Aaron Gember, Anand Krishnamurthy, Saul St John, Robert Grandl, Xiaoyang Gao, Ashok Anand, Theophilus Benson, Aditya Akella, and Vyas Sekar. Stratos: A network-aware orchestration layer for middleboxes in the cloud. Technical report, Technical Report, 2013.
- [19] Xiaoke Wang, Chuan Wu, Franck Le, Alex Liu, Zongpeng Li, and Francis Lau. Online vnf scaling in datacenters. *arXiv preprint arXiv:1604.01136*, 2016.
- [20] Po-Wen Chi, Yu-Cheng Huang, and Chin-Laung Lei. Efficient nfV deployment in data center networks. In *2015 IEEE International Conference on Communications (ICC)*, pages 5290–5295. IEEE, 2015.
- [21] Zhen Xiao, Weijia Song, and Qi Chen. Dynamic resource allocation using virtual machines for cloud computing environment. *IEEE Transactions on parallel and distributed systems*, 24(6):1107–1117, 2013.
- [22] Wenrui Ma, Carlos Medina, and Deng Pan. Traffic-aware placement of nfV middleboxes. In *2015 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2015.
- [23] Ermin Wei and Asuman Ozdaglar. Distributed alternating direction method of multipliers. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 5445–5450. IEEE, 2012.
- [24] Praveen Tammana, Rachit Agarwal, and Myungjin Lee. Cherrypick: Tracing packet trajectory in software-defined datacenter networks. In *Proceedings of the 1st ACM SIGCOMM Symposium on Software Defined Networking Research*, page 23. ACM, 2015.