

Dynamic Programming of Network Slices in Software-Defined Metro-Core Optical Networks

Sebastian Troia, *Politecnico di Milano*, Alberto Cibari, *Politecnico di Milano*, Rodolfo Alvizu, *Politecnico di Milano* and Guido Maier, *Politecnico di Milano*

Abstract—Nowadays networks are the basis of our communication providing a great number of services. As a consequence, the traffic is increasing and there is a growing demand for new services that require stringent constraints on capacity, latency and jitter to provide an appropriate Quality of Service (QoS) to end users. In order to cope with these requirements, network infrastructure needs to evolve from a static and closed architecture towards a more scalable, dynamic and agile one. Software-Defined Networking and Network Function Virtualization allow to provide different services, each one with its own QoS constraints, independent and secure, thanks to the network slicing concept, the main subject of this work. Network slicing allows to segment the underlying physical network into different logical networks to provide data transport customized to specific services. In this paper, we propose two mathematical models able to dynamically provision network slices on the physical network, complying with their QoS requirements for their instantiation and routing of traffic. The proposed models aim at minimizing a linear combination of probability of blocking traffic requests, energy consumption of physical network devices and interruption of service due to the reconfiguration of the slices. Taking advantage of traffic signatures from a city’s mobile network, the goal is to predict *how* and *when* to reconfigure slices already deployed in the network with the aim to optimize the resource allocation in the underlying physical network.

Index Terms—Software Defined Networking (SDN), Network Function Virtualization (NFV), Network Slicing (NS), 5G, Mixed Integer Linear Programming (MILP), mobile traffic signatures

I. INTRODUCTION

The revolution of the communication system is just about to happen. Software-Defined Networking (SDN) and Network Function Virtualization (NFV) are finally going mainstream. SDN is a well-known network paradigm that decouples network-device hardware from the control decisions. The network intelligence is logically centralized in software-based controllers (the control plane), and network devices become simple packet forwarding devices (the data plane) that can be programmed via open interfaces. On top of the control plane it is possible to run a large set of applications used to improve the network resource allocation and give flexibility to the physical network [1]. Network Function Virtualization (NFV) is the paradigm of transferring network functions from dedicated hardware appliances to software-based applications running on commercial off-the-shelf equipment [2]. The benefits of this approach are manifold: 1) network devices are no longer proprietary but are open to host software from different vendors; 2) scalability and flexibility of networks increase; 3) CapEx and OpEx can be reduced [3].

The softwarization of network functions, namely Virtual Network Functions (VNFs), and network control procedures

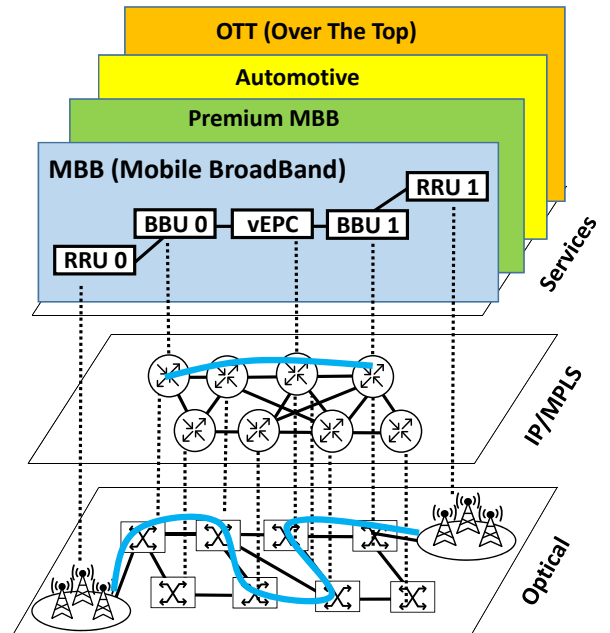


Fig. 1: Metro-core network model. We show the general framework of the IP-over-optical network considered in this work, along with an example of how a Mobile Broadband (MBB) network slice is mapped into the service, IP/MPLS and optical network layer. The MBB handles all the mobile cellular internet traffic in the network and it is composed by 3 kinds of VNF: Remote Radio Unit, Baseband Unit (BBU) and virtual Evolved Packet Core (vEPC).

led to the rise of Network Slicing (NS). NS, based on the concept of network abstraction and virtualization, introduces programmability, flexibility, and modularity to traditional networks [4]. Applying NS, the physical topology is virtually divided into multiple independent virtual networks (VNs), referred to as network slices or, simply, slices. Within a single slice, a set of Service Function Chains (SFCs) and network resources, e.g., bandwidth, are dynamically allocated while meeting specific Quality of Service (QoS) requirements. A SFC is defined as a VNF Forwarding Graph (VNF-FG) which is a sequence of VNFs interconnected to provide a complex network service with specific functionalities. In other words, it consists in steering traffic flows across switches and VNFs in an ordered fashion to fulfill a more sophisticated network service.

According to the 3rd Generation Partnership Project (3GPP)

[5], NS is considered a key driver for 5G networks to serve vertical industries with widely different service needs, in term of latency, reliability and capacity. It does so by exposing isolated partitions of network resources and services realizing multiple independent virtual networks. The modern network/service provider configures network elements to create one or different slices holding a specific set of SFCs. In all cases, each network slice is defined with a variable set Key Performance Indicators (KPIs), such as availability, delay and packet loss.

The allocation of resources to a network slice can be performed either statically or dynamically. In static slicing, a fixed set of resources is allocated to a virtual network for its entire lifetime. This implies that the allocated resources should support the peak service requirements. In dynamic slicing the amount of resources allocated to a virtual network varies according to the actual service needs [6].

In this paper, we propose two mathematical models based on Mixed Integer Linear Programming (MILP) to dynamically optimize the resource allocation of multiple network slices in a metro-core network. Specifically, the transport network being considered carries traffic also from mobile access networks in order to support different kind of 5G applications, such as: mobile broadband, mobile video and cloud data (see section V). In this work, we are interested only on the traffic related to the access networks. In particular, we focus on IP-over-optical based metro-core network following the infrastructure proposed by the European project called Metro-Haul [7]. It comprises metro nodes interconnected by a high capacity dynamic and flexible optical network. The nodes are considered as mini datacenter hosting both Information Technology (IT) and Telecommunication (TLC) equipment, following the 5G Multi-access Edge Computing (MEC) [8] model, so that they can support the instantiation of VNFs.

Fig. 1 shows an example of a metro-core network considered in this work. Given a request to provision a slice, we must consider the availability of resources on three network layers, such as: 1) computing resources in the service layer, e.g. processing and memory; 2) bandwidth in the IP/MPLS layer; 3) and finally optical resources in the transport layer, e.g. fibers and wavelengths. We developed two optimization models: the first is a complete MILP formulation, by which we aim to find the optimal solution to the optimization problem being considered; the second one is called two-step approach. It proposes a solution close to that of the first model but with advantages in terms of execution time. We tested the two models in near-real network scenarios, where we considered: 1) real traffic traces from the mobile network of Milan urban area (Italy); 2) two metro-core network topologies where we show the deployment of numerous types of network slice.

The paper is organized as follow: Section II presents related works. Section III presents the optimization models. Section IV introduces the network scenarios. Section V describes the whole set of network slices implemented in this work. Numerical results are reported in Section VI. Finally, conclusions are presented in Section VII.

II. RELATED WORK

Many attempts have been done on dynamic programming of access, metro and core networks for service chains and network slices resources allocation. The problem of VNF placement and traffic routing for SFC provisioning has been the subject of deep investigation in the last years. [6] proposes an optimization model for the minimization of wavelength-resource usage during a defined period of time. For each experiment they randomly place the VNFs in the network and inject traffic according to a day/night profile. Every time this traffic profile changes (12 hours), a Mixed Integer Linear Programming (MILP) optimization procedure is rerun and the network is reconfigured. In [9] the authors developed algorithms for VNF placement and traffic routing with protection on a single node and link. They performed end-to-end protection and compare the results with an unprotected case. Their model aims at the minimization of the number of active nodes where a VNF is placed, which is an indicator of Capital Expenditure (CapEx) and Operating Expenditure (OpEx) costs.

Authors in [10], address the problem of SFC provisioning by designing a mathematical model based on ILP which scales well with the number of nodes and requests. The goal is to find the best trade-off between the conflicting goals of minimizing infrastructure/bandwidth resources, end-to-end latency and minimizing the replica numbers of VNFs and SFCs (allowing the sharing of VNFs among different SFCs). In [11], authors show how to use NFV coupled with an SDN to improve the energy efficiency of networks. They propose an ILP formulation, and two heuristic models that rely on joint routing and VNF placement. The goal is to minimize the network energy consumption while placing service functions using generic hardware along the paths followed by flows.

In [12] the authors propose a context-free formalization of NF and SFC concepts and present a Mixed Integer Quadratically Constrained Model (MIQCP) for the optimal placement of VNFs and SCs. They deal for the first time with a formal modeling of the service chaining problem, which finds the placement of VNFs and chains them together considering resource limitations of the network. Authors in [13] designed a dynamic programming algorithm to jointly place VNFs and route traffic between them. Also some heuristic algorithms for VNF placement have been already proposed, as in [14]. In [15], the authors proposed a virtual machine allocator for cloud computing data center. They allocate a set of virtual machines on servers that are interconnected through a three-tier fat-tree network topology. The goal is to prevent overloading of computing devices (i.e. allocating more resource than servers' availability) while reducing servers and switches power consumption. They solved the problem using a particular kind of heuristics called Multi Objective Genetic Algorithm (MOGA) is quite often able to effectively solve complex problems.

Dynamic placement of VNFs is addressed in a limited set of works. In [16], the authors provide an algorithm for dynamic VNF placement for SFC provisioning in a metro network. At each time instant, a certain number of users request a specific SFC, and based on the current condition of the

network, VNFs are placed in order to minimize the blocking probability. The algorithm performs VNF placement in such a way that the bandwidth requirements of links, computational requirements of the NFV-nodes and latency requirements of requested SFC are satisfied, and wavelength continuity at each node is enforced. Authors of the work in [17] provide an on-line algorithm for VNF scaling to dynamically provision network services in a datacenter network. In [18] a Mixed Integer Programming formulation and a heuristic algorithm are provided to dynamically provision SFC, again in a datacenter network, where an appropriate resource management is done based on number of users requesting SCs. The authors of [19] consider dynamic SC provisioning for two types of users in the network, new users and existing users that can change location in the network and change their requested SC. So, they propose at first an ILP model for service provisioning with the objective to maximize the profit of the service provider; then, to reduce time complexity, they provide a more scalable model based on column generation [20]. A specific use case of optimal cache deployment for Video-on-Demand delivery as been addressed by authors in [21]. The aim is to find the optimal deployment of caches which minimizes overall network resource occupation. They prove that deploying all the available storage capacity in nearest cache locations does not guarantee the minimal resource occupation.

A. Paper Contribution

This work focuses on the creation and analysis of two mathematical models able to dynamically optimize the resource allocation of network slices. We developed two optimization models: the first is a complete MILP formulation, by which we aim to find the optimal solution to the optimization problem being considered; the second one is a model called two-step approach. It proposes a solution close to that of the first model but with advantages in terms of execution time.

Our contributions in this paper are summarized as follows:

1) *Complete MILP formulation model*: The first model, called complete MILP formulation, provides a complete mathematical formulation taking into consideration the power consumption of the network elements, blocking probability of SFCs traffic requests and reconfiguration penalty due to the re-allocation of SFCs in the IP-over-optical network. Each of these parameters are weighted in order to give freedom of setting a specific policy in the optimization procedure.

2) *Two-step approach*: The previous model is divided in two steps in order to decrease the solving time and complexity. The first step (ILP formulation) solves the VNF placement problem while the second step (MILP formulation) performs Routing and Wavelength Assignment (RWA). Splitting the original overall problem in two steps reduces the number of variables and equations, thus reducing the computation time.

3) *Mobile traffic signatures exploitation*: Reconfigurations are necessary because of the traffic variation in the network slices. In [22] we published a study in which we used real traffic traces from the mobile network of Milan urban area (Italy) to extract typical traffic patterns, called signatures. The signature describes a regularity in the variation of up and

down-stream mobile traffic in a specific area of the city during 24 hours. In this work, we took advantage of these signatures in order to predict how and when to reconfigure slices, already deployed in the network, given their traffic variations over the day.

III. OPTIMIZATION MODELS

The optimization models are multi-objective path formulations focused on the mapping of slices on a metro-core network. Each network slice comprises an independent set of VNFs that support the requirements of a particular SFC.

For each traffic request injected in the SFCs, the model creates multiple lightpaths on the optical layer, from source to the destination node. The model is meant to be executed several times during the day at hourly intervals to optimize resource allocation for a given traffic profile. The topology is a metro-core network composed by a service layer, IP/MPLS layer and optical transport layer, as shown in Fig. 1.

As mentioned in section I, we took as reference the metro-core network architecture proposed by the Metro-Haul European project [7]. The network comprises metro nodes interconnected by a high capacity dynamic and flexible optical network. They are named AMEN (Access Metro Edge Network Nodes) and MCEN (Metro Core Edge Nodes), according to their location in the topology. Transmission distances range between $50km$ and $200km$. To meet the Metro-Haul objectives, both storage and computing resources will be necessary both in the AMEN and the MCEN in order to host VNFs and enable NS. The Metro-Haul network is designed to support several slices according to the 5G-MEC model.

A. Complete MILP formulation

The model comprises a multi-objective function composed by a linear combination of weighted variables. Thanks to these weights we can define different kind of optimization policies according to the user needs. The multi-objective function is composed by:

- Blocking penalty: it is added when an SFC is not routed.
- Power consumption of the IP/MPLS layer.
- Power consumption of the Optical layer.
- Reconfiguration penalty: it is added because of the service disruption generated when the model changes the mapping of an SFC.

The goal of the complete MILP formulation is to find an optimal slice mapping that minimizes the objective function. We now describe the MILP considering the input parameters of the problem, its objective function, variables and constraints.

The optimization model is stated as a multi-objective path formulation focused on the mapping of slices on a metro-core network. Each network slice comprises a set of SFCs. For each traffic request injected in the SFCs, the model creates multiple paths, on the optical layer, from source to the destination node, routing the traffic on multiple fibers. The model is meant to be executed multiple times in a consecutive order to find the configuration of a network for a given traffic pattern.

1) *Optical layer*: The optical network is modelled as a bi-directional graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where \mathcal{V} represent the set of physical nodes $v \in \mathcal{V}$ while \mathcal{E} represent the set of physical links (u, v) . These nodes are composed by white-box ROADMs, following the OPEN ROADM project ¹. The ROADMs can either: drop or terminate a lightpath (send it to the IP layer), add or initiate a lightpath and switch a lightpath. The last operation is also known as optical bypass that allows to reduce power consumption bypassing the usage of the IP layer for routing tasks and avoiding opto-electronic and electro-optic (OEO) conversions. Each physical link has a length and a latency contribution due to the signal propagation time. They also have associated a number of fibers F_{uv} and a maximum number of wavelengths per fiber W . All the wavelengths on a link (u, v) have a bandwidth capacity L_{uv} .

2) *IP/MPLS layer*: The IP/MPLS layer is modeled as a complete bi-directional graph $\hat{\mathcal{G}} = (\hat{\mathcal{V}}, \hat{\mathcal{E}})$ with loops on nodes. The set $\hat{\mathcal{V}}$ represent the set of IP/MPLS nodes $\hat{v} \in \hat{\mathcal{V}}$. $\hat{\mathcal{E}}$ is the set of links (\hat{u}, \hat{v}) between each pair of nodes $\hat{u} \in \hat{\mathcal{V}}$ and $\hat{v} \in \hat{\mathcal{V}}$. Each node \hat{v} has associated a number of transmitters and receivers, denoted as $T(\hat{v})$ and $R(\hat{v})$.

3) *Service layer*: All the nodes are considered to have a mini data center following the 5G-MEC [8]. This architecture allows to instantiate network functions in the metro network. The service layer is thus modelled as a bi-directed graph $\bar{\mathcal{G}} = (\bar{\mathcal{D}}, \bar{\mathcal{E}})$, where $\bar{\mathcal{D}}$ represent the set of NFs $\bar{v} \in \bar{\mathcal{D}}$. Every node $\bar{v} \in \bar{\mathcal{D}}$ can be placed in the network a maximum number of times, denoted as $n_{\bar{v}}$. Also, it can be placed only in specific nodes denoted as $l_{\bar{v}\hat{v}}$ (\hat{v} is a node of the IP/MPLS Layer). Each link $(\bar{u}, \bar{v}) \in \bar{\mathcal{E}}$ represent a possible way to connect two NFs.

4) *Power consumption model*: We refer to the power consumption model in [23] of the different elements of the layers taken in consideration in the formulation.

5) *Input Data*: The input parameters that are used to describe the *static* parts of the problem are listed in this section: the graph topology, the penalties for service disruption and blocking probability, the power consumption model used by nodes and links, but also the *dynamic* parts which describe the traffic input matrix.

a) *Service Layer*: This layer is formed by sets of slices and relative associated chains. It does the mapping of the NF on the IP nodes of the IP/MPLS layer.

- $\bar{\mathcal{D}}$: set of all the possible Network Functions (NFs) d that can be used and placed in the network
- $\bar{\mathcal{G}}(\bar{\mathcal{D}}, \bar{\mathcal{E}})$: a bi-directed graph
- $(\bar{u}, \bar{v}) \in \bar{\mathcal{E}}$: links between two different NFs.
- \mathcal{R} : set of slices $r \in \mathcal{R}$
- b^r : latency requirement of slice r
- $\eta_{\bar{u}\bar{v}}^r$: binary. ($\eta_{\bar{u}\bar{v}}^r = 1$ if there exist some traffic between a NF \bar{u} and another NF \bar{v} , 0 otherwise)
- $h_{\bar{u}\bar{v}}^r$: volume of traffic requested between $(\bar{u}, \bar{v}) \in \bar{\mathcal{E}}$ in the slice $r \in \mathcal{R}$. This input defines how the NFs must be connected, and thus the chains that have to be created. In fact given a slice r , and the links $(\bar{u}, \bar{v}) \in \bar{\mathcal{E}}$ and

$(\bar{w}, \bar{x}) \in \bar{\mathcal{E}} \mid \bar{v} = \bar{w}$, if $h_{\bar{u}\bar{v}}^r \neq 0$ and $h_{\bar{w}\bar{x}}^r \neq 0$ then we have a chain from the NFs $\bar{u} \rightarrow \bar{v} \rightarrow \bar{x}$.

- $m_{\bar{v}}^r$: binary ($m_{\bar{v}}^r = 1$ if $\bar{v} \in \bar{\mathcal{D}}$ is used by slice $r \in \mathcal{R}$, otherwise 0)
- $n_{\bar{v}}$: tells the maximum number of times a NF $\bar{v} \in \bar{\mathcal{D}}$ can be placed in the network.
- $l_{\bar{v}\hat{v}}$ binary ($l_{\bar{v}\hat{v}} = 1$ if $\bar{v} \in \bar{\mathcal{D}}$ can be placed on $\hat{v} \in \hat{\mathcal{V}}$, 0 otherwise).

b) *IP/MPLS layer*: This layer handles the creation of virtual tunnels between the various virtual functions of the mapped chains of the service layer.

- $\hat{\mathcal{G}}(\hat{\mathcal{V}}, \hat{\mathcal{E}})$: a bi-directed graph where $\hat{\mathcal{V}} = \mathcal{V}$ and $\hat{\mathcal{E}} \supseteq \mathcal{E}$, since the graph is complete
- $\hat{v} \in \hat{\mathcal{V}} = 1, 2, \dots, \hat{V}$: set of MPLS layer nodes
- $(\hat{u}, \hat{v}) \in \hat{\mathcal{E}}$: end points of a tunnel seen at the MPLS layer
- $T(\hat{v})$: available transmitters in node $\hat{v} \in \hat{\mathcal{V}}$
- $R(\hat{v})$: available receivers in node $\hat{v} \in \hat{\mathcal{V}}$
- $\delta_{\hat{u}\hat{v}, p}^{\hat{u}\hat{v}}$: binary matrix. If $\delta_{\hat{u}\hat{v}, p}^{\hat{u}\hat{v}} = 1$ then the link (\hat{u}, \hat{v}) of lightpath (\hat{u}, \hat{v}) is in path p , otherwise 0.
- $\mathcal{P}_{\hat{u}\hat{v}}$ = set of optical paths between the IP/MPLS nodes \hat{u} and \hat{v}
- ρ_p : minimum number of fibers in the path $p \in \mathcal{P}_{\hat{u}\hat{v}}$
- $\sigma_p^{\hat{u}\hat{v}}$: delay of path $p \in \mathcal{P}_{\hat{u}\hat{v}}$ between IP/MPLS nodes \hat{u} and \hat{v}
- $L_{\hat{u}\hat{v}}$ = maximum rate of the wavelengths for virtual tunnel $(\hat{u}, \hat{v}) \in \hat{\mathcal{E}}$

c) *Optical layer*: This layer represent the physical topology of the network and it handles the RWA of the virtual tunnels created in the IP/MPLS layer.

- $\mathcal{G}(\mathcal{V}, \mathcal{E})$: a bi-directed graph representing the optical topology
- $v \in \mathcal{V} = 1, 2, \dots, V$: set of Optical layer nodes
- $(u, v) \in \mathcal{E}$: end points of a directed physical fiber link
- F_{uv} : available fibers in link $(u, v) \in \mathcal{E}$
- L_{uv} : maximum line rate of the wavelengths for link $(u, v) \in \mathcal{E}$
- W : maximum number of wavelength channels per fiber
- β^{IP} : power consumption of the IP/MPLS node based on the total capacity. $\beta^{IP} = 10W/Gbps$.
- $\beta^{Optical}$: power consumption of the transponder/muxponder of the optical node. $\beta^{Optical} = 75W$ for 80Gbps
- μ : power consumption of the WDM terminal with 80 wavelengths. $\mu = 120W$
- γ and ζ : power consumption of the OXC with 80 wavelength channels. $\gamma = 85W$ for each mux/demux for add/drop, and $\zeta = 100W$ for each amplifier.

d) *Service disruption and blocking probability*:

- $\mathcal{I}_{\hat{u}\hat{v}}^{\bar{u}\bar{v}, r}$ traffic route used by the working paths of the request r from \bar{u} to \bar{v} at time $t-1$. The set $\mathcal{I}_{\hat{u}\hat{v}}^{\bar{u}\bar{v}, r}$ is defined by considering the paths used in the previous routing (time period $t-1$) $\lambda_{\hat{u}\hat{v}}^{\bar{u}\bar{v}, r}$ to realize all the connections of demand (\bar{u}, \bar{v}) of slice r .
- θ^r service disruption penalty (due to change of paths) for each slice $r \in \mathcal{R}$
- Ω Blocking penalty

¹Website: <http://www.openroadm.org/home.html> (last access: 17/07/2019)

6) *Variables:* In the following section we describe the variables.

a) *Service layer:* The following variables handle the mapping of the chains and NF on the IP/MPLS layer.

- $y_{\bar{v},\hat{v}}^r$ binary variable ($y_{\bar{v},\hat{v}}^r = 1$ if NF $\bar{v} \in \mathcal{D}$ of service layer request belonging to slice r is mapped to node $\hat{v} \in \hat{\mathcal{V}}$, 0 otherwise).
- $\hat{y}_{\bar{v},\hat{v}}$ binary variable ($\hat{y}_{\bar{v},\hat{v}} = 1$ if NF $\bar{v} \in \mathcal{D}$ of service layer request is mapped to node $\hat{v} \in \hat{\mathcal{V}}$, 0 otherwise).
- $B_{\bar{u}\bar{v}}^r$: binary variable $B_{\bar{u}\bar{v}}^r = 1$ if the traffic request of slice r between (\bar{u}, \bar{v}) is successfully routed.

b) *IP/MPLS layer:* These variables handle the creation of tunnels between the various NFs.

- $\lambda_{\hat{u}\hat{v}}^{\bar{u}\bar{v},r}$: binary variable, $\lambda_{\hat{u}\hat{v}}^{\bar{u}\bar{v},r} = 1$ if the traffic request between $(\bar{u}, \bar{v}) \in \hat{\mathcal{E}}$ of slice r uses the virtual link $(\hat{u}, \hat{v}) \in \hat{\mathcal{E}}$.
- $i_{\hat{u}\hat{v}}^{\bar{u}\bar{v},r}$: binary variable, $i_{\hat{u}\hat{v}}^{\bar{u}\bar{v},r} = 1$ if the new route $\lambda_{\hat{u}\hat{v}}^{\bar{u}\bar{v},r}$ is different from the route of the previous model execution $\mathcal{I}_{\hat{u}\hat{v}}^{\bar{u}\bar{v},r}$.

c) *Optical layer:* These are the variables for the mapping between the IP/MPLS layer and the optical layer. In particular they handle the RWA of the tunnels created in the layer above.

- $P_p^{\hat{u}\hat{v},w,r}$ integer: it is the number of tunnels between nodes $(\hat{u}, \hat{v}) \in \hat{\mathcal{E}}$ of slice r routed through path $p \in \mathcal{P}_{\hat{u}\hat{v}}$, on wavelength $w \in \mathcal{W}$.
- $\hat{P}_p^{\hat{u}\hat{v},r}$ binary ($\hat{P}_p^{\hat{u}\hat{v},r} = 1$ if tunnel $(\hat{u}, \hat{v}) \in \hat{\mathcal{E}}$ of slice r is routed through path $p \in \mathcal{P}_{\hat{u}\hat{v}}$).
- $S_{\hat{u}\hat{v}}^r$: tells the maximum delay for a path between end-points (\hat{u}, \hat{v}) on the slice r
- f_{uv} amount of active fibers in link $(u, v) \in \mathcal{E}$
- $V_{\hat{u}\hat{v}}$ number of active wavelengths shared by the tunnels on the virtual link from node \hat{u} to node \hat{v} .
- $V_{\hat{u}\hat{v}}^{w,r}$ integer: the number of virtual tunnels from node \hat{u} to node \hat{v} on wavelength $w \in \mathcal{W}$ in slice $r \in \mathcal{R}$
- $T_{\hat{u}\hat{v}}^{w,r}$, continuous: it is the percentage of traffic used in the tunnel from node \hat{u} to node \hat{v} on wavelength $w \in \mathcal{W}$ in slice $r \in \mathcal{R}$. This percentage value can be greater than 1. This means that the model creates multiple tunnels for the same traffic on the same wavelength. This can be translated in the usage of multiple fibers in the same physical links.

7) *MILP formulation:* The objective function and all the constraints are presented and explained in this section. In Table I, the expression (1) is a multi-objective function composed by a linear combination of weighted variables. Thanks to these weights it is possible to model different kind of optimization policies according to the user needs. We refer to the power consumption model in [23]. Constraints (2) and (3) enforce that every NF is mapped on a physical node among the available ones. Expressions (4) and (5) prevent the use of a NF more than a certain number of times; such number is specified as input of the MILP. In each node of the MPLS Layer, the number of tunnels coming in or going out from the nodes must be less than the number of transmitters or receivers in the node, thus is enforced by constraints (6) and (7). (8) ensures that the number of virtual links from two nodes in the MPLS Layer is the sum of the tunnels on each wavelength between the two

nodes. Constraints (9) (10) find the percentage of capacity used and the usage of each MPLS tunnel and wavelength for each demand. The constraint (11) ensures that all the paths flowing on a node for each wavelength do not exceed the number of active fibers present in the optical link. (12) puts a limit to the number of fibers for each link. (13) ensures that there is at least one physical path for each optical tunnel of the IP/MPLS layer, while (14) and (15) map the the path on the physical layer for each IP/MPLS tunnels with different source and destination nodes. (16) creates a tunnel between two nodes only if there exist traffic between the NFs associated to the MPLS nodes. The expression (17) ensures that the traffic is successfully routed and (18) tells us if there exists at least one path between a source and a destination node for a certain SFC. For each tunnel, equation (20) searches the longest optical path in terms of propagation delay.

(21) is the latency constraint for each SFC and ensures that all the traffic demands do not exceed the end-to-end latency requirement. In this work, we do not consider a fixed latency split between the VNFs of the SFCs. The total delay computation is performed as the sum of delay of each part of the SFC.

Expression (19) is a linearization of the expression $i_{\hat{u}\hat{v}}^{\bar{u}\bar{v},r} = (\mathcal{I}_{\hat{u}\hat{v}}^{\bar{u}\bar{v},r} - \lambda_{\hat{u}\hat{v}}^{\bar{u}\bar{v},r})^2$. The variable $i_{\hat{u}\hat{v}}^{\bar{u}\bar{v},r}$ assigned in this equation is minimized by the objective function and it prevents the disruption of a tunnel when the model takes as input the previous configuration of the tunnels.

B. Two-step approach

In the two-step approach, the original problem is divided in two sub-problems: 1) the first one solves the VNF placement; 2) the second solves the RWA. The result of the first step is passed to the second one in order to decrease its complexity.

The two-step approach starts by solving an ILP that involves the first two layers of the network: service and IP/MPLS. It focuses on the positioning of the VNFs on the IP/MPLS layer taking into account the delay constraints of each slice. The objective function maximizes the number of successfully routed chains and minimize the reconfiguration penalty.

The solution of the positioning problem provides the set of traffic demands that is the input to the RWA problem. The second objective function to minimize is the total power consumption of the network considering the power used by the IP/MPLS and the optical components.

We present an alternative simpler formulation of the problem. The difference with the previous model is that the MILP is divided in two parts in order to decrease the solving time and complexity. These two parts are the VNF positioning problem (III-B1) and the Routing and Wavelength Assignment problem (III-B2). The result of the first problem is passed to the second one in order to decrease its complexity, since the RWA is the most computationally expensive part of the model.

1) *VNF positioning problem:* The two-step approach starts by solving an ILP that involves the first two layers of our problem statement (1): service and IP/MPLS layers. It focuses on the positioning of the VNFs on the IP/MPLS layer taking into account the capacity constraints of the links of the network

TABLE I: Complete MILP formulation

Objective: Minimize blocking and consumed bandwidth	
$\min \Omega \sum_{r \in \mathcal{R}} (1 - B^r) + (\mu_{uv} + \gamma_{uv} + \zeta_{uv}) \sum_{(u,v) \in \mathcal{E}} f_{uv} + (L_{\hat{u}\hat{v}} \beta^{IP} + \beta_{\hat{u}\hat{v}}^{Optical}) \sum_{(\hat{u},\hat{v}) \in \hat{\mathcal{V}}} V_{\hat{u}\hat{v}} + \sum_{r \in \mathcal{R}} \sum_{(\hat{u}\hat{v}) \in \hat{\mathcal{E}}} \sum_{(\bar{u}\bar{v}) \in \bar{\mathcal{E}}} \theta^r i_{\hat{u}\hat{v}}^{\bar{u}\bar{v},r}$ (1)	
Node mapping, MPLS tunnels, Tunnels mapping	Link mapping and Service disruption
$\sum_{\hat{v} \in \hat{\mathcal{V}}} \hat{y}_{\bar{v},\hat{v}}^r = m_{\bar{v}}^r \quad \forall \bar{v} \in \mathcal{D}, r \in \mathcal{R}$ (2)	$\sum_{p \in \mathcal{P}_{\hat{u}\hat{v}} (u,v) \in \mathcal{E}} \sum_{r \in \mathcal{R}} \sum_{(\hat{u},\hat{v}) \in \hat{\mathcal{E}}} \delta_{uv}^{\hat{u}\hat{v},p} P_p^{\hat{u}\hat{v},w,r} \leq f_{uv}, \quad \forall (u,v) \in \mathcal{E}, w \in \mathcal{W}$ (11)
$\hat{y}_{\bar{v},\hat{v}} \leq l_{\bar{v},\hat{v}}; \quad \forall \bar{v} \in \mathcal{D}, \hat{v} \in \hat{\mathcal{V}}, r \in \mathcal{R}$ (3)	$f_{uv} \leq F_{uv}, \quad \forall (u,v) \in \mathcal{E}$ (12)
$\sum_{\hat{v} \in \hat{\mathcal{V}}} \hat{y}_{\bar{v},\hat{v}} \leq n_{\bar{v}} \quad \forall \bar{v} \in \mathcal{D}$ (4)	$\sum_{p \in \mathcal{P}_{\hat{u}\hat{v}}} P_p^{\hat{u}\hat{v},w,r} = V_{\hat{u}\hat{v}}^{w,r}, \quad \forall (\hat{u},\hat{v}) \in \hat{\mathcal{E}} \hat{u} \neq \hat{v}, w \in \mathcal{W}, r \in \mathcal{R}$ (13)
$\hat{y}_{\bar{v},\hat{v}}^r \leq \hat{y}_{\bar{v},\hat{v}}$ $\forall \bar{v} \in \mathcal{D}, \bar{r} \in \mathcal{R}, \hat{v} \in \hat{\mathcal{V}}$ (5)	$\sum_{w \in \mathcal{W}} \sum_{p \in \mathcal{P}_{\hat{u}\hat{v}}} P_{uv}^{\hat{u}\hat{v},w,r} \geq \lambda_{\hat{u}\hat{v}}^{\bar{u}\bar{v},r}, \quad \forall (\bar{u},\bar{v}) \in \bar{\mathcal{E}}, (\hat{u},\hat{v}) \in \hat{\mathcal{E}} \hat{u} \neq \hat{v}, r \in \mathcal{R}$ (14)
$\sum_{\hat{v} \in \hat{\mathcal{V}}: (\hat{u},\hat{v}) \in \hat{\mathcal{E}}, \hat{u} \neq \hat{v}} V_{\hat{u}\hat{v}} \leq T(\hat{u}) \quad \forall \hat{u} \in \hat{\mathcal{V}}$ (6)	$P_p^{\hat{u}\hat{v},w,r} \leq \rho_p \sum_{(\bar{u}\bar{v}) \in \bar{\mathcal{E}}} \lambda_{\hat{u}\hat{v}}^{\bar{u}\bar{v},r} \quad \forall (\hat{u},\hat{v}) \in \hat{\mathcal{E}} \hat{u} \neq \hat{v}, p \in \mathcal{P}_{\hat{u}\hat{v}}, r \in \mathcal{R}, w \in \mathcal{W}$ (15)
$\sum_{\hat{u} \in \hat{\mathcal{V}}: (\hat{u},\hat{v}) \in \hat{\mathcal{E}}, \hat{u} \neq \hat{v}} V_{\hat{u}\hat{v}} \leq R(\hat{v}) \quad \forall \hat{v} \in \hat{\mathcal{V}}$ (7)	$\lambda_{\hat{u}\hat{v}}^{\bar{u}\bar{v},r} \leq \eta_{\bar{u}\bar{v}}^r * y_{\bar{u},\bar{u}}^r * y_{\bar{v},\bar{v}}^r \quad \forall (\bar{u},\bar{v}) \in \bar{\mathcal{E}}, (\hat{u},\hat{v}) \in \hat{\mathcal{E}}, r \in \mathcal{R}$ (16)
$\sum_{w \in \mathcal{W}, r \in \mathcal{R}} V_{\hat{u}\hat{v}}^{w,r} = V_{\hat{u}\hat{v}} \quad \forall (\hat{u},\hat{v}) \in \hat{\mathcal{E}}, \hat{u} \neq \hat{v}$ (8)	$B^r = \sum_{(\hat{u}\hat{v}) \in \hat{\mathcal{E}}} \lambda_{\hat{u}\hat{v}}^{\bar{u}\bar{v},r} \quad \forall (\bar{u},\bar{v}) \in \bar{\mathcal{E}}, r \in \mathcal{R}$ (17)
$\sum_{(\bar{u},\bar{v}) \in \bar{\mathcal{E}}} h_{\bar{u}\bar{v}}^r \lambda_{\hat{u}\hat{v}}^{\bar{u}\bar{v},r} = \sum_{w \in \mathcal{W}} T_{\hat{u}\hat{v}}^{w,r} L_{\hat{u}\hat{v}}$ $\forall (\hat{u},\hat{v}) \in \hat{\mathcal{E}} \hat{u} \neq \hat{v}, r \in \mathcal{R}$ (9)	$\rho_p \hat{P}_p^{\hat{u}\hat{v},r} \geq P_p^{\hat{u}\hat{v},w,r} \quad \forall (\hat{u},\hat{v}) \in \hat{\mathcal{E}} \hat{u} \neq \hat{v}, p \in \mathcal{P}_{\hat{u}\hat{v}}, r \in \mathcal{R}, w \in \mathcal{W}$ (18)
$T_{\hat{u}\hat{v}}^{w,r} \leq V_{\hat{u}\hat{v}}^{w,r}$ $\forall (\hat{u},\hat{v}) \in \hat{\mathcal{E}}, w \in \mathcal{W}, r \in \mathcal{R}$ (10)	$\mathcal{I}_{\hat{u}\hat{v}}^{\bar{u}\bar{v},r} + \lambda_{\hat{u}\hat{v}}^{\bar{u}\bar{v},r} - 2\mathcal{I}_{\hat{u}\hat{v}}^{\bar{u}\bar{v},r} \lambda_{\hat{u}\hat{v}}^{\bar{u}\bar{v},r} = i_{\hat{u}\hat{v}}^{\bar{u}\bar{v},r} \quad \forall (\hat{u},\hat{v}) \in \hat{\mathcal{E}}, (\bar{u},\bar{v}) \in \bar{\mathcal{E}}, r \in \mathcal{R}$ (19)
Latency constraints	
$S_r^{\hat{u}\hat{v}} \geq \sigma_p^{\hat{u}\hat{v}} \hat{P}_p^{\hat{u}\hat{v},r}; \quad \forall r \in \mathcal{R}, (\hat{u},\hat{v}) \in \hat{\mathcal{E}} \hat{u} \neq \hat{v}, p \in \mathcal{P}_{\hat{u}\hat{v}}$ (20)	
$\sum_{(\hat{u}\hat{v}) \in \hat{\mathcal{E}} \hat{u} \neq \hat{v}} S_r^{\hat{u}\hat{v}} \leq b^r, \quad \forall r \in \mathcal{R}$ (21)	

and the delay constraints of each slice, but keeping an eye on the RWA problem. The objective function is the following:

Minimize:

$$\Omega \sum_{r \in \mathcal{R}} (1 - B^r) + \sum_{r \in \mathcal{R}} \sum_{(\hat{u}\hat{v}) \in \hat{\mathcal{E}}} \sum_{(\bar{u}\bar{v}) \in \bar{\mathcal{E}}} \theta^r i_{\hat{u}\hat{v}}^{\bar{u}\bar{v},r}$$

It maximizes the number of successfully routed chains and minimize the reconfiguration penalty. It is important to notice that it does not minimize energy consumption. The constraints are the same as the complete problem (III-A7). This model removes only a variable (f_{uv}) and the set of constraints relative to this variable, such as (11) and (12).

2) *RWA problem:* The solution of the VNF positioning problem provides the set of traffic demands to the second ILP that regards the RWA problem. The two-step approach considerably simplifies the complete formulation problem, however it may lead to sub-optimal solutions. The following objective function tries to minimize the total power consumption of the network considering the power used by the IP/MPLS and optical components.

Minimize:

$$(\mu_{uv} + \gamma_{uv} + \zeta_{uv}) \sum_{(u,v) \in \mathcal{E}} f_{uv} + (L_{\hat{u}\hat{v}} \beta^{IP} + \beta_{\hat{u}\hat{v}}^{Optical}) \sum_{(\hat{u},\hat{v}) \in \hat{\mathcal{V}}} V_{\hat{u}\hat{v}}$$

This formulation is similar to the complete one in Section III-A7, the only differences are in the objective function, that takes into consideration only the power consumption. There are no constraints about the node mapping (2)-(5) and service disruption (19), since they are already been solved by the

positioning problem. Moreover, the λ s in the MILP are inputs of the problem and not variables.

IV. NETWORK SCENARIOS

Considering the Metro-Haul project, we have designed two network scenarios. The first one represents a basic network in which we have two rings and a mesh core network, see Fig. 2. The second is composed by four different rings simulating a smart city network, see Fig. 3.

A smart city incorporates information and communication technologies (ICT) to enhance the quality and performance of urban services such as energy, transportation and utilities in order to enhance the quality of living for its citizens through smart technology [24]. Smart city mainly focuses on applying the next-generation information technology, including 5G, to all walks of life, enabling different kind of network services, such as: smart industry, virtual reality and remote surgery. Specifically, 5G technology allows the application of these network services that require strict QoS.

A. Basic network

The topology represented in Fig. 2 is composed by two ring networks connected by a core mesh.

It is composed by four MCENs. Each pair of MCENs is connected to a metro ring. One of the MCEN contains a Central Office (CO) with an Internet eXchange Point (IXP) and a Data Center (DC). Each link of the core mesh is 40km long. The other two rings represent metropolitan network segment. The AMENs (Fig. 2) are the nodes from which

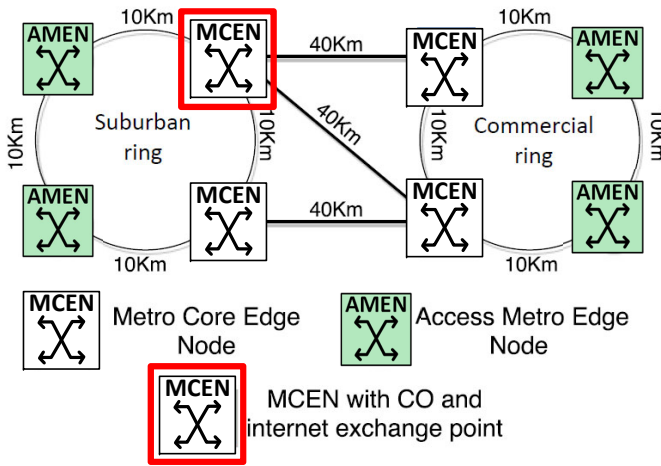


Fig. 2: Basic network topology

the traffic is injected in the network. The links in the two rings have a length of 10km . Each physical link contains two fibers, each one with 20 wavelengths with a capacity of 10Gbps/wavelength .

B. Smart city network

In Fig. 3, we used a network topology typical of an urban metro-core network [25] [26] that supports typical smart city services, such as smart industry, virtual reality applications, etc.

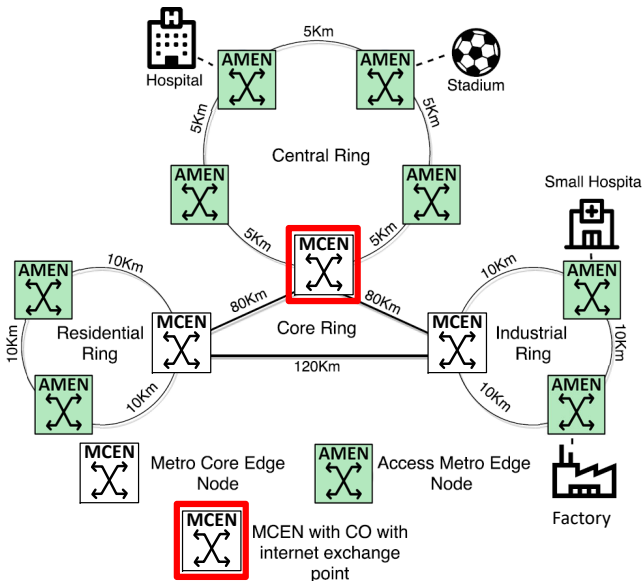


Fig. 3: Smart city network topology

The core ring represents the backbone of the network where most of traffic flows into the global Internet. It is composed by three MCENs, where one of them contains a CO (with an IXP) as in the basic topology. Each core link contains two fibers with 20 wavelengths each at $100\text{Gbps/wavelength}$. The other

three rings are located into different areas of the city. Each of them contains some AMENs that inject traffic in the network and a MCEN that connects the ring to the core network. The central ring represents the core of the city where we can find a stadium and a big hospital. This ring is characterized by dense traffic throughout. The industrial ring, that is composed by three nodes, contains a small hospital and a factory. Here the traffic is not as dense as the central ring, but it is high during the working hours (from 9AM to 5PM). Last, the residential ring is characterized by high traffic during the early hours of the day (from 6AM to 10AM) and evening/night times (from 4PM to 11PM) and it is composed by three nodes.

V. NETWORK SLICES

Following the Metro-Haul guidelines [7], we have defined a list of 6 network slices. For each slice we show the peak of traffic it uses, its maximum allowed latency and the SFCs associated to them. We divided the list in two parts according to the type of traffic: variable and constant. In Table II we show the list of slices with variable traffic. For each of them we defined a signature that describes its traffic variation over 24 hours [22]. Likewise, in Table III we list the slices that have constant traffic, thus there is no traffic variation when they are active.

A. Slices with variable traffic

Considering Table II, the *Mobile Broadband (MBB)* slice handles all the mobile cellular internet traffic in the network, except for the video traffic. In the radio access network (RAN), there are two main deployment options. The traditional deployment where the baseband processing is performed at the antenna site, and the centralized RAN (C-RAN) deployment where baseband processing units (BBUs) are centralized in a BBU pool, and digitized radio signals are transmitted to remote radio units (RRUs) over a fronthaul network [27]. The virtual Evolved Packet Core (vEPC) functions can be placed at different locations depending on the requirements defined by 3GPP [5] specifications. In this work, we followed the C-RAN functional split in which the SFC is composed by: RRU (Remote Radio Unit), BBU (Base Band Unit) and vEPC (virtual Evolved Packet Core), see Table II. The traffic, instead, is modelled by using as reference the real traffic traces generated by citizens in the city of Milan (Italy). In [22] we performed a study in which we used these traces to extract typical traffic patterns, called signatures. In particular, there are two kinds of traffic signatures. The *commercial* one presents peaks during the afternoon hours. Then we have the *suburban* traffic that presents peaks during the morning and night hours. For both *commercial* and *suburban* traffic their peak is 50Gbps , see Fig. 4.

Mobile video slice handles 75% of mobile data traffic in the network². Differently from the MBB slice, the mobile video includes the cache as VNF in the SFC.

The *cloud data* slice emulates the traffic flows among distributed DCs. These data centers host applications for social

²According to Ericsson network traffic measurements [28], in 2022 video will account for around 75% of mobile data traffic.

TABLE II: List of network slices with variable traffic.

Mobile broadband (MBB)	
Latency	100 ms
SFCs	1) RRU - BBU - vEPC 2) vEPC - BBU - RRU
Traffic volume	50 Gbps
Mobile video	
Latency	20 ms
SFCs	1) Cache - vEPC - BBU - RRU 2) RRU - BBU - vEPC - Cache
Traffic volume	37.5 Gbps
Cloud data	
Latency	50 ms
SFC	DC - DC
Traffic volume	500 Gbps

networking, e-commerce, and video hosting. Cloud traffic is characterized by a high volume during night hours and a lower one during day hours. Fig. 4 shows how the traffic volume changes during the 24 hours of the day reaching a peak of 500Gbps [29].

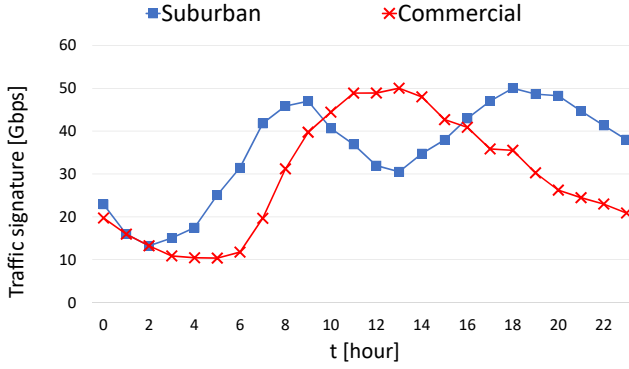


Fig. 4: Commercial and cloud traffic signatures.

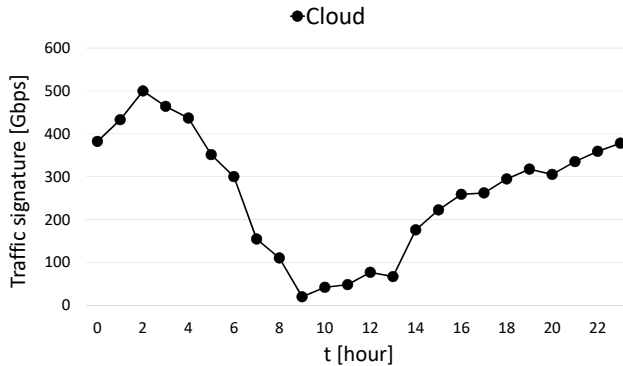


Fig. 5: Suburban traffic signature.

B. Slices with constant traffic

Considering Table III, the *smart industry* slice represents an environment where machinery and equipment are able to improve processes through automation and self-optimization. Thus it is possible to create a lot of different SFCs, one for

each smart application the factories implement [30], see Table III. The main SFCs envisioned for the smart industry are:

- *Robot*: this SFC is a model for using robotic systems remotely. It comprises the following VNFs: Industry, NAT, Firewall and Robot application. The NAT and firewall ensure security to the Industry.
- *Massive internet of things (mIoT)*: it is a model for using IoT technology trend with feature of the mobility and the transitioning to serve smart environments. This field integrates IoT concept with mobile applications.
- *Video surveillance*: the video surveillance SFC provides the services and resources for the management of the video surveillance of an industrial plant.
- *Automated vehicles*: this SFC represents the system for handling automated guided vehicles for material transportation.
- *Virtual Reality - Core*: this chain represents the streaming of the stadium cameras data to a streaming server, where the data is processed and encoded for the distribution to the users. In this scenario the stadium contains 8 cameras that stream the football matches.
- *Virtual Reality - User*: this chain represents the streaming of the stadium cameras content to the users spread in the network.
- *Remote surgery*: this slice represent a remote surgery case [31] [32]. A doctor needs to make a surgery to a patient in another hospital building. To do so it is possible to use this technology called tele-surgery, where the surgeon views the patient through a 2D/3D camera, and instruct a machine in another hospital on how to move.

VI. SIMULATIONS AND RESULTS

Simulations spans over a complete day with one hour intervals, thus the optimization models, complete and two-step, are executed 24 times. At hour 0 the models are executed to deploy a certain number of SFCs; then, they run at each hour of the day in order to find the best balance for: 1) maximizing the number of routed SFCs, 2) minimizing the number of IP/MPLS tunnel reconfigurations, 3) minimizing the physical and IP/MPLS layer power consumption. To test the proposed models, we have defined two use cases:

- 1) Use case 1: we have run the complete MILP and the two-step approach to find the best resource allocation in order to deploy the slices listed in Table II: 4 *MBB*, 3 *cloud data* and 4 *mobile video* in the basic network scenario.
- 2) Use case 2: we used the whole set of slices considered in the previous use case and the ones described in Table III to be implemented in the smart city network. As for the use case 1, we run both optimization models assuming that the slices of Table III are active all the day.

In order to show the performance of the optimization models, we have analyzed the results in terms of power consumption, number of reconfigurations and run time. We run the simulations in an Ubuntu 16.04 x64 machine with 32GB of RAM, Intel(R) Core(TM) i7-7700 CPU @ 3.6GHz using CPLEX 12.6.1 solver through the Python API (python 2.7.12).

TABLE III: List of network slices with constant traffic.

Smart industry - Robot	
Latency	100 ms
SFCs	1) Industry - NAT - Firewall - Robot App 2) Robot App - Firewall - NAT - Industry
Traffic volume	0,1 Gbps per km^2
Smart industry - Massive internet of things (mIoT)	
Latency	1 ms
SFCs	1) Industry - NAT - Firewall - mIoT App 2) mIoT App - Firewall - NAT - Industry
Traffic volume	1,1 Gbps per km^2
Smart industry - Video surveillance	
Latency	10 ms
SFCs	1) Industry - NAT - Firewall - Video App 2) Video App - Firewall - NAT - Industry
Traffic volume	5 Gbps per km^2
Smart industry - Automated vehicles	
Latency	50 ms
SFCs	1) Industry - NAT - Firewall - Vehicle App 2) Vehicle App - Firewall - NAT - Industry
Traffic volume	1 Gbps per km^2
Virtual Reality - Core	
Latency	50 ms
SFCs	1) Stadium - VR Content - VR Streaming 2) VR Streaming - VR Content - Stadium
Traffic volume	1 Gbps per camera
Virtual Reality - User	
Latency	50 ms
SFCs	1) Stadium - VR Content - VR Streaming 2) VR Streaming - VR Content - Stadium
Traffic volume	800 Mbps
Remote surgery	
Latency	350 ms
SFCs	1) Hospital - E-Health - Small Hospital 2) Small Hospital - E-Health - Hospital
Traffic volume	2 Gbps

A. Use Case 1: Complete vs. two-step

In this use case, we considered the basic network in Fig. 2 as reference topology. We have run the complete MILP and the two-step approach considering the mobile traffic signatures showed in Fig. 4.

We want to compare the complete MILP model and the two-step approach. We dimensioned the network resources so to avoid any blocking of connections in order to focus on the effects of the reconfiguration penalty on the optimization. We show how the two models perform assuming that they have the same reconfiguration penalty θ set initially to a high value, and then to a low value. Since it is a free input parameter, we have defined the low and upper bound experimentally by running the models with different values of θ . As a result, $\theta = 10$ corresponds to the highest reconfiguration penalty, while the values between 0 and 3 correspond to a lowest ones. Considering $\theta = 10$, we expect to have a low number of reconfigurations or no reconfigurations at all throughout the day. As depicted in Fig. 6, the two-step approach reconfigures the IP/MPLS tunnels just three times during the day, for a total of 270 reconfigurations per day. This effect is a consequence of the objective function of the VNF positioning problem. Choosing $\theta = 3$, we have a high number of reconfigurations, that are in total 850.

Let's see what is the solution of the complete MILP model. Its objective function includes both energy and reconfiguration penalty, so it has a wider decision space than the two-step

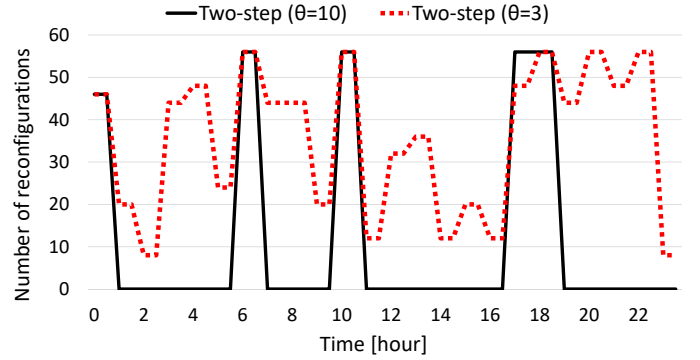


Fig. 6: Number of reconfigurations of VNF positioning with high and low reconfiguration penalty performed by the two-step approach.

approach. Fig. 7 shows the number of reconfigurations for both cases: high ($\theta = 10$) and low ($\theta = 3$) reconfiguration penalty. In particular, considering $\theta = 10$ we obtained 192 total reconfigurations per day; instead, for $\theta = 3$ we got a total of 282 reconfigurations.

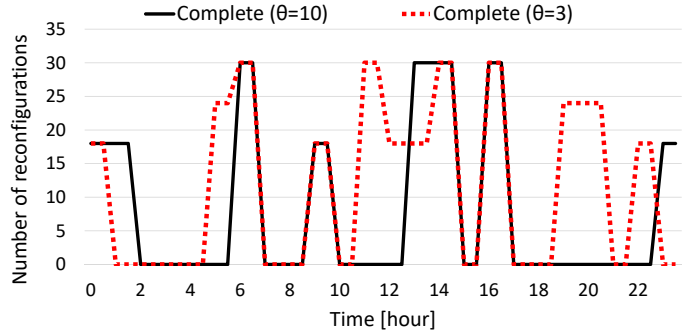


Fig. 7: Number of reconfigurations of VNF positioning with high and low reconfiguration penalty performed by the complete model.

In Fig. 8, we show the power consumption at each hour of the day of the network for both models with high reconfiguration penalty ($\theta = 10$). As expected, the complete MILP model is able to find a better result compared to the two-step approach. In this particular scenario, if we analyze the cumulative power consumption, the power optimality gap of the two-step model at the end of the simulations is about $100kW/h$, which is 15% of the optimal solution found by the complete model. Differently, choosing $\theta = 3$, the two-step approach performs better in terms of power consumption respect to the complete model. The power optimality gap at the end of the simulations is about $88kW/h$, which is 21% of the optimal solution found by the two-step approach.

B. Use Case 1: reconfiguration predictions

So far we have obtained the results by considering the mobile traffic signatures derived by the work in [22]. We tested the two-step approach with several days of real traffic to see if we are able to predict the future reconfiguration time

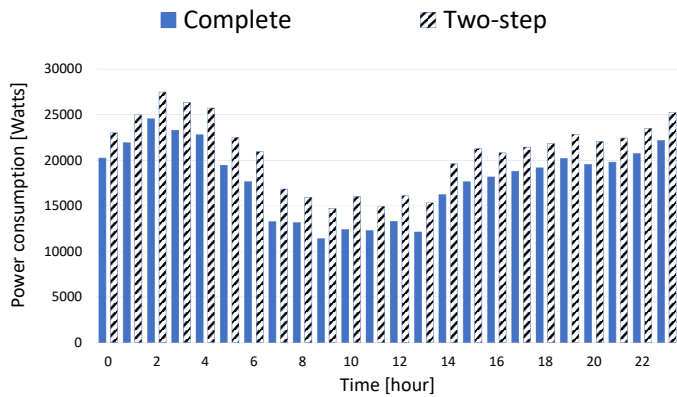


Fig. 8: Power consumption with high reconfiguration penalty ($\theta = 10$).

instants. In Fig. 9 we compare the number of reconfigurations considering the signatures and three days of real traffic data. As expected, the two-step approach performs the reconfigurations during the same time instants with respect to the case in which we consider the signatures. This result confirms the effectiveness of the use of traffic signatures to predict the moments when the network will have to be reconfigured. In addition, a network/service operator is able to predict what will be the use of their network resources and will be able to proactively intervene to avoid congestions and to reduce the blocking probability of traffic requests.

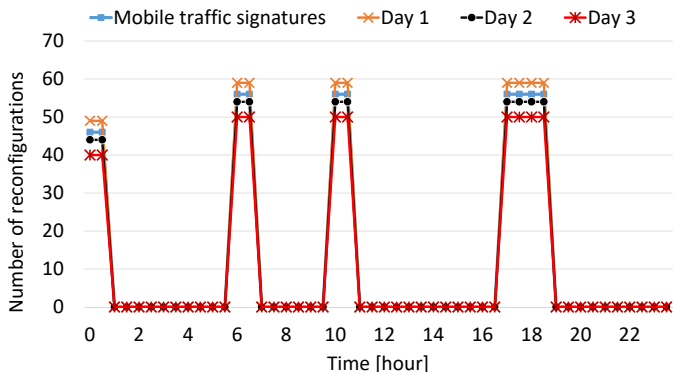


Fig. 9: Number of reconfigurations of VNF positioning with high reconfiguration penalty $\theta = 10$ performed by the two-step approach during three days of simulation.

C. Use Case 1: VNF movements

Another interesting result is how the VNFs move during the simulations of the two-step approach with high reconfiguration penalty ($\theta = 10$). In particular we can focus on the vEPC VNF of the MBB chain. This VNF is shared among all the chains of this type, thus it can change its position according to the traffic load injected by the chains it shares. It is possible to see from Fig. 10 that it moves three times in the core mesh during the 24 iterations of the simulation. It starts from node 8 then passes to node 5 in the 6th iteration then on node 7 in the 10th and it returns to node 8 till the end of the simulation,

seven iterations later. It follows the tidal effect of the traffic moving closer to the most loaded ring.

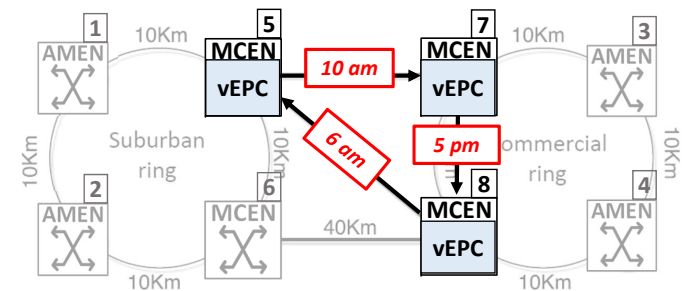


Fig. 10: vEPC movement during 24 iterations of the model for the use case 1.

Considering Fig. 11, the traffic volume of the cloud chain between nodes 7 and 2 is split in two different paths. This is caused by the high capacity request of the chain. If we look at the peak traffic of this slice, 500Gbps, it is possible to show that one path is not enough to route all the traffic. Each physical link is composed by 2 fibers, each one with 20 wavelengths at 10Gbps, for a total of 400Gbps per link. Then, 400Gbps are routed on the path through nodes 7 \rightarrow 8 \rightarrow 6 \rightarrow 2, and 100Gbps from path 7 \rightarrow 5 \rightarrow 1 \rightarrow 2. Splitting the traffic in two optical paths may generate a problem of out-of-sequence of the received packets at the IP/MPLS layer. This problem can be solved by implementing a suitable buffering system in order to compensate the delay of the two optical paths avoiding out-of-sequence packets.

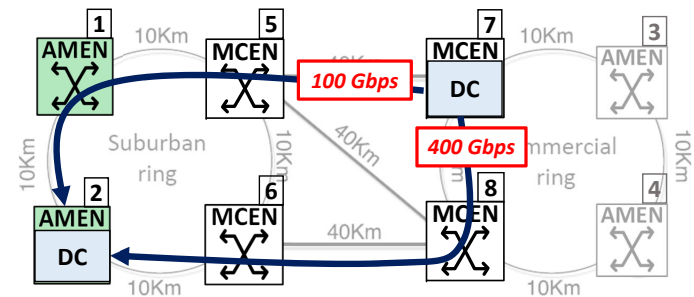


Fig. 11: How DC to DC traffic for use case 1 is split among different lightpaths if the capacity constraints are not met on a single path.

D. Use case 2

This is the most complex scenario because we used the whole set of slices described in Table II and III. All the slices, except for the ones listed in Table II, have a fixed constant traffic injected in the network. The MBB, mobile video and cloud data follow the traffic signature as in the previous use case, instead the slices described in Table III, except for the virtual reality slice, are active and generate traffic for 24 hours. VNFs of the virtual reality slice are active and generate traffic only between the 1PM to 5PM and from 8PM to 11PM to the end of the simulation. We assume this slice as active when

there is a football match in the stadium of the smart city. In addition, the reconfiguration penalty for each chain is set to a meaningful value. In particular the e-health slice has an high penalty, since it needs to have an high availability ($\theta = 10$), so it will not be reconfigured easily; the industrial slices have a value of $\theta = 1$; and the mobile video, MBB and virtual reality do not have a penalty ($\theta = 0$).

We compared the results of the two optimization models in terms of power consumption and number of reconfigurations. Fig. 12 shows that for the first six iterations both models find the same power consumption. This effect is due to the fact that the traffic injected in the network is not high and do not cause congestion on links, therefore both models are able to route the traffic among the VNFs on their shortest path. But when it starts to increase in the 7th hour, the result of the two-step approach starts to diverge from the optimal one. At the end of the simulation its gap from the optimum value is about 4%.

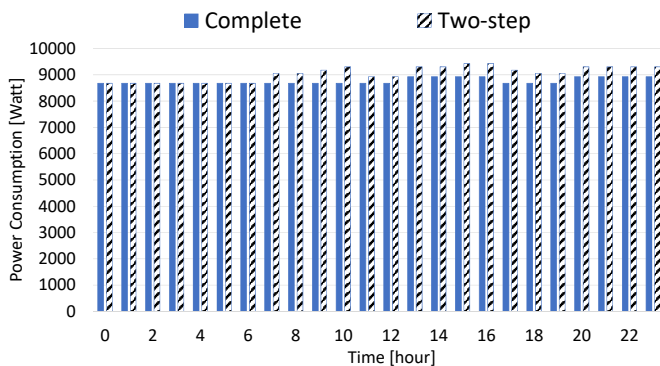


Fig. 12: Power consumption for the use case 2.

From this simulation we can deduce that in a case where the network is not congested the two-step approach performs better than the complete model. Moreover, both models make the same reconfigurations at the same time intervals that are: hour 13th, 17th and 20th respectively; but the two-step model is able to find it 20 times faster than the complete one.

E. Discussion on complexity

The complexity of the complete MILP model comprises two kinds of formulation: the first one, i.e. the VNF positioning problem, is a flow formulation for the IP/MPLS layer and it has been proved to be NP-hard [33]; the second, i.e. RWA problem, is a path formulation for the optical transport layer and it has been proven to be NP-complete [34].

With respect to the complete MILP formulation, in the two-step approach there are less variables and constraints³ given that it considers just the virtual IP/MPLS and physical-optical topology section of the complete MILP formulation. The complexity of this model rely mainly on the RWA formulation. The improvement given by the two-step approach is assessed by in the run time of the optimization problem. In particular, it is about 20 times faster than the complete MILP model. The last takes 37 hours to solve the 24 instances of the problem

³In particular, there are $N(N-1)$ less variables and constraints, where N is the number of nodes. See Appendix A for more details.

(one for each hour of the day), while the two-step model takes less than 2 hours (113 minutes). The number counts of variables and constraints can be found in Appendix A.

VII. CONCLUSIONS

This paper proposes effective optimization mathematical models to dynamically provision network slices in metro-core optical network. Exploiting the programmability and full network visibility leveraged by SDN in the mobile metro-core network, this technique can be deployed as a planning module to perform reconfiguration of the network based on mobile traffic signatures. Our models were tested by considering two network topologies in line with the Metro-Haul European project [7]. Results demonstrate that the use of signatures represents an essential component to optimize the network with dynamic optical routing or other advanced techniques. The proposed optimization models are able to implement network slices in a metro-core network, positioning different VNFs and routing the traffic requested among them. The models consider also the network power consumption and the constraints on capacity and latency of the traffic among the network functions. Both models are able to position the chain and route their traffic. The two-step model is very efficient, but its solution is not optimal. The complete model, instead, finds the best solution but after a long execution time.

With our models, we are able to configure the resource of a multi layer network in a dynamic way. This guarantees that all the traffic requests and chains are routed, if permitted by the capacity and latency constraints of the networks. This is possible because the VNFs can be moved in different nodes for freeing resources for other chains. One of the causes of this movement is the variation of traffic volume. In addition this reconfiguration behaviour is determined also by the penalties given to the reconfiguration of slices. Slices with an high reconfiguration penalty have a less probability to see their VNFs repositioned and their lightpaths reconfigured than slices with low reconfiguration penalties. This mechanism of changing reconfiguration penalties is used to give priority to slices and guarantee their QoS. Indeed, high penalty can be translated to high priority.

ACKNOWLEDGMENT

The work leading to these results has been supported by the European Community under grant agreement no. 761727 Metro-Haul project.

APPENDIX A

NUMBER COUNTS OF VARIABLES AND CONSTRAINTS

To evaluate the computational burden in terms of run time, we computed the number of constraints and variables of the proposed formulations. Table IV summarizes the variables used in the formulations. The complete MILP model comprises two kinds of formulation: the first one is a flow formulation for the IP/MPLS layer, and the second is a path formulation for the optical transport layer. Table V shows the overall number of variables according to their role in the formulation. We can notice that the number of variables increase in a polynomial

way, in particular the ones of the physical-optical topology and traffic route section. The total number of constraints is showed in Table VI. They grow according to the number of nodes and deployed VNFs; in big \mathcal{O} notation it is $\mathcal{O}(N^2)$. If the network dimension is small, then it can be $\mathcal{O}(D^2)$. With respect to the complete MILP formulation, in the two step approach there are $N(N-1)$ less variables and constraints given that it considers just the virtual IP/MPLS and physical-optical topology section. The improvement given by the two-steps approach is visible in the run time of the optimization problem. In particular, it is about 20 times faster than the complete MILP model. The last takes 37 hours to solve the 24 instances of the problem (one for each hour of the day), while the two-steps model takes less than 2 hours (113 minutes).

TABLE IV: Formulations parameters

Name	Description
N	Number of nodes in the network
R	Number of slices
D	Number of network functions
W	Number of wavelengths per fiber
P	Number of precalculated paths

TABLE V: Number of variables of the complete model

Section	Variables
Virtual function mapping	$DN(1+R)$
Virtual IP/MPLS topology	$N^2(1+2WR)$
Physical-optical topology	$N^2R(1+P+PW)+N(N-1)$
Traffic route	$D(D-1)R(1+2N^2)$

TABLE VI: Number of constraints of the complete model

Section	Constraints
Node mapping	$2N^2R+D(1+R)$
IP/MPLS Tunnels	N^2+N
Link Mapping	$N(N-1)(1+W+WR+2WRP)+$ $+D(D+1)R(1+N(2N-1))$
Tunnels Mapping	$N(N-1)(R+W+WR)$
Latency constraints	$R(1+N(N-1)P)$

REFERENCES

- [1] D. Kreutz, F. M. Ramos, P. Verissimo, C. E. Rothenberg, S. Azodolmoly, and S. Uhlig, "Software-defined networking: A comprehensive survey," *Proceedings of the IEEE*, vol. 103, no. 1, pp. 14–76, 2015.
- [2] H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, "Nfv: State of the art, challenges and implementation in next generation mobile networks (vepc)," *arXiv preprint arXiv:1409.4149*, 2014.
- [3] M. Chiosi, D. Clarke, P. Willis, A. Reid, J. Feger, M. Bugenhagen, W. Khan, M. Fargano, C. Cui, H. Deng *et al.*, "Network functions virtualisation: An introduction, benefits, enablers, challenges and call for action," in *SDN and OpenFlow World Congress*, vol. 48. sn, 2012.
- [4] T. Soenen, R. Banerjee, W. Tavernier, D. Colle, and M. Pickavet, "Demystifying network slicing: From theory to practice," in *2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*. IEEE, 2017, pp. 1115–1120.
- [5] T. rep. 3GPP, "Technical specification group services and system aspects; service requirements for the 5g system," Tech. Rep., 2017.
- [6] M. R. Raza, M. Fiorani, A. Rostami, P. Ohlen, L. Wosinska, and P. Monti, "Benefits of programmability in 5g transport networks," in *Optical Fiber Communication Conference*. Optical Society of America, 2017, pp. M2G–3.
- [7] (2019) Metro-haul project. [Online]. Available: <https://metro-haul.eu/>
- [8] (2019) 5g multi-access edge computing (mec). [Online]. Available: <https://www.etsi.org/technologies/multi-access-edge-computing>
- [9] A. Hmaity, M. Savi, F. Musumeci, M. Tornatore, and A. Pattavina, "Protection strategies for virtual network functions placement and service chains provisioning," *Networks*, vol. 70, no. 4, pp. 373–387, 2017.
- [10] N. Huin, B. Jaumard, and F. Giroire, "Optimal network service chain provisioning," *IEEE/ACM Transactions on Networking*, vol. 26, no. 3, pp. 1320–1333, June 2018.
- [11] N. Huin, A. Tomassilli, F. Giroire, and B. Jaumard, "Energy-efficient service function chain provisioning," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 10, no. 3, pp. 114–124, March 2018.
- [12] S. Mehraghdam, M. Keller, and H. Karl, "Specifying and placing chains of virtual network functions," in *2014 IEEE 3rd International Conference on Cloud Networking (CloudNet)*. IEEE, 2014, pp. 7–13.
- [13] C. Ghribi, M. Mechtri, and D. Zeghlache, "A dynamic programming algorithm for joint vnf placement and chaining," in *Proceedings of the 2016 ACM Workshop on Cloud-Assisted Networking*. ACM, 2016, pp. 19–24.
- [14] R. Cohen, L. Lewin-Eytan, J. S. Naor, and D. Raz, "Near optimal placement of virtual network functions," in *2015 IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 2015, pp. 1346–1354.
- [15] G. Portaluri and S. Giordano, "Multi objective virtual machine allocation in cloud data centers," in *2016 5th IEEE International Conference on Cloud Networking (Cloudnet)*, Oct 2016, pp. 107–112.
- [16] L. Askari, A. Hmaity, F. Musumeci, and M. Tornatore, "Virtual-network-function placement for dynamic service chaining in metro-area networks," in *2018 International Conference on Optical Network Design and Modeling (ONDM)*. IEEE, 2018, pp. 136–141.
- [17] X. Wang, C. Wu, F. Le, A. Liu, Z. Li, and F. Lau, "Online vnf scaling in datacenters," in *2016 IEEE 9th International Conference on Cloud Computing (CLOUD)*. IEEE, 2016, pp. 140–147.
- [18] A. Leivadreas, M. Falkner, I. Lambadaris, and G. Kesidis, "Resource management and orchestration for a dynamic service chain steering model," in *2016 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2016, pp. 1–6.
- [19] J. Liu, W. Lu, F. Zhou, P. Lu, and Z. Zhu, "On dynamic service function chain deployment and readjustment," *IEEE Transactions on Network and Service Management*, vol. 14, no. 3, pp. 543–553, 2017.
- [20] J.-J. Pedreno-Manresa, P. S. Khodashenas, M. S. Siddiqui, and P. Pavon-Marino, "On the need of joint bandwidth and nfv resource orchestration: A realistic 5g access network use case," *IEEE Communications Letters*, vol. 22, no. 1, pp. 145–148, 2018.
- [21] O. Ayoub, F. Musumeci, D. Andreoletti, M. Mussini, M. Tornatore, and A. Pattavina, "Optimal cache deployment for video-on-demand delivery in optical metro-area networks," in *2018 IEEE Global Communications Conference (GLOBECOM)*, Dec 2018, pp. 1–6.
- [22] S. Troia, G. Sheng, R. Alvizu, G. A. Maier, and A. Pattavina, "Identification of tidal-traffic patterns in metro-area mobile networks via matrix factorization based model," in *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 2017, pp. 297–301.
- [23] W. Van Heddeghem, F. Idzikowski, W. Vereecken, D. Colle, M. Pickavet, and P. Demeester, "Power consumption modeling in optical multilayer networks," *Photonic Network Communications*, vol. 24, no. 2, pp. 86–102, 2012.
- [24] H. Qin, H. Li, and X. Zhao, "Development status of domestic and foreign smart city," *Global Presence*, vol. 9, pp. 50–52, 2010.
- [25] N. Cvijetic, "Optical network evolution for 5g mobile applications and sdn-based control," in *16th Int. Telecommun. Netw. Strategy Planning Symp. (Networks)*, Sept 2014, pp. 1–5.
- [26] Huawei Technologies, "Technological Developments and Trends of Optical Networks," White Paper, 2016. [Online]. Available: <http://www.huawei.com/>
- [27] P. Ohlén, B. Skubic, A. Rostami, M. Fiorani, P. Monti, Z. Ghebretseae, J. Mårtensson, K. Wang, and L. Wosinska, "Data plane and control architectures for 5g transport networks," *Journal of Lightwave Technology*, vol. 34, no. 6, pp. 1501–1508, 2016.
- [28] (2019) Ericsson network traffic measurements. [Online]. Available: <https://www.ericsson.com/en/mobility-report/mobile-traffic-by-application-category>
- [29] P. Teli, M. V. Thomas, and K. Chandrasekaran, "Big data migration between data centers in online cloud environment," *Procedia Technology*, vol. 24, pp. 1558–1565, 2016.
- [30] M. Wollschlaeger, T. Sauter, and J. Jasperneite, "The future of industrial communication: Automation networks in the era of the internet of things and industry 4.0," *IEEE Industrial Electronics Magazine*, vol. 11, no. 1, pp. 17–27, 2017.

- [31] S. E. Butner and M. Ghodoussi, "Transforming a surgical robot for human telesurgery," *IEEE Transactions on Robotics and Automation*, vol. 19, no. 5, pp. 818–824, Oct 2003.
- [32] Dr. Roger Smith, CTO for Florida Hospital's Nicholson Center, "Research In Lag Time Set To Determine The Future Of Telesurgery," White Paper, 2015. [Online]. Available: <https://techcrunch.com/2015/06/14/lag-time-research-set-to-determine-the-future-of-telesurgery/>
- [33] B. Kar, E. H. Wu, and Y. Lin, "Energy cost optimization in dynamic placement of virtualized network function chains," *IEEE Transactions on Network and Service Management*, vol. 15, no. 1, pp. 372–386, March 2018.
- [34] I. Chlamtac, A. Ganz, and G. Karmi, "Lightpath communications: an approach to high bandwidth optical wan's," *IEEE Transactions on Communications*, vol. 40, no. 7, pp. 1171–1182, July 1992.