IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 19, 2020, accepted October 30, 2020, date of publication November 4, 2020, date of current version November 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3035749

Reconfiguration of Optical-NFV Network Architectures Based on Cloud Resource Allocation and QoS Degradation Cost-Aware Prediction Techniques

VINCENZO ERAMO[1], (Member, IEEE), FRANCESCO GIACINTO LAVACCA[2], TIZIANA CATENA[1], AND FLAVIO DI GIORGIO[1]

[1]DIET, "Sapienza" University of Rome, 00184 Rome, Italy
[2]Fondazione Ugo Bordoni, 00161 Rome, Italy

Corresponding author: Vincenzo Eramo (vincenzo.eramo@uniroma1.it)

ABSTRACT The high time required for the deployment of cloud resources in Network Function Virtualization network architectures has led to the proposal and investigation of algorithms for predicting traffic or the necessary processing and memory resources. However, it is well known that whatever approach is taken, a prediction error is inevitable. Two types of prediction errors can occur that have a different impact on the increase in network operational costs. In case the predicted values are higher than the real ones, the resource allocation algorithms will allocate more resources than necessary with the consequent introduction of an over-provisioning cost. Conversely, when the predicted values are lower than the real values, the allocation of fewer resources will lead to a degradation of QoS and the introduction of an under-provisioning cost. When over-provisioning and under-provisioning costs are different, most of the prediction algorithms proposed in the literature are not adequate because they are based on minimizing the mean square error or symmetric cost functions. For this reason we propose and investigate a forecasting methodology in which it is introduced an asymmetric cost function capable of weighing the costs of over-provisioning and under-provisioning differently. We have applied the proposed forecasting methodology for resource allocation in a Network Function Virtualization architectures where the Network Function Virtualization Infrastructure Point-of-Presences are interconnected by an elastic optical network. We have verified a cost savings of 40% compared to solutions that provide a minimization of the mean square error.

INDEX TERMS Network function virtualization, computing resources, bandwidth resources, elastic optical networks.

I. INTRODUCTION

The Network Function Virtualization (NFV) [1]–[3] was introduced a few years ago to reduce the software maintenance and updating costs of traditional middleboxes. It is based on the execution of Virtual Machines (VM) in datacenters called Network Function Virtual Infrastructure-Point of Presence (NFVI-PoP). The VMs execute software implementing network functions. The network service is composed by a set of network functions referred to as Service Function Chain (SFC). The NFV paradigm has the advantage of allowing a dynamic and flexible allocation of resources (processing, memory and disk resources) necessary to support a service function. The resource allocation problems have been largely investigated [4]–[11] as well as the interconnection problem of the NFVI-PoPs with optical networks [12]–[17] when the support of high bit rate SFCs is needed. While NFV allows for a cloud resource flexible reconfiguration [18], Elastic Optical Network (EON) allows for a flexible bandwidth resource reconfiguration thanks to the allocation of consecutive frequency slots of 6.25 GHz or 12.5 GHz [19]–[21].

Two approaches are possible for the reconfiguration of resources:

- reactive approach: the reconfiguration is activated as soon as traffic change is detected [22]–[27]; the

The associate editor coordinating the review of this manuscript and approving it for publication was Bong Jun David Choi.

technique is ineffective due to the high time needed to reconfigure cloud resources which can take tens of minutes and cause Quality of Service (QoS) degradation;
- proactive approach: the reconfiguration is triggered in advance based on a traffic or necessary resources prediction [28].

Among the reactive approaches, Ghaznavi *et al.* [29] propose consolidation algorithms based on horizontal scaling techniques in which the processing capacity is varied by increasing/decreasing the Virtual Network Function Instance (VNFI) without changing the processing capacity allocated to each VNFIs. An over-sized static resource allocation to the VNFIs may avoid reconfigurations but with this over-allocation the financial benefits would occur only in low cloud resource cost scenario when the objective of the network provider is only to avoid reconfiguration in order not to pay QoS degradation penalty to users.

To avoid complex NFV state management issue, solutions based on vertical scaling techniques [30] in which the VNFIs are dimensioned to achieve the processing capacity required by the traffic, have been also investigated; when the traffic increases/decreases, rather than adding/removing VNFIs, their processing capacity is increased/decreased.

The dynamic allocation of resources to the VNFI leads to high reconfiguration times if the operation is not performed in advance. For this reason most recent research on the reconfiguration of NFV network architectures follows the proactive approach and are based on the application of Artificial Intelligence (AI) techniques [28].

Unfortunately, prediction techniques are not able to predict exactly the traffic or the necessary bandwidth and cloud resources because there is always an innovative component that cannot be predicted. When prediction errors occur, there is an increase in operational cost, which can be of two types:
- Over-Provisioning (OP) cost: it is the cost of additional bandwidth and cloud resources that the network operator needs to use when the prediction algorithm produces traffic or needed resources overestimates;
- Under-Provisioning (UP) cost: it is the cost of compensation to the user that the network operator has to pay for the QoS degradation that occurs when the prediction algorithm produces traffic or needed resources underestimates.

The costs mentioned above may have a different impact on the operational cost and that depends on the service type to be supported. The reconfiguration techniques based on prediction and proposed in literature fail to minimize the operational cost because they are mainly based on the minimization of a symmetric error function (i.e. Mean Squared Error (MSE), Mean Absolute Error (MAE),...) and therefore they are not able to weigh differently the UP and OP costs.

The main contribution of this work is to propose a prediction technique which, aware of the fact that traffic cannot be accurately predicted, tries to overestimate or underestimate traffic in relation to the values of OP and UP costs. This objective is achieved by minimizing an asymmetric cost function characterized by a parameter that takes into account the OP and UP costs.

The proposed methodology can be applied for any prediction technique, both traditional and those based on the application of AI. In this article we illustrate the proposed methodology for the following two prediction techniques: the first one based on Seasonal Autoregressive Integrated Moving Average (SARIMA) traditional models; the second one based on the application of Long Short Term Memory (LSTM) neural networks.

To our best knowledge, only Bega [31]–[33] proposes a solution for Mobile Network Resource Orchestration in which the different values of the OP and UP costs are taken into account. Deep Cognitive framework is proposed for the resource allocation to slicing in a 5G mobile environment. It is based on a deep learning technique in which the cost function attributes a rising cost as the amount of over-allocated resources increases and a constant penalty, that is independent of the lost traffic amount, when a QoS degradation occurs.

In this article we also propose a solution in which OP and UP costs are considered and our work differs from [31]–[33] in the following points:
- our solution is proposed for a NFV network scenario where the data center and network model is well detailed and articulated and the application is mainly based on the NFV implementation of middleboxes;
- the proposed procedures are based on SARIMA and LSTM prediction techniques that are different from those proposed in [31] where convolutional neural networks are considered;
- the function costs are characterized by parameters that can be set for a cost penalty not necessarily constant but related to the amount of traffic lost during the under-provisioning periods.

The paper is organized as follows. The related work is mentioned in Section II. We describe the problem statement in Section III. The SARIMA traffic forecasting technique based on asymmetric cost function is illustrated in Section IV. The asymmetric LSTM traffic forecasting technique is described in Section V. The numerical results, reported in Section VI, show the effectiveness of the proposed technique with respect to MSE-based traditional forecasting techniques in an NFV network environment. Appendix A describes an extension of the European Telecommunications Standards Institute (ETSI) NFV architecture for the support of the proposed prediction and reconfiguration algorithms. Appendix B reports the evaluation of a parameter of the SARIMA-based forecasting technique with asymmetric cost function.

II. RELATED WORK AND RESEARCH MOTIVATION

Reactive resource reconfiguration [12]–[14], [26] approaches in NFV networks have shown all their limits in terms of QoS degradation due to the high time needed to change the allocated cloud resources (increase/decrease of cores allocated, instantiation/removing of VNFIs,...) [28]. For this reason recently the focus has been on a proactive approach where

cloud resources can be reconfigured in advance thanks to a traffic [34] or allocated resource prediction [28]. A traffic prediction-based approach is proposed in [34], [35] in the case of NFV networks in which the NFVI-PoP are interconnected by an EON [34], [35]; the SFC traffic parameters are predicted in [36], [37] by applying a LSTM recurrent network.

Tang *et al.* [38] proposes a traffic prediction method for scaling resources in NFV environments based on traffic modeling with an Autoregressive Moving Average (ARMA); the predicted traffic values are obtained by minimizing MSE.

Among the solutions based on the prediction of the resources to be allocated, Farahnakian *et al.* [39] proposes regressive algorithms for estimating memory and processing consumption in cloud datacenters; the proposed solutions are based on Linear Regression [40] and K-Nearest Neighbor Regression (K-NNR) [41] methods that notoriously determine the prediction by minimizing symmetric error functions. A VNF migration algorithm is proposed and investigated in [42]; it is based on a deep belief network framework to predict the future resource requirements; the authors show how the proposed solution can obtain better estimates of CPU resources than a solution based on Back Propagation Neural Network [43] in terms of MSE. Some solutions [28], [44]–[46] have been proposed on the prediction of host load in cloud infrastructures; these solutions are based on time series forecasting with LSTM recurring neural networks; however, all are based on minimizing MSE.

Other approaches have been proposed that are based on machine learning classification procedures; for example in Rahman's work [47] the classification problem is to choose the best VNFI resource scaling actions to minimize operational cost and QoS degradation.

All the above mentioned solutions have the ambition to predict exactly the traffic or resources to be allocated. For this reason are based on the minimization of symmetric cost functions such as MSE. Unfortunately, there are random components that are not predictable and that leads to an unavoidable prediction error. Such a mistake leads to higher operational costs. For example, if the predicted traffic is higher than the real traffic, the resources will be over-sized and this will lead to an OP cost; in the opposite case less resources will be allocated and this will lead to a QoS degradation and to an UP cost characterized by the compensation due to the user.

Our research objective is to propose and evaluate a prediction-based allocation technique in which both UP and OP costs are takes into account.

A preliminary result on the advantages of the proposed SARIMA prediction technique is illustrated in [48]. The following contributions are added in this manuscript:

- an extensive description of the SARIMA prediction model with asymmetric loss function is reported;
- an innovative prediction algorithm based on an LSTM recurrent neural network with asymmetric cusp loss function is added;
- extensive numerical results are reported in which the operational costs of an NFV network with resource allocation based on SARIMA and LSTM are evaluated.
- an extension of ETSI NFV architecture with proposed traffic prediction and resource allocation modules is reported and described in Appendix A.

III. PROBLEM STATEMENT

The objective of the paper is to propose and evaluate a solution for the cloud and bandwidth resource allocation in NFV environments in which the traffic offered is not a-priory known but it is predicted according to a prediction technique aiming at minimizing the total operational cost. Two cost components are considered for a predicted SFC traffic: i) Cloud Resource Allocation Cost; ii) QoS Degradation Cost occurring when the traffic is incorrectly predicted, less resources are allocated and the network operator must pay a compensation cost to the user due to lost traffic.

We report the reference scenario in Fig. 1.a where four NFVI-PoPs interconnected by an EON are represented in a given traffic scenario. Five VNFIs are instantiated: NFVI-PoP #1 hosts VNFI #1 and VNFI #2, NFVI-PoP #2 hosts VNFI #3 and VNFI #4 and NFVI-PoP #3 hosts VNFI #5. Processing resources, represented by black rectangles, are allocated to the VNFIs. Three optical paths are also set up to interconnect the tuples VNFI #1 and VNFI #5, VNFI #2 and VNFI #3, VNFI #4 and VNFI #5. The interconnection is realized by allocating Frequency Slots (FS) whose number is depending on: i) the bandwidth to be allocated between the VNFIs; ii) the optical path length that determines the best modulation system according to the optical signal quality to be guaranteed [12].

In a dynamic traffic scenario, the cloud and bandwidth resources have to be re-allocated according to the current traffic conditions. The following operations can be performed: i) horizontal [1] or vertical scaling [26] of the cloud resources allocated to the VNFIs; ii) migrations of VNFIs towards a different NFVI-PoP [18]; iii) reconfigurations of optical paths by changing the routing and increasing/decreasing the number of allocated FSs [12]. We report an example of reconfiguration in Fig. 1.b in the case of a traffic increase between the VNFIs #1 and #5. For handling this increase the cloud resources allocated to the two VNFIs are increased by applying a vertical scaling technique that lead to increase their processing capacity of an amount represented with a grey rectangle in Fig. 1.b. Furthermore the optical path bandwidth is increased by allocating other two FSs in the network links on which the optical path is set up and represented with dotted blue lines in Fig. 1.b.

Reactive reconfiguration approaches are not suited in NFV environments especially due to the high time needed to reconfigure the cloud resources [34]. For this reason traffic prediction is needed to allocate in advance the cloud resources. Unfortunately the traffic cannot be predicted exactly and the prediction error may lead to resource over/under provisioning with a consequently increase in operational network cost.

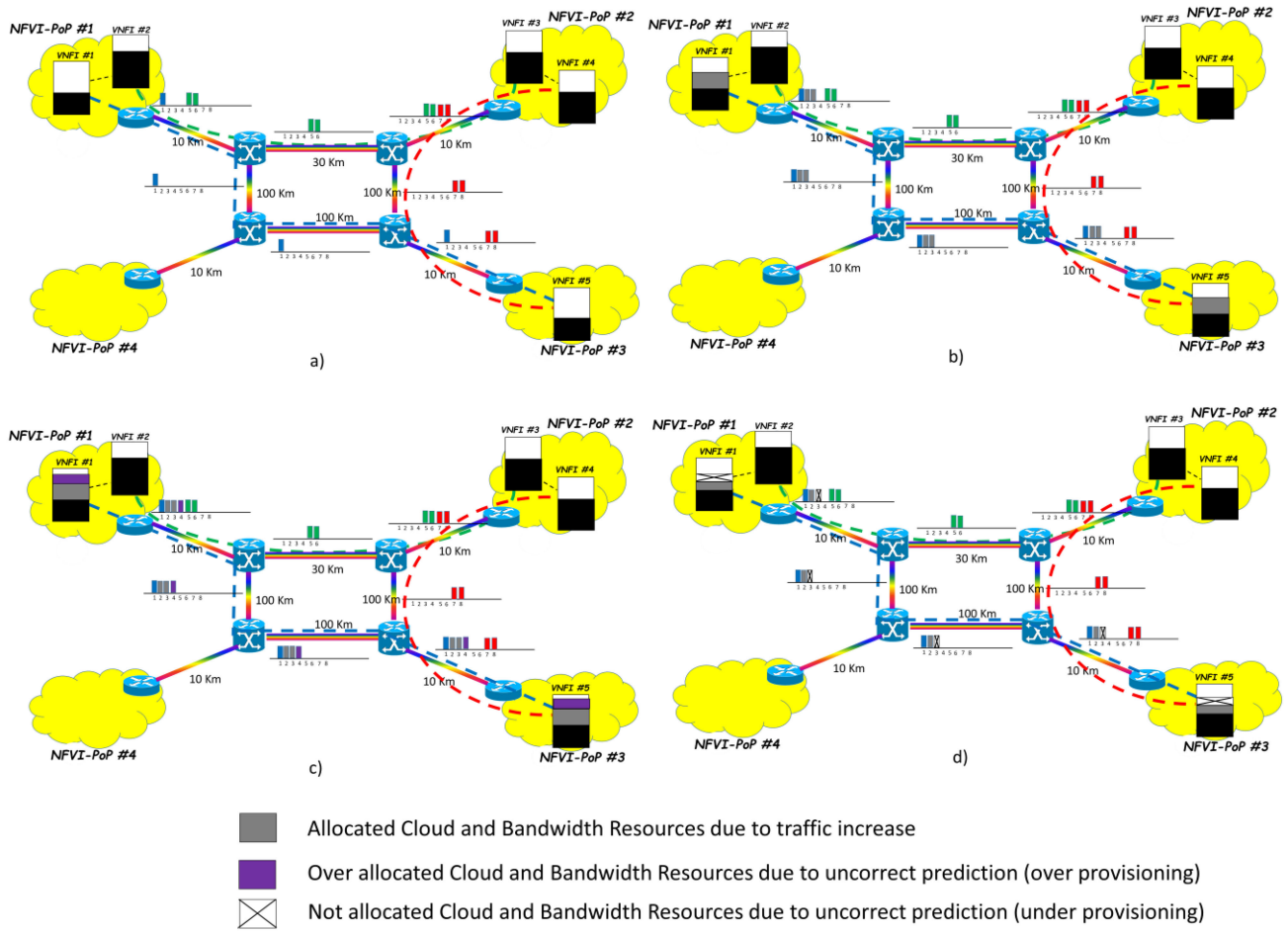


FIGURE 1. Cloud and bandwidth resource allocation in an NFV environment with four NFVI-PoPs (a); resource reconfiguration when a traffic increase between the VNFI #1 and VNFI #5 (b); resources over provisioning when a traffic prediction error is made (c); resources under provisioning when a traffic prediction error is made (d).

Over provisioning occurs when the predicted traffic is higher than the real one; in this case more cloud and bandwidth resources than needed are allocated; an example of over provisioning is illustrated in Fig. 1.c where the additional cloud and bandwidth resources are reported with violet rectangles; obviously the allocation of unnecessary resources leads to a cost increase.

Under provisioning occurs when the predicted traffic is lower than the real one; in this case less resources than needed are allocated as illustrated in Fig. 1.d where the lack of needed resource is represented with crossed rectangles; the under provisioning leads to QoS degradation due to the traffic amount which will inevitably be lost because of the lack of resources; that will determine a cost increase for the network operator due to the compensation cost to be paid to the user for the lost traffic.

From the shown example, we can observe that because the errors in predicting traffic are inevitable, the impact on cost increase is not only dependent on the absolute value of the error but positive and negative errors can differently impact on the cost increase depending on the values of the resource allocation and QoS degradation costs. In particular if the resource allocation costs are higher than the QoS degradation ones, the errors made by the algorithm should lead to predict lower traffic values than the real ones; conversely the algorithm should behave in the opposite way when the QoS degradation costs are higher than the resource allocation ones.

The prediction algorithms are based on the minimization of an error function referred to as loss function. Most of the solutions proposed in literature are based on symmetric loss functions (i.e. MSE, MAE) that are not able to optimize the total cost as previously explained. For this reason we propose solutions with asymmetric loss function and characterized by parameters whose setting depends on the resource allocation and QoS degradation costs. The setting of the parameters is based on the observation of past traffic values and allows for a total cost minimization.

Next we illustrate the cloud infrastructure, network and traffic models in Subsection III-A. The prediction framework is described in Subsection III-B.

A. CLOUD INFRASTRUCTURE, NETWORK AND TRAFFIC MODELS

We represent with the graph $\bar{G} = (\bar{V}, \bar{L})$ the NFVI-PoPs interconnected by the EON, where the set $\bar{L}$ denotes the optical links and the set $\bar{V}$ denotes the union of three sets: i) $\bar{V}_{NP}$ containing the NFVI-PoPs; ii) $\bar{V}_A$ containing the access nodes in which the traffic is originated/terminated; iii) $\bar{V}_S$ containing the optical switches.

The NFVI-PoPs are equipped with cloud resources characterized by processing cores. We denote with $N_{\bar{v}}$ the number of cores assigned to the NFVI-PoP $\bar{v} \in \bar{V}_{NP}$.

VNFIs, supported by VMs, are activated to support the execution of Service Functions (SFs) as Firewall (FW), Load Balancer (LB), Network Address Translation (NAT),... We assume vertical processing resource scaling where the processing cores assigned to the VNFIs can be changed over time according to the VNFI current load. In particular if F SFs are supported then F VNFI types can be instantiated. For the i-*th* ($i = 1, \ldots, F$) type VNFI we denote with:

- $C_i^{pr,max}$ (Gbps) ($i = 1, \ldots, F$): the maximum processing capacity that can be assigned to i-*th* type VNFI;
- n_i^c ($i = 1, \ldots, F$): number of cores assigned to i-*th* type VNFI when the maximum processing capacity is provided;
- $C_{i,k}^{pr} = \frac{k}{n_i^c} C_i^{pr,max}$ (Gbps) ($i = 1, \ldots, F, k = 1, \ldots, n_i^c$): the processing capacity assigned to i-*th* type VNFI when k cores are assigned to i-*th* type VNFI.

We also denote with $c_{\bar{v}}^{core}$ the core cost expressed in (\$/h) and characterizing the cost of renting one processing core for one hour in the NFVI-PoP $\bar{v} \in \bar{V}_{NP}$. We also introduce the average core cost c_{av}^{core} expressed by:

$$c_{av}^{core} = \frac{1}{|\bar{V}_{NP}|} \sum_{\bar{v} \in \bar{V}_{NP}} c_{\bar{v}}^{core} \tag{1}$$

The traffic demand is characterized by the SFCs whose bandwidth is variable over time. N SFCs are generated; the i-*th* SFC ($i = 1, \ldots, N$) is characterized by R_i SFs and we introduce the binary variable $s_i(j, p)$ ($i = 1, \ldots, N; j = 1, \ldots, R_i; p = 1, \ldots, F$) assuming the value 1 if the j-*th* SF executed is of p-*th* type.

We characterize the SFCs with the average bandwidth offered in Time Intervals (TI) of duration T_s. In particular we denote with $b_j(i)$ the offered average bandwidth of the i-*th* SFC ($i = 1, \ldots, N$) in the j-*th* ($j = 1, 2, \cdots$) TI.

The cloud resource allocation cost for the $i - th$ SFC is denoted with $C_{RA,i}$; it is expressed in (\$/Gb) and it characterizes the average cost for the cloud resource allocation needed to the SFC bandwidth of one Gb. This cost can be easily expressed as:

$$C_{RA,i} = \sum_{j=1}^{R_i} \sum_{p=1}^{F} c_{av}^{core} \frac{n_p^c}{C_p^{pr,max}} s_i(j, p) \quad \forall i \in [1..N] \tag{2}$$

Expression (2) can be justified as follows:

- the i-*th* SFC allocation cost $C_{RA,i}$ is given by the sum of the allocation costs of each of the R_i SFs composing the SFC;
- the support of one Gb of traffic for p-*th* type SF requires the allocation of $\frac{n_p^c}{C_p^{pr,max}}$ cores each of which has an average cost of c_{av}^{core};
- the cost evaluated in each term of expression (2) has to be included if the j-*th* SF of the i-*th* SFC is of p-*th* type that is if $s_i(j, p)$ equals 1.

Finally we denote the QoS degradation cost with C_{QoS}; it is expressed in (\$/Gb) and characterizes the cost to be paid by the network operator when resources are not allocated for a SFC bandwidth of one Gb.

To limit their number, the VNFIs are shared among the SFCs. The VNFIs are instantiated and connected with optical paths. The SFCs are routed through the VNFIs so as to execute the SFs of each SFC. The VNFIs and their interconnection can be represented by the graph $G = (V, L)$ where the set of nodes V characterizes the VNFIs and the set L contains elements representing the logical links interconnecting the VNFIs.

B. CLOUD AND BANDWIDTH PROVISIONING FRAMEWORK WITH TRAFFIC PREDICTION

A resource allocation algorithm has the objective to determine an embedding $\Gamma(\bar{G}, G)$ of the VNFI graph $G = (V, L)$ into the physical graph $\bar{G} = (\bar{V}, \bar{L})$ by determining: i) in which NFVI-PoP any VNFI is executed; ii) the cloud (processing) resources to be assigned to the VNFIs; iii) in which optical path any logical link has to be routed; iv) the number of FSs to be allocated on the chosen optical path. When traffic variations over time occur, cloud and bandwidth reconfigurations are needed to reduce the costs. Some reconfiguration techniques have been proposed. For instance the solution proposed in [12] leverages the following techniques: i) migration of VNFIs towards lowest cost NFVI-PoPs; ii) vertical cloud resource scaling by increasing/decreasing the number of cores allocated to the VNFIs. To apply the techniques, embedding changes of the VNFI graph $G = (V, L)$ into the physical graph $\bar{G} = (\bar{V}, \bar{L})$ are needed and depending on the the processing capacities $f_v^{(j)}$ ($j = 1, 2, \cdots$) requested by the nodes $v \in V$ and the requested bandwidth $f_e^{(j)}$ ($j = 1, 2, \cdots$) by the links $e \in L$ of the VNFI graph in the j-*th* TI ($j = 1, 2, \cdots$). The processing capacities and the link bandwidths are depending on the offered SFC bandwidths and for this reason they are not a-priori known. We propose and investigate a reconfiguration solution based on the prediction of the offered SFC bandwidths. Because it is not possible to determine the traffic exactly, we propose a solution that underestimates or overestimates the traffic according to the values of the resource allocation and QoS degradation costs.

The main steps performed by the framework for the cloud and bandwidth resource provisioning are illustrated in Algorithm 1. The inputs are: the physical graph $\bar{G} = (\bar{V}, \bar{L})$,

Algorithm 1 Cloud and Bandwidth Resource Provision Framework With Traffic Prediction

1: **Input:**
Cloud Infrastructure and Network Graph $\bar{G} = (\bar{V}, \bar{L})$
SFC bandwidths: $b_j(i)$ $(i = 1, \ldots, N, j = 0, \ldots, n)$
VNFI graph: $G = (V, L)$
/*Multi-step ahead SFC bandwidth Prediction*/

2: **Predict** the SFC bandwidths $\widehat{b}_{n+j}(i)$ $(i = 1, \ldots, N, j = 1, \ldots, h)$ with asymmetric loss function
/*Cloud and Bandwidth Resource Reconfiguration*/

3: **Evaluate** the estimated bandwidths $\widehat{f}_e^{(n+j)}$ and the estimated processing capacities $\widehat{f}_v^{(n+j)}$ of the links $e \in L$ and nodes $v \in V$ of the VNFI graph in the TIs $n+1, \ldots, n+h$

4: **Reconfigure** the bandwidth and the cloud resources by applying the NORR/ONRCA algorithms [12] and evaluating the embeddings $\Gamma_{n+j}(\bar{G}, G)$ $(j = 1, \ldots, h)$ in the ITs $n+1, \ldots, n+h$

5: **Output:** $\Gamma_{n+j}(\bar{G}, G)$ $(j = 1, \ldots, h)$

the SFC bandwidths $b_j(i)$ $(i = 1, \ldots, N, j = 0, \ldots, n)$ known up to TI n and the VNFI graph $G = (V, L)$. Next a multi-step ahead prediction of the SFC offered is performed in step 2 by predicting the next h SFC bandwidth values $\widehat{b}_{n+j}(i)$ $(i = 1, \ldots, N, j = 1, \ldots, h)$. That allows for the evaluation in step 3 of an estimate of the link bandwidths $\widehat{f}_e^{(n+j)}$ and the nodes processing capacities $\widehat{f}_v^{(n+j)}$ of the VNFI graph in the TIs $n+1, \ldots, n+h$. The knowledge of these estimated values and the application of cloud and bandwidth resource reconfiguration algorithms allow in step 4 for the determination of h new embeddings $\Gamma_{n+j}(\bar{G}, G)$ $(j = 1, \ldots, h)$ to be applied in the TIs $n+1, \ldots, n+h$. We apply the reconfiguration algorithms proposed in [12] referred to as NFV/Optical Resource Reconfiguration (NORR) and Optical Network Reconfiguration Costs Aware (ONRCA). Finally the framework returns the evaluated embeddings $\Gamma_{n+j}(\bar{G}, G)$ $(j = 1, \ldots, h)$.

IV. SFC BANDWIDTH FORECASTING BASED ON SARIMA AND ASYMMETRIC LINEX COST FUNCTION

To simplify the notations, next we drop the i parameter characterizing the offered SFCs; we will explain the traffic forecasting procedures for a generic SFC.

We propose an SFC bandwidth forecasting procedure based on: i) characterizing the SFC bandwidth values $\{b_j \; j = 1, \ldots, n\}$ as a time series; ii) modeling the time series with a SARIMA process; iii) forecasting the observed bandwidth values $\hat{b}_{n+j}$ at time $n+j$ $(j = 1, \ldots, h)$ of a SARIMA in the case in which an asymmetric cost function of the error $b_{n+j} - \hat{b}_{n+j}$ $(j = 1, \ldots, h)$ is minimized. The main steps of the proposed methodology are illustrated in Fig. 2 and explained in the next Subsections IV-A-IV-D. The following steps are performed: i) in the first step, illustrated in Subsection IV-A, trend and seasonality, due to the traffic

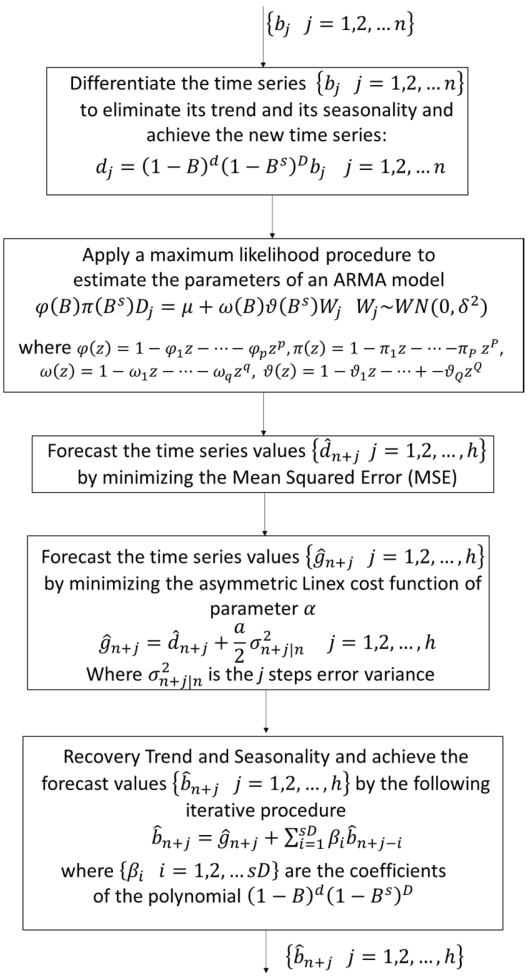


FIGURE 2. Main Steps of the SFC Bandwidth Forecasting Procedure.

cycle-stationarity, are removed from the times series and a stationary time series is achieved; ii) in the second step, illustrated in Subsection IV-B, the stationary time series is modeled as an Autoregressive Moving Average (ARMA) process by estimating the ARMA model parameters with a maximum likelihood procedure; iii) in the third step, illustrated in Subsection IV-C, the time series forecasting is performed by minimizing the conditioned expectation of the asymmetric cost function of the forecasting error; iv) in the four step, illustrated in Subsection IV-D, the trend and seasonality are recovered.

Finally we illustrate in Subsection IV-E how to set the parameter of the asymmetric cost function so as to achieve bandwidth forecast values allowing for the minimization of the cloud resource allocation and QoS costs.

A. TREND AND SEASONALITY ELIMINATION PROCEDURE

The traffic is non-stationary [26] and has trend and seasonality components. For instance it is well known that the traffic has a daily seasonality component. The trend and seasonality components can be eliminated by differentiating the time

series $\{b_j \ j = 1, \ldots, n\}$. To perform this differentiation we introduce the operator B^k that delays k times the values of the time series, that is $B^k b_j = b_{j-k}$. The differentiated time series $\{d_j \ j = 1, \ldots, n\}$ can be expressed as follows [49]:

$$d_j = (1 - B)^d (1 - B^s)^D b_j \quad j = 1, \ldots, n \tag{3}$$

where s is the seasonal parameter that may be chosen equal to 24 if a typical daily traffic profile is considered, d and D are the number of times in which the time series $\{b_j \ j = 1, \ldots, n\}$ is differentiated to eliminate the trend and the seasonality respectively. If the parameters s, d and D are appropriately chosen, the time series $\{d_j \ j = 1, \ldots, n\}$ can be made stationary [49].

B. PROCEDURE OF ARMA PARAMETERS IDENTIFICATION AND ESTIMATION

The second step of the proposed methodology consists in modeling the stationary time series $\{d_j \ j = 1, \ldots, n\}$ with an Autoregressive Moving Average (ARMA) process $\{D_j \ j = 1, \ldots, \}$, that is expressed by the following expression [49]:

$$\varphi(B)\pi(B^s)D_j = \mu + \omega(B)\vartheta(B^s)W_j \quad W_j \sim WN(0, \delta^2) \tag{4}$$

wherein:

- $\varphi(B) = 1 - \varphi_1 z - \cdots - \varphi_p z^p$ and $\omega(B) = 1 - \omega_1 z - \cdots - \omega_q z^q$ are the autoregressive and moving average components respectively allowing for the characterization of correlation between the values of the time series belonging to different seasons;
- $\pi(B) = 1 - \pi_1 z - \cdots - \pi_P z^P$ and $\vartheta(B) = 1 - \vartheta_1 z - \cdots - \vartheta_Q z^Q$ are the autoregressive and moving average components respectively allowing for the characterization of correlation between the values of the time series belonging to a same season;
- μ is a parameter linked to the average value of the time series and it equals zero in the case of zero average time series;
- $\{W_j \ j = 1, \ldots, \}$ is a white noise with zero average and standard deviation δ.

The identification of the ARMA model involves the choice of the following parameters:

- $p_{max}, P_{max}, q_{max}, Q_{max}$: they are the maximum values of the parameters p, P, q, Q and are determined from the observation of the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the original time series $\{d_j \ j = 1, \ldots, n\}$ [49];
- $p_{opt}, P_{opt}, q_{opt}, Q_{opt}, \mu_{opt}, \delta_{opt}, (\varphi_j^{opt} \ j = 1, \ldots, p_{opt}), (\omega_j^{opt} \ j = 1, \ldots, P_{opt}), (\pi_j^{opt} \ j = 1, \ldots, q_{opt}), (\vartheta_j^{opt} \ j = 1, \ldots, Q_{opt})$: they are the final values of the ARMA model and are determined by applying Maximum Likelihood (ML) procedures for $(p, P, q, Q) \in [1..p_{max}, 1..P_{max}, 1..q_{max}, 1..Q_{max}]$ and by choosing the orders that minimize the Alaike Information Criterion (AICC) statistic [49].

In the application of the ML procedure the values $\{d_j \ j = 1, \ldots, n\}$ of the original time series are used; at the end we also check that the residuals, given by the difference between the original and ARMA values, are uncorrelated [49].

C. PROCEDURE OF TIME SERIES FORECASTING

All of the prediction-based resource allocation algorithms in NFV environments aims at exactly forecasting either the traffic [34] or the resources [28] to be allocated. They are based on the minimization of the conditioned expectation of a symmetric cost function of the forecast error $d_{n+j} - \hat{d}_{n+j}$ $(j = 1, 2, \ldots, h)$. The classical case is the minimization of the conditioned expectation of MSE, that is $E_n[(D_{n+j} - \hat{d}_{n+j})^2]$ where the symbol $E_n[*]$ is the expectation conditioned to the knowledge of the values $\{d_j \ j = 1, \ldots, n\}$.

The choice of symmetric cost functions leads to equally weight positive and negative errors. Conversely being aware that an exact traffic prediction is not possible, our objective is to make mistakes where it is more convenient according to the cloud resource allocation the QoS degradation costs. For this reason we consider asymmetric cost functions and because of its simplicity we choose the LINEX function [50]. That leads to the minimization of the conditioned expectation $E_n[L(D_{n+j} - \hat{d}_{n+j})^2]$ where the LINEX function $L(x)$ is defined as follows:

$$L(x) = \exp^{ax} - ax - 1 \tag{5}$$

In particular notice that: i) for $a > 0$ $(a < 0)$ the error $d_{n+j} - \hat{d}_{n+j}$ $(j = 1, \ldots, h)$ has higher cost when it is positive (negative); ii) for $|a|$ increasing the difference in cost of positive and negative errors grows.

It is possible to prove [50] that the LINEX optimal predictor $\hat{g}_{n+j}$ $(j = 1, \ldots, h)$ has the following expression:

$$\hat{g}_{n+j} = \hat{d}_{n+j} + \frac{a}{2}\sigma^2_{n+j|n} \quad j = 1, \ldots, h \tag{6}$$

wherein $\hat{d}_{n+j}$ $(j = 1, \ldots, h)$ is the MSE optimal predictor and $\sigma^2_{n+j|n}$ $(j = 1, \ldots, h)$ is the conditioned error variance $E_n[(D_{n+j} - \hat{d}_{n+j})^2]$ whose the iterative evaluation is reported in Appendix B.

From Figure 2 we can notice how the the forecasting values $\hat{g}_{n+j}$ $(j = 1, \ldots, h)$ in the asymmetric cost function case are achieved by evaluating the forecasting values $\hat{d}_{n+j}$ $(j = 1, \ldots, h)$ that minimizes the conditioned expectation of MSE and then by applying the expression (6).

D. TREND AND SEASONALITY RECOVERY PROCEDURE

The final step of the proposed methodology consists in recovering the trend and the seasonality to the predicted time series $\hat{d}_{n+j}$ $(j = 1, \ldots, h)$. From the initial transformation of expression (3), we can obtain the following expression [49]:

$$b_{n+j} = d_{n+j} + \sum_{i=1}^{d+sD} \beta_i b_i \quad j = 1, \ldots, h \tag{7}$$

where the coefficients β_i $(i = 1, \ldots, d + sD)$ are the coefficients of the polynomial $(1 - B)^d(1 - B^s)^D$ [49].

From eq. (7) we can write the following expression of the predicted values $\hat{b}_{n+j}$ $(j = 1, \ldots, h)$:

$$\hat{b}_{n+j} = \hat{g}_{n+j} + \sum_{i=1}^{d+sD} \beta_i \hat{b}_{n+j-i} \quad j = 1, \ldots, h \tag{8}$$

The values $\hat{b}_{n+j}$ $(j = 1, \ldots, h)$ can be recursively evaluated from expression (8) with $j = 1, 2, \ldots, h$ and by taking into account that $\hat{b}_{n+j-i} = b_{n+j-i}$ for $(i = j, j+1, \ldots, d+Ds)$.

E. SETTING OF THE LINEX COST FUNCTION PARAMETER

We need to set the parameter a of the LINEX function. The value of the parameter determines the shape of the asymmetric cost function and has to be chosen so as to optimize the sum of the cloud resource allocation and the QoS degradation costs. To evaluate the optimal value a_{opt}, instead of using all of the time series $\{b_j \ j = 1, \ldots, n\}$ to identify ARMA process, we split the time series in two sets: the first one is used to estimate the ARMA parameters and the second one is used to evaluate the parameter a_{opt}. The pseudo-code of the procedure for the setting of the parameter a is illustrated in Algorithm 2. We assume to choose the parameter a_{opt} in the interval $[a_{min}, a_{max}]$. The procedure chooses (line 2) the index $1 < p < n$ so that the time series $\{b_j \ j = 1, \ldots, p\}$ is used for the ARMA parameters estimation (line 3), while the time series $\{b_j \ j = p+1, \ldots, n\}$ is used to evaluate the parameter value a_{opt} (line 4). Next for each value of a, the sum $C(a)$ of the cloud resource allocation and QoS degradation costs (lines 5-10) is evaluated. Finally the value a_{opt} minimizing $C(a)$ and the optimum cost C_{opt} are determined and returned as output (line 13).

Algorithm 2 Procedure for the Setting of the Parameter a of the LINEX Function

1: **Input:**
Input: $a_{min}, a_{max}, \{b_j \ j = 1, \ldots, n\}$
***/*Splitting of the set** $\{b_j \ j = 1, \ldots, n\}$*/*
2: **Choose** the index p so that the time series $\{b_j \ j = 1, \ldots, p\}$ is used for ARMA model identification and the time series $\{b_j \ j = p+1, \ldots, n\}$ is used to evaluate the optimal value a_{opt} of the LINEX function
/*ARMA model Identification*/
3: **Identify** the ARMA model parameters by using the time series $\{b_j \ j = 1, \ldots, p\}$
/*Time Series Forecasting*/
4: **Evaluate** the predicted values $\hat{b}_j$ $(j = p+1, \ldots, n)$
***/*Evaluation of the parameter value** a_{opt}*/*
5: **Initialize** $a_{opt} = \infty$, $C_{opt} = \infty$
6: **for** $a \in [a_{min}..a_{max}]$ **do**
7: **Evaluate** $C(a) = \sum_{j=p+1}^{n} (I(b_j - \hat{b}_j) C_{QoS} + I(\hat{b}_j - b_j) C_{RA})$
***/*$I(x)$ is the indicator function that is** $I(x) = 1$ ***for** $x > 0$ ***and** $I(x) = 0$ ***for** $x < 0$*/*
8: **if** $C(a) < C_{opt}$ **then**
9: $a_{opt} = a$
10: $C_{opt} = C(a)$
11: **end if**
12: **end for**
13: **Output:** a_{opt}, C_{opt}

V. ASYMMETRIC LOSS FUNCTION-BASED LSTM PREDICTION ALGORITHM

The L unfolded stages version of the LSTM prediction algorithm is illustrated in Fig. 3.a and consists of the following two layers:

- the LSTM prediction layer: it performs the time series prediction by providing the storage of the internal states; we consider the case of a single layer composed by L LSTM Cell Blocks (LCB) referred to as LCB_j $(j = n - L + 1, \ldots, n)$;
- the feed forward network layer: it evaluates from the output of the last LSTM layer the h steps ahead predicted bandwidth values $\hat{b}_{n+j}$ $(j = 1, \ldots, h)$ stored in the vector $\hat{b}_{n,h}$.

The SFC bandwidth predictions are performed by the LSTM layer which has as inputs the SFC bandwidth values b_j $(j = n - L + 1, \ldots, n)$. The output vector h_n is processed by a feed forward neural network which provides to evaluating the vector $\hat{b}_{n,h}$ of predicted SFC bandwidth values.

In the LSTM layer the state variable s_j $(j = n - L + 1, \ldots, n)$ is also updated. In the LSTM Cell Block LCB_j, shown in Fig. 3.b, the state variable s_j in the j-th TI depends on the following variables: i) the SFC bandwidth value b_j; ii) the output h_{j-1} in the $(j-1)$-th TI; iii) the state variable s_{j-1} in the $(j-1)$-st TI.

The LSTM innovative idea is to introduce the forget and input gates that decide which components of the state vector has to be deleted (forget gate) and preserved (input gate). This operation is learned through the training of the weight matrices W_{fh}, W_{fx}, W_{gh}, W_{gx} and biasing vectors d_f, d_g. The output gate is also introduced in LSTM neural networks. It is characterized by the matrices W_{oh} and W_{ox}, the biasing vector d_o and controls what information encoded in the state variable is sent to the output h_j of the LSTM Cell Block LCB_j.

If W_{ih}, W_{ix} and d_i denote the weight matrices and biasing vector for the input, we can write the following expressions for the evaluation of the state variable s_j and the output h_j of the LSTM Cell Block LCB_j:

$$\begin{aligned} s_j = {} & s_{j-1} \times \sigma(W_{fh} h_{j-1} + W_{fx} b_j + d_f) \\ & + \sigma(W_{ih} b_j + W_{ix} b_j + d_i) \times \varphi(W_{gh} h_{j-1} \\ & + W_{gx} b_j + d_g) \end{aligned} \tag{9}$$

$$h_j = \varphi(s_j) \times \sigma(W_{oh} h_{j-1} + W_{ox} b_j + d_o) \tag{10}$$

where $\sigma(*)$ represents the sigmoid activation function, while $\varphi(*)$ represents the *tanh* activation function.

All of the LSTM-based traffic prediction algorithms proposed in literature, [28] are based on the minimization of a symmetric cost function of the errors $e_{n+j} = b_{n+j} - \hat{b}_{n+j}$

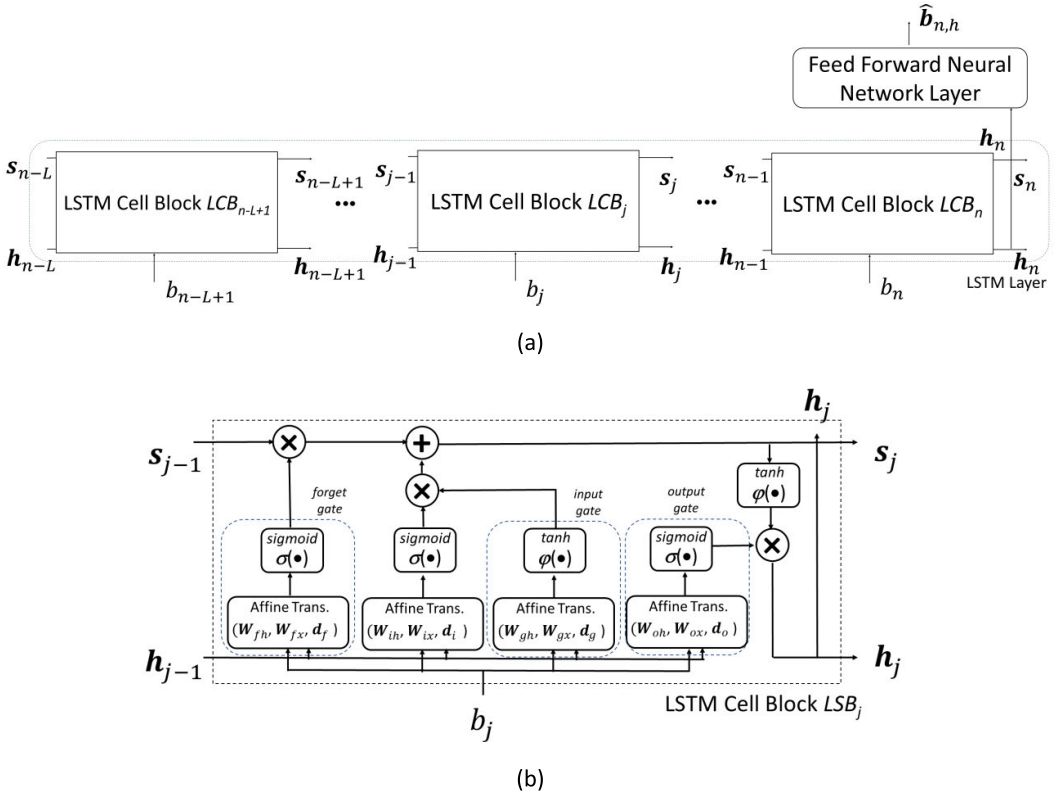


FIGURE 3. LSTM Prediction Framework (a). LSTM Cell Block LCB_j (b).

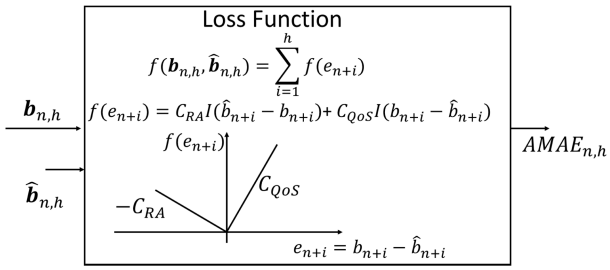


FIGURE 4. Definition of the Asymmetric Loss Function.

$(j = 1, \ldots, h)$. We consider asymmetric cost functions and because of its simplicity we choose a cusp linear loss function as represented in Fig. 4 where the slopes are dependent on the resources allocation cost C_{RA} and QoS degradation cost C_{QoS} both defined in (\$/Gb). As reported in Fig. 4 the training process minimizes the Asymmetric Mean Absolute Error $AMAE_{n,h}$ expressed by:

$$AMAE_{n,h} = \frac{1}{h} \sum_{j=1}^{h} (C_{RA} I(\hat{b}_{n+j} - b_{n+j}) + C_{QoS} I(b_{n+j} - \hat{b}_{n+j})) \quad (11)$$

where $I(x)$ is the indicator function that is $I(x) = 1$ for $x > 0$ and $I(x) = 0$ for $x < 0$.

VI. NUMERICAL RESULTS

We will evaluate the effectiveness of the asymmetric cost function-based SARIMA and LSTM forecasting model in predicting the requested SFC bandwidth when both the cloud resource allocation and QoS degradation costs are considered. The SARIMA and LSTM forecasting technique will be applied in a real scenario to evaluate the operation cost of an NFV network and compare it to the case in which an MSE traditional forecasting technique is applied.

We describe the simulation environment in Subsection VI-A. The application of the asymmetric cost function-based SARIMA forecasting technique to real traffic data is illustrated in Subsection VI-B. Finally we will show in Subsection VI-C the effectiveness of the proposed SARIMA and LSTM forecasting solutions when it is applied to allocate the resources in an NFV network whose NFVI-PoPs are interconnected by an EON.

A. SIMULATION ENVIRONMENT

The numerical results will be provided for the values of the simulation parameters reported in Table 1. SFCs composed by the sequence of the followings four SFs are considered: Firewall (FW), Intrusion Detection System (IDS), Network Address Translator (NAT) and Proxy.

The effectiveness of the proposed prediction techniques are evaluated for the ABILENE network shown in Fig. 5

TABLE 1. Input parameters.

Parameter	Description	Value
B_{FS}	Bandwidth of one Frequency Slot	6.25 GHz
L_{BPSK}	Maximum Length Path for BPSK	3000 km
L_{QPSK}	Maximum Length Path for QPSK	1500 km
L_{8QAM}	Maximum Length Path for 8QAM	750 km
L_{16QAM}	Maximum Length Path for 16QAM	375 km
S	Number of FSs	50, 600
N_{NP}	Number of NFVI-PoPs	4
$N_{\bar{v}}$	Number of cores in any NFVI-PoP	3072
c_{av}^{core}	Average Core Cost	1 \$/h
T_s	Time duration of a TI	6 1 h
C_{QoS}	QoS degradation cost (\$ to be paid for one lost traffic Gb)	{0.0025,0.025,0.25}
$w = \frac{C_{RA}}{C_{QoS}}$	Ratio of resource allocation cost C_{RA} to the QoS degradation cost C_{QoS}	{0.1,1,10}

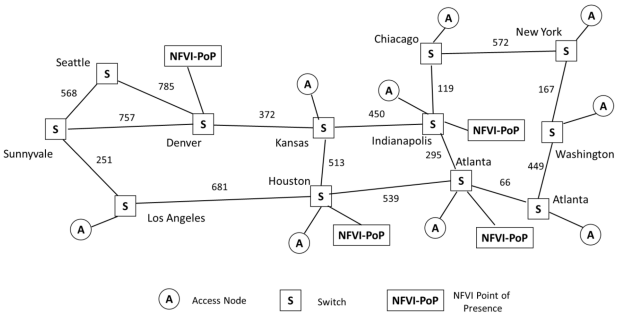


FIGURE 5. ABILENE Network Topology. The distances are expressed in km.

composed by 12 optical switches and 15 links. The network is equipped with four NFVI-PoPs located in the cities of Atlanta, Denver, Houston and Indianapolis. Each NFVI-PoP is equipped with 3072 cores whose average cost per hour is $c_{av}^{core} = 1$\$/h. We assume that one SFC is established for each tuple of access nodes reported in Fig. 5. We assume as SFC bandwidth values the real ones reported in [51] for the ABILENE network. In particular we consider the traffic values measured at hourly intervals. These values are used to forecast the future traffic values according to the procedure illustrated in Sections IV and V. The SFs are supported by four types of VNFIs whose characteristics, that is maximum processing capacity and the number of allocated cores, are reported in Table 2.

Each optical link is organized in 600 FSs each one with a bandwidth occupancy of 6.25 GHz. The connection between VNFIs is supported by optical paths with a choice of one of the following modulation systems: Binary Phase Shift Keying (BPSK), Quadrature Phase Shift Keying (QPSK), 8 Quadrature Amplitude Modulation (8QAM) and 16 Quadrature Amplitude Modulation (16QAM). The maximum optical

TABLE 2. Maximum processing capacity and allocated number of cores for the software modules implementing FW, IDS, NAT and Proxy.

VNFI Software	Maximum Processing Capacity ($C^{pr,max}$)	Number of cores allocated (n^c)
FW	30 Gbps	130
IDS	30 Gbps	400
NAT	30 Gbps	70
Proxy	30 Gbps	130

path length is 3000 km, 1500 km, 750 km and 375 km for BPSK, QPSK, 8QAM and 16QAM respectively [12]. The bandwidth efficiency factor is 1, 2,3 and 4 for BPSK, QPSK, 8QAM and 16QAM respectively [12].

B. SARIMA BANDWIDTH FORECASTING OF A SINGLE SFC BY MINIMIZING AN ASYMMETRIC COST FUNCTION

We evaluate the proposed forecasting technique for the time series reported in Fig. 6 reporting the hourly bandwidth values requested by the SFC instantiated between the nodes Chicago and Indianapolis. The time series is composed by 480 traffic values measured in the weekdays from May 31st 2004 to June 27th 2004 [51]. We have organized the time series into three sets: i) the first 240 values are used for the parameters estimation of the SARIMA model; ii) the next 120 values are used to evaluate the parameter a_{opt} of the LINEX function as illustrated in Subsection IV-E; iii) the last 120 values are used for the test phase in which the real and one-step predicted values are compared.

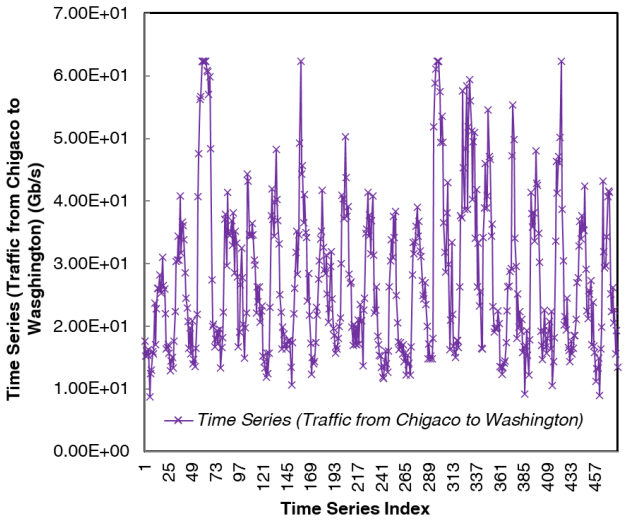


FIGURE 6. Time Series of the hourly bandwidth values requested by the SFC instantiated between the nodes Chicago and Indianapolis. The time series is composed by 480 traffic values measured in the weekdays from May 31st 2004 to June 27th 2004 [51].

The choice of the core costs and processing capacities of Tables 1 and 2 leads to a cloud resource allocation cost $C_{RA} = 0.025$\$/Gb for the SFC considered according to the expression (2).

We carry out the comparison for the following values of the QoS degradation cost:

- C_{QoS} = 0.0025$/Gb; that corresponds to the case in which the OP cost is higher than the UP one;
- C_{QoS} = 0.025$/Gb; that corresponds to the case in which the OP and UP costs are equal;
- C_{QoS} = 0.25$/Gb; that corresponds to the case in which the UP cost is higher than the OP one.

We also introduce the parameter $w = \frac{C_{RA}}{C_{QoS}}$; its value equals 10, 1 and 0.1 when C_{QoS} equals 0.0025, 0.025 and 0.25 respectively.

By applying the procedure illustrated in Subsection IV-B we have estimated the best parameters of the SARIMA model; this study has led to the choice of the following parameter values: i) the value of the parameter s has been chosen equal to 24 due to the traffic daily periodicity; ii) both the differentiation parameters d and D for the trend and seasonality elimination have been chosen equal to 1; iii) the maximization of the likelihood function for the ARMA model illustrated in Subsection IV-B has led to the choice of the following parameter values: $p = 16$, $q = 8$, $P = 1$, $Q = 1$.

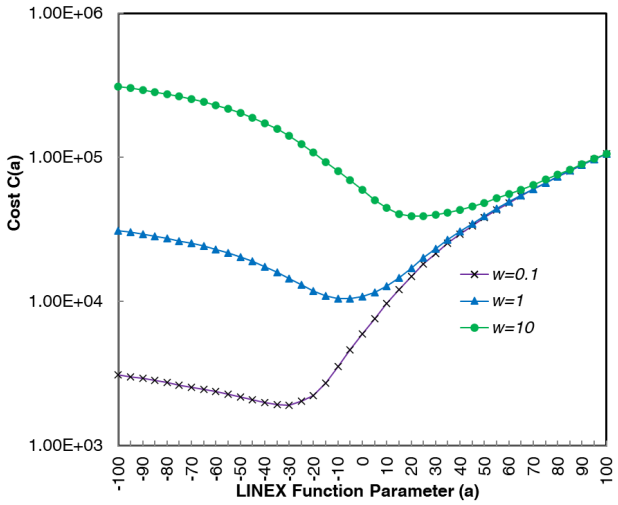


FIGURE 7. Cost function $C(a)$ as a function of the LINEX parameter a for w equal to 0.1, 1 and 10.

To determine the optimal parameter a_{opt} of the LINEX function we evaluate, considering the 120 values of the times series of indexes from 241 to 360, the cost function $C(a)$ introduced in Subsection IV-E for values of a in the range $[-100,100]$; the parameter value a_{opt} is determined by choosing the value of a minimizing $C(a)$. In particular we report in Fig 7 the function $C(a)$ for a in the range $[-100,100]$ and for values of the parameter w equal to 0.1, 1 and 10. The minimization operation leads to choose for a_{opt} the values -25, -2 and 20 for w equal to 0.1, 1 and 10 respectively. We can remark from Fig. 7 that:

- when $w = 0.1$ and consequently the OP cost is lower than the UP one, the procedure for the choice of a_{opt} behaves correctly by determining a negative value a_{opt} so that the LINEX function expressed by expression (5) provides a lower cost when the real traffic is higher than the predicted one and a higher cost in the opposite case;
- when $w = 10$ and consequently the OP cost is higher than the UP one, the value a_{opt} is positive and the LINEX function gives more weight to errors in which the real traffic is higher than the predicted one;
- when $w = 1$ and consequently the cloud resources allocation and QoS degradation costs C_{RA} and C_{QoS} are equal, the parameter a_{opt} is near to zero and provide a balanced cost function.

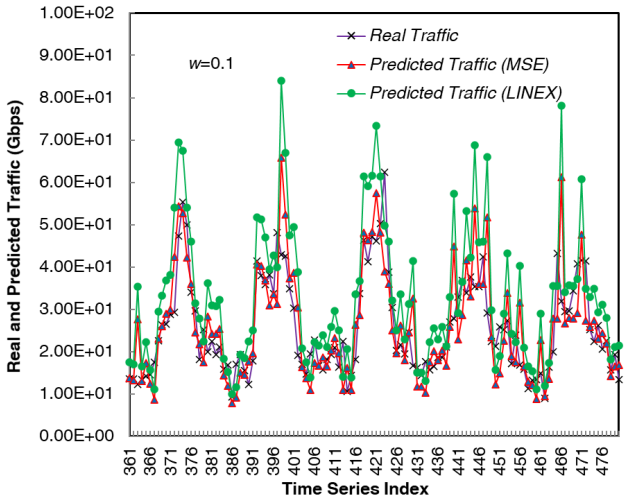


FIGURE 8. Comparison of the Real, MSE and LINEX prediction values for $w = 0.1$.

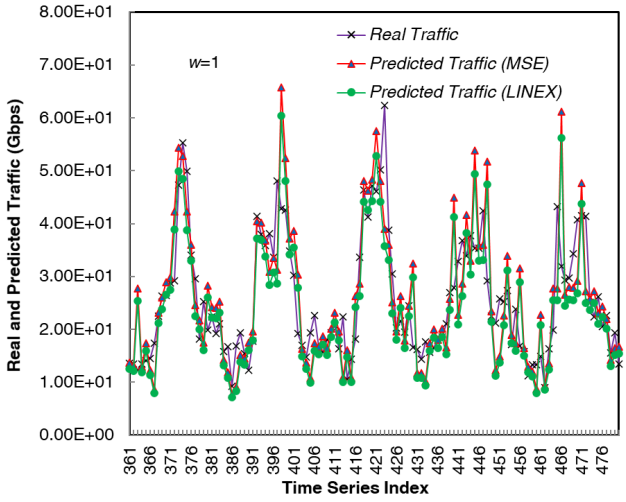


FIGURE 9. Comparison of the Real, MSE and LINEX prediction values for $w = 1$.

The comparison between real and predicted values of the time series from index 361 to 480 is reported in Figs 8-10 for w equal to 0.1, 1 and 10 respectively. We report the prediction values when the MSE and a LINEX cost functions

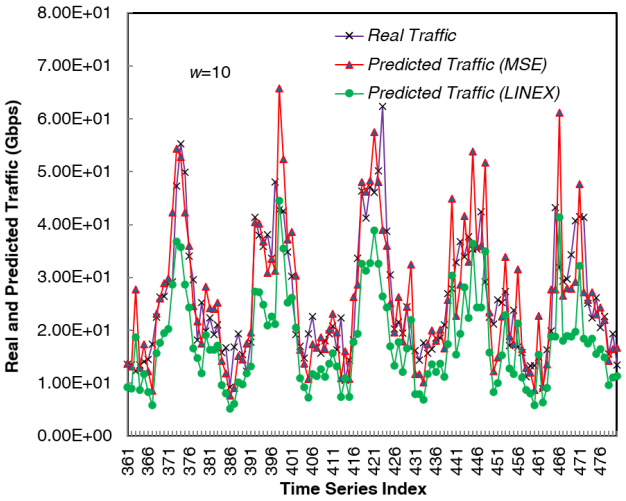


FIGURE 10. Comparison of the Real, MSE and LINEX prediction values for $w = 10$.

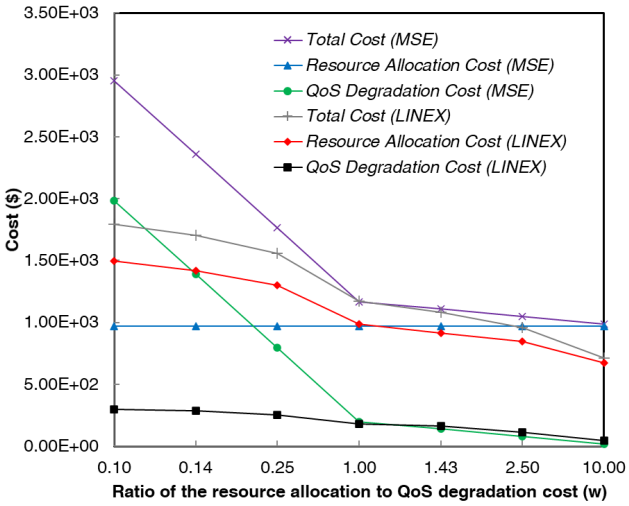


FIGURE 11. Cost in allocating resources for the NFV network of Fig. 5 when w varies from 0.1 to 10. The total, resource allocation and QoS degradation costs are reported when the allocation algorithms use MSE and LINEX predicted SFC bandwidth values.

are minimized. From Figs 8-10 we can remark that the MSE predictions are very near to the real time series values but they do not allow us to reach the goal of over-estimating or under-estimating the time series values according to the values of OP and UP costs; conversely the LINEX predictions allows for a correct operation mode by overestimating (Fig. 8) and underestimating (Fig. 10) when w equal 0.1 and 10 respectively. Finally you can notice how the MSE and LINEX predictions methodology behave similarly (Fig. 9) in the case of $w = 1$ that is when the OP and UP costs are equal and a balanced cost function is the best choice.

C. QoS DEGRADATION AND CLOUD RESOURCE ALLOCATION COST EVALUATION IN NFV NETWORK ENVIRONMENT

We compare the operation cost of the NFV network reported in Fig. 5 in the case of an SARIMA traffic prediction with minimization of MSE and the LINEX function respectively. The cost values have been achieved as follows:

- the real traffic values from May 31st 2004 to June 13th 2004 [51] have been used to evaluate the parameter values of the SARIMA model according to the procedures illustrated in Subsections IV-A-IV-D;
- the real traffic values from June 14th 2004 to June 20th 2004 [51] have been used to evaluate the optimal parameter a_{opt} of the LINEX function according to the procedure illustrated in Subsection IV-E;
- the total cost has been evaluated for the period from June 21th 2004 to June 27th 2004 when the optical bandwidth and cloud resources are allocated and reconfigured on the basis of the predicted traffic values and by applying the reconfiguration algorithms described in [12].

We report the cost in Fig. 11 in the cases in which the SARIMA traffic predictions are performed with the minimization of the MSE and LINEX function. Three cost components are reported as a function of the parameter w: the total cost, the cloud resource allocation cost and the QoS degradation cost. From the results reported in Fig. 11 we can remark that:

- the proposed forecasting solution based on the asymmetric cost function allows for total costs lower than or equal to the one of the MSE-based forecasting solution; the total costs of the two solutions are equals only for $w = 1$ that is when the OP and UP costs are equal; as a matter of example the total costs of the LINEX and MSE solutions for $w = 0.1$ are 1794$ and 2955$ with 40% cost advantage of our proposed prediction solution;
- the better performance in cost total of the LINEX predictions for w lower than 1 is due to the fact that it reduces the resource under-provisioning periods as highlighted from the QoS degradation costs that are lower with respect to the MSE-based prediction solution;
- the better performance in cost total of the LINEX predictions for w higher than 1 is a consequence of the reduction in the over-provisioning periods that, as highlighted in Fig. 11, leads to resource allocation costs lower with respect to the MSE-based prediction solution.

Next we show in Fig. 12 the cost comparison when the traffic predictions are performed with SARIMA and LSTM approaches respectively. The LSTM predicted values are evaluated by applying the proposed traffic forecasting algorithm illustrated in Section V and from the knowledge of the real requested SFC bandwidth values from May 31st 2004 to June 20th 2004 [51]. The real traffic values are used for the LSTM training. To reduce the training times we have considered an LSTM network with the following parameters [36]: i) the number N_{nr} of neurons equals 8; ii) the loop-back parameter L equals 24; iii) the batch size N_{sz} equals 24; iv) the total number N_{ep} of epochs has been fixed

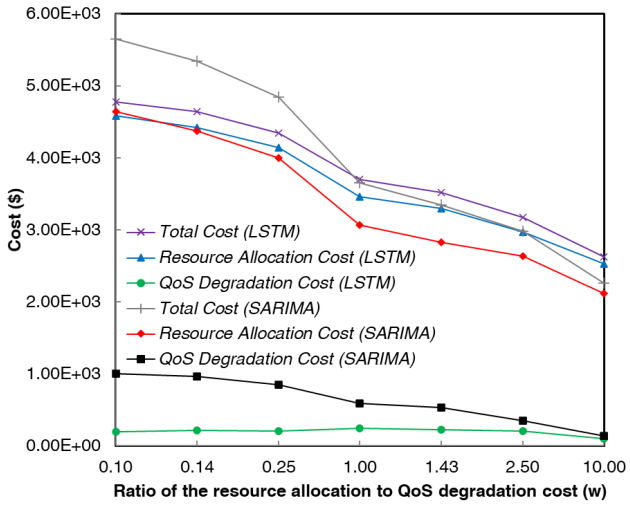


FIGURE 12. Cost in allocating resources for the NFV network of Fig. 5 when w varies from 0.1 to 10. The total, resource allocation and QoS degradation costs are reported when the allocation algorithms use the LSTM and SARIMA predicted SFC bandwidth values.

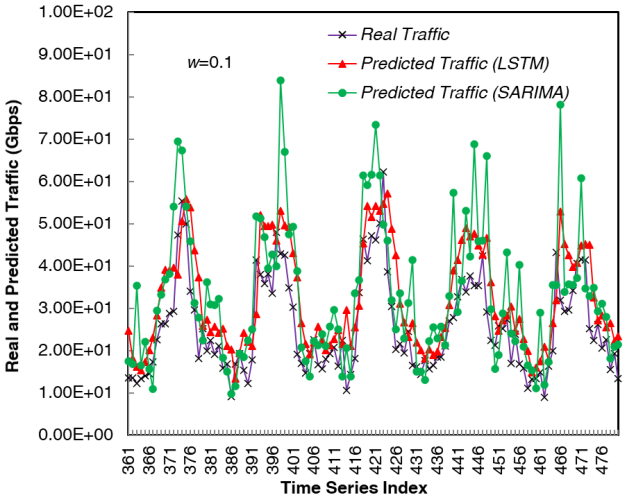


FIGURE 13. Comparison of the Real, LSTM and SARIMA prediction values for $w = 0.1$.

to 20, that is LSTM training process is executed 20 times to find the best model to perform forecasting.

We can notice from Fig. 12 as the application of an AI prediction technique as LSTM allows for lower cost when w is lower than or equal to 1 that is when the UP costs are larger than the OP ones. This is justifiable for the best prediction that the LSTM approach allows to obtain with respect to the SARIMA one as shown in Fig. 13 where we report the real, LSTM and SARIMA predicted traffic values between the nodes Chicago and Indianapolis of the network of Fig. 5 for the period from June 21th 2004 to June 27th 2004. We can observe from Fig. 13 that the application of asymmetric cost function allows both SARIMA and LSTM prediction techniques to provide predicted traffic values larger than the real ones but the LSTM values are closer to the real ones.

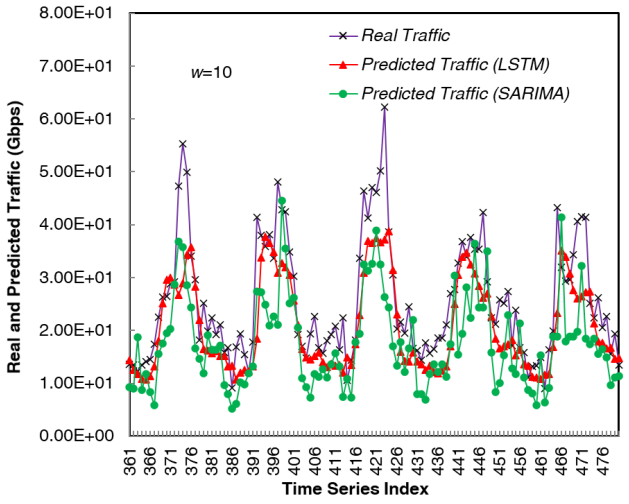


FIGURE 14. Comparison of the Real, LSTM and SARIMA prediction values for $w = 10$.

We report in Fig. 14 the real, LSTM and SARIMA predicted traffic values for the case $w = 10$ that is when the UP costs are lower than the OP ones. We can still observe that the proposed asymmetric SARIMA and LSTM prediction techniques works correctly underestimating the real traffic and as before LSTM approach provides results closer to the real ones with respect to SARIMA. That allows SARIMA, as shown in Fig. 12, to achieve lower resource allocation costs as well as total cost slightly lower than LSTM in the case w larger than 1.

VII. CONCLUSION

We have proposed and investigated traffic prediction techniques in which the predicted values takes into account the OP and UP costs in NFV networks. Since all prediction techniques make prediction errors then the proposed techniques aim to predict under-estimates or over-estimates of traffic depending on whether the OP cost is lower or higher than the UP one respectively. The techniques have been applied to traditional and AI-based prediction algorithms by defining appropriate loss functions. In particular the SARIMA and LSTM prediction algorithms have been considered with LINEX and cusp loss functions respectively. The proposed solutions have been applied to evaluate the operational cost of an Abilene network equipped with four NFVI-PoPs. We have verified how the proposed asymmetric loss functions allows for a cost reduction that can reach the 40% in some cases. Furthermore we have also shown how the LSTM technique is more effective than SARIMA one in reducing the total cost especially when the OP costs are lower then the UP one.

APPENDIX A
EXTENSION OF THE ETSI NFV ARCHITECTURE FOR THE SUPPORT OF THE TRAFFIC PREDICTION AND RESOURCE RECONFIGURATION SOLUTIONS

We show an extension of the ETSI NFV architecture [52] for the support of the proposed LSTM and SARIMA prediction

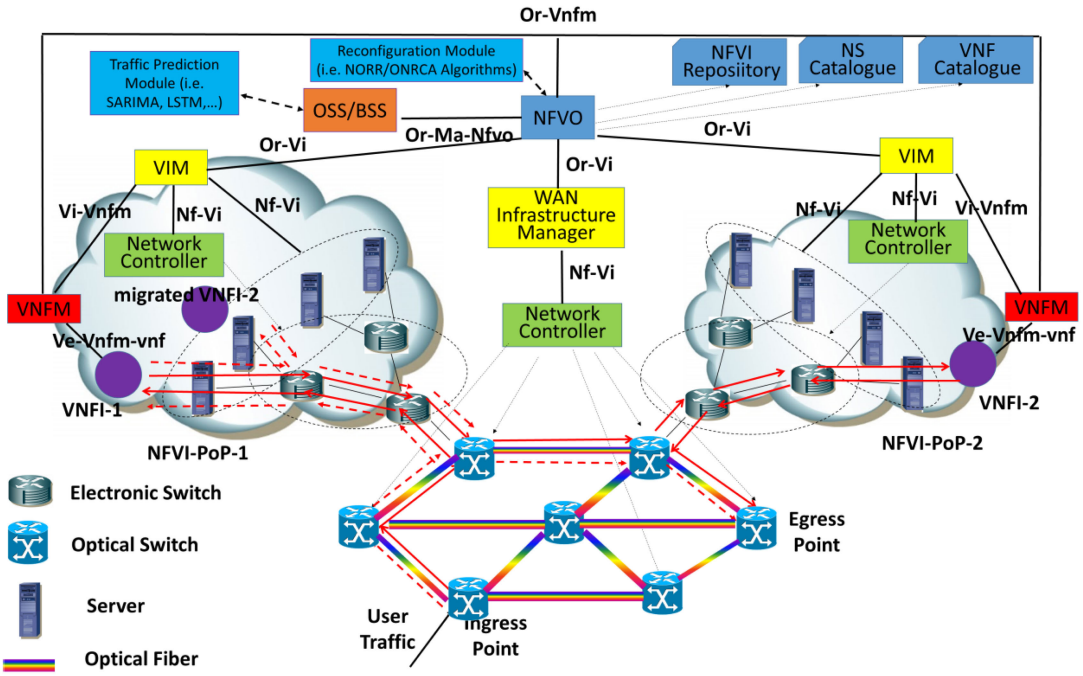


FIGURE 15. ETSI compliant architecture supporting the traffic prediction and resource reconfiguration solutions.

and reconfigurations algorithms. In particular the application of the traffic prediction and resource reconfiguration algorithms will occur as illustrated in Fig. 15 where we report the main functional blocks of the NFV architecture enriched with some operations that allow for the support of proposed algorithms. In the reported example, the processing, RAM and disk memory resources are handled in two NFVI-PoPs. The management and the orchestration of virtualized resources are handed by the Virtual Infrastructure Manager (VIM). In the considered scenario a specialized VIM is also introduced referred to as WAN Infrastructure Manager (WIM) typically used to establish connectivity between access switches in different NFVI-PoPs. VIM and WIM are also helped by Network Controllers to configure both virtual and legacy electronic and optical switches in order to support the concatenation of VNFs of a Network Service (NS). The VNF Managers (VNFM) are responsible for the lifecycle management of VNF Instances and perform functions like VNF instantiation and termination, VNF instance scaling out/in and up/down. The NFV Orchestrator (NFVO) has the responsibilities of the orchestration of NFVI resources across multiple VIM and the lifecycle management of NS. It performs the main following main functions: i) on-boarding of NSs and VNFs in NS and NVF catalogues respectively; ii) storing in the VNFI repository of information about the allocated and consumed cloud and bandwidth resources for the NS instances; iii) NS instantiation and termination and NS instance scaling out/in and up/down. Operation Support System/Business Support System (OSS/BSS) allows for the legacy device management and it provides to submit requests to the NFVO as the ones to on-boarding NS and VNF, to instantiating, terminating and resource scaling any NS and VNF instance. ETSI reports some reference points (Or-Ma-Nfvo, Or-Vi, Nf-Vi, Vi-Vnfm, Or-Vnfm, Ve-Vnfm-vnf) in which some interfaces are defined. Through these interfaces, the functional blocks can call up some operations which allows for the NFV management and orchestration. Next we illustrate how we can support the proposed reconfiguration solution in the NFV architecture scenario depicted in Fig. [52]. First of all, two modules are added to the ETSI NFV architectures. The first one is devoted to estimate traffic; the module performs the estimation by applying algorithms such as SARIMA and LSTM illustrated in this article. The estimated traffic data are sent by the OSS/BSS to the NFVO. The second module provides to the application of the reconfiguration algorithms on the basis of the estimated traffic data; it determines how to reconfigure the NSs so as to minimize the sum of the cloud, bandwidth and reconfigurations costs. The main procedures involved to support the algorithms are:

- A NS Descriptor (NSD) describing a NS in which the cloud and bandwidth resources can be re-allocated is defined; a request is presented by the OSS/BSS to the NFVO for on-boarding the NSD; the request is presented by using the operation On-board Network Service Descriptor of the Network Service Descriptor interface defined by ETSI; the NFVO inserts the NSD in the NSD catalogue and acknowledges the Network Service on-boarding;
- NFVO receives from the OSS/BSS requests for the instantiation of NSs; the requests specify some NSD

descriptor parameters characterizing Access Points, Egress Points, VNFs and cloud and bandwidth resources to be allocated; NFVO receives from OSS/BSS a request to instantiate a Network Service using the operation Instantiate Network Service of the Network Service Lifecycle Management interface defined by ETSI; the NFVO provides to instantiate the NS by contacting all of the NFV actors (VIM, WIM, VNFM,...) and if the operation is successful it acknowledges to the OSS/BSS the completion of the NS instantiation; a NS instance is represented with red arrows in Fig. 15 and involve two VNFIs (VNFI-1 and VNFI-2);

- the estimated traffic data are periodically sent by the OSS/BSS to the NFVO that executes the proposed reconfiguration algorithms so as to minimize the sum of the cloud, bandwidth and reconfiguration costs [7]; next the NFVO activates the VIM, VNFM, WIM and Network Controllers to reconfigure the NSs according to the outputs of the algorithm; for instance we have reported in dashed red arrows the NS reconfigured; in this case VNF-2 is migrated toward the NFVI-PoPs whose cloud resource cost is lower than the one of the NFVI-PoP-2.

APPENDIX B
EVALUATION OF THE TERM $\sigma^2_{n+j|n}$ $(j = 1, \ldots, h)$

The ARMA process expressed by (4) can be equivalently written as follows:

$$\varphi^*(B)D_t = \mu + \omega^*(B)W_t \quad W_t \sim WN(0, \delta^2) \tag{12}$$

where $\varphi^*(B)$ and $\omega^*(B)$ are polynomials of degree $p^* = p+sP$ and $q^* = q+sQ$ respectively. It has been shown [50] that the LINEX predictor has the following expression:

$$\begin{aligned}\hat{g}_{n+j} &= \hat{d}_{n+j} + \frac{a}{2}\sigma^2_{n+j|n} \\ &= \mu - \sum_{i=1}^{j-1}\varphi_i^* E_n[d_{n+j-i}] - \sum_{i=1}^{p^*}\varphi_i^* d_{n+j-i} \\ &\quad - \sum_{i=1}^{q^*}\omega_i^* w_{n+j-i} + \frac{a}{2}\sigma^2_{h+j|h} \quad j = 1, \ldots, h\end{aligned} \tag{13}$$

where the expression of $\sigma^2_{h+j|h}$ is the following:

$$\begin{aligned}\sigma^2_{n+j|n} &= \sum_{i=1}^{j-1}(\varphi_i^*)^2\sigma^2_{n+i|n} + \sum_{i=1}^{j-1}(\omega_i^*)^2\delta^2 \\ &\quad + 2\sum_{i=1}^{j-2}\sum_{k=i+1}^{j-1} E_n[\tilde{e}_{n+j-i|n}\tilde{e}_{n+j-k|n}]\varphi_i^*\varphi_k^* \\ &\quad - 2\sum_{i=1}^{j-1}\sum_{k=i}^{j-1} E_n[\tilde{e}_{n+j-i|n}w_{n+j-k}]\varphi_i^*\omega_k^*\end{aligned} \tag{14}$$

where $\tilde{e}_{n+j-i|n} = d_{n+j-i} - \hat{d}_{n+j-i}$ $(j = 1, \ldots, h)$.

From expression (14) we can notice how the term $\sigma^2_{n+j|n}$ $j = 1, \ldots, h$ can be recursively evaluated starting from $j = 1$ as long as a recursive evaluation of the terms $E_n[\tilde{e}_{n+j-i|n}\tilde{e}_{n+j-k|n}]$ and $E_n[\tilde{e}_{n+j-i|n}w_{n+j-k}]$ can be accomplished.

In particular we need to evaluate the values $\tilde{\sigma}_{i,j} = E_n[\tilde{e}_{n+i|n}w_{n+j}]$ $(i, j = 1, \ldots, h)$ and $\tilde{\rho}_{i,j} = E_n[\tilde{e}_{n+i|n}\tilde{e}_{n+j|n}]$.

By taking account of 13 and after some algebra we achieve the iterative procedures in Algorithm 3 and Algorithm 4 for the evaluation of $\tilde{\sigma}_{i,j}$ and $\tilde{\rho}_{i,j}$ $(i, j = 1, \ldots, h)$ respectively.

Algorithm 3 Iterative Procedure for the Evaluation of $\tilde{\sigma}_{i,j}$ $(i, j = 1, \ldots, h)$

$\tilde{\sigma}_{1,1} = \delta^2$
for $i \in [1..h]$ **do**
  for $j \in [1..i-1]$ **do**
    /*$\delta_D(\bullet)$ ***is the discrete Dirac impulse****/
    $\tilde{\sigma}_{i,j} = \sum_{s=1}^{p^*}\varphi_s^*\tilde{\sigma}_{i-s,j} - \sum_{s=1}^{q^*}\delta_D(s-(h+p))\omega_i^*\delta^2$
  end for
  $\tilde{\sigma}_{i,i} = \delta^2$
end for
Output: $\tilde{\sigma}_{i,j}$ $(i, j = 1, \ldots, h)$

Algorithm 4 Iterative Procedure for the Evaluation of $\tilde{\rho}_{i,j}$ $(i, j = 1, \ldots, h)$

$\tilde{\rho}_{1,1} = \delta^2$
for $i \in [1..h]$ **do**
  for $j \in [1..i-1]$ **do**
    $\tilde{\rho}_{i,j} = \sum_{s=1}^{i-1}\sum_{t=1}^{j-1}\varphi_s^*\varphi_t^*\tilde{\rho}_{i-s,j-t} +$
    $- \sum_{s=j-i}^{j-1}\omega_s^*\omega_{i-j+s}^*\delta^2 +$
    $- \sum_{s=1}^{i-1}\sum_{t=max(0,j-i+s)}^{j-1}\varphi_s^*\omega_t^*\tilde{\rho}_{i-s,j-t} +$
    $- \sum_{s=0}^{i-1}\sum_{t=min(j-1,j-i+s)}^{j-1}\varphi_t^*\omega_s^*\tilde{\rho}_{i-t,j-s}$
  end for
  $\tilde{\rho}_{i,i} = \delta^2$
end for
Output: $\tilde{\rho}_{i,j}$ $(i, j = 1, \ldots, h)$

REFERENCES

[1] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, "Network function virtualization: State-of-the-Art and research challenges," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 236–262, 1st Quart., 2016.
[2] B. Yi, X. Wang, K. Li, S. K. Das, and M. Huang, "A comprehensive survey of network function virtualization," *Comput. Netw.*, vol. 133, pp. 212–262, Mar. 2018.
[3] A. A. Barakabitze, A. Ahmad, R. Mijumbi, and A. Hines, "5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges," *Comput. Netw.*, vol. 167, Feb. 2020, Art. no. 106984.
[4] J. Pei, P. Hong, M. Pan, J. Liu, and J. Zhou, "Optimal VNF placement via deep reinforcement learning in SDN/NFV-enabled networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 2, pp. 263–278, Feb. 2020.
[5] B. Farkiani, B. Bakhshi, and S. A. MirHassani, "A fast near-optimal approach for energy-aware SFC deployment," *IEEE Trans. Netw. Service Manage.*, vol. 16, no. 4, pp. 1360–1373, Dec. 2019.
[6] J. Chen, J. Chen, R. Hu, and H. Zhang, "ClusVNFI: A hierarchical clustering-based approach for solving VNFI dilemma in NFV orchestration," *IEEE Access*, vol. 7, pp. 173257–173272, 2019.

[7] V. Eramo and F. G. Lavacca, "Computing and bandwidth resource allocation in multi-provider NFV environment," *IEEE Commun. Lett.*, vol. 22, no. 10, pp. 2060–2063, Oct. 2018.

[8] Y. Yu, X. Bu, K. Yang, H. K. Nguyen, and Z. Han, "Network function virtualization resource allocation based on joint benders decomposition and ADMM," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1706–1718, Feb. 2020.

[9] M. Karimzadeh-Farshbafan, V. Shah-Mansouri, and D. Niyato, "Reliability aware service placement using a viterbi-based algorithm," *IEEE Trans. Netw. Service Manage.*, vol. 17, no. 1, pp. 622–636, Mar. 2020.

[10] I. El Mensoum, O. Abdul Wahab, N. Kara, and C. Eds.,trom, "MuSC: A multi-stage service chains embedding approach," *J. Netw. Comput. Appl.*, vol. 159, Jun. 2020, pp. 1–18.

[11] T. M. Pham, S. Fdida, T. T. L. Nguyen, and H. N. Chu, "Modeling and analysis of robust service composition for network functions virtualization," *Comput. Netw.*, vol. 166, pp. 1–16, Jan. 2020.

[12] V. Eramo and F. G. Lavacca, "Proposal and investigation of a reconfiguration cost aware policy for resource allocation in multi-provider NFV infrastructures interconnected by elastic optical networks," *J. Lightw. Technol.*, vol. 37, no. 16, pp. 4098–4114, Aug. 15, 2019.

[13] F. Matera, A. Schiffini, M. Guglielmucci, M. Settembre, V. Eramo, "Numerical investigation on design of wide geographical optical-transport networks based on n×40-Gb/s transmission," *J. Lightw. Technol.*, vol. 21, no. 2, pp. 456–465, Feb. 2003.

[14] R. Zhu, S. Li, P. Wang, Y. Tan, and J. Yuan, "Gradual migration of co-existing Fixed/Flexible optical networks for cloud-fog computing," *IEEE Access*, vol. 8, pp. 50637–50647, 2020.

[15] M. Garrich, F.-J. Moreno-Muro, M.-V. Bueno Delgado, and P. Pavon Marino, "Open-source network optimization software in the open SDN/NFV transport ecosystem," *J. Lightw. Technol.*, vol. 37, no. 1, pp. 75–88, Jan. 1, 2019.

[16] T. Shen, M. Zhu, J. Gu, X. Ren, B. Chen, and C. Shi, "When virtual network functions are deployed in network resource virtualized elastic optical networks," *Proc. SPIE*, vol. 11435, Mar. 2020, Art. no. 114350J.

[17] H. Xuan, S. Wei, Y. Feng, D. Liu, and Y. Li, "Bi-level programming model and algorithm for VNF deployment with data centers placement," *IEEE Access*, vol. 7, pp. 185760–185772, 2019.

[18] V. Eramo, E. Miucci, and M. Ammar, "Study of reconfiguration cost and energy aware VNE policies in cycle-stationary traffic scenarios," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1281–1297, May 2016.

[19] *Spectral Grids for WDM Applications: DWDM Frequency Grid*, document Rec. G.694.1, Jun. 2002. [Online]. Available: https://www.itu.int/rec/T-REC-G.694.1-201202-I/en

[20] A. Agrawal, V. Bhatia, and S. Prakash, "Low-Crosstalk-Margin routing for spectrally-spatially flexible optical networks," *IEEE Commun. Lett.*, vol. 24, no. 4, pp. 835–839, Apr. 2020.

[21] N. Shahriar, M. Zulfiqar, S. R. Chowdhury, S. Taeb, M. Tornatore, R. Boutaba, J. Mitra, and M. Hemmati, "Disruption-minimized re-adaptation of virtual links in elastic optical networks," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, San Diego, CA, USA, Mar. 2020, pp. 1–3.

[22] Y. Liu, H. Lu, X. Li, and D. Zhao, "An approach for service function chain reconfiguration in network function virtualization architectures," *IEEE Access*, vol. 7, pp. 147224–147237, 2019.

[23] M. Karimzadeh-Farshbafan, V. Shah-Mansouri, and D. Niyato, "A dynamic reliability-aware service placement for network function virtualization (NFV)," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 2, pp. 318–333, Feb. 2020.

[24] B. Yi, X. Wang, M. Huang, and K. Li, "Design and implementation of network-aware VNF migration mechanism," *IEEE Access*, vol. 8, pp. 44346–44358, 2020.

[25] X. Fu, F. R. Yu, J. Wang, Q. Qi, and J. Liao, "Dynamic service function chain embedding for NFV-enabled IoT: A deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 507–519, Jan. 2020.

[26] V. Eramo and F. G. Lavacca, "Optimizing the cloud resources, bandwidth and deployment costs in multi-providers network function virtualization environment," *IEEE Access*, vol. 7, pp. 46898–46916, 2019.

[27] W. Ma, J. Beltran, D. Pan, and N. Pissinou, "Placing traffic-changing and partially-ordered NFV middleboxes via SDN," *IEEE Trans. Netw. Service Manage.*, vol. 16, no. 4, pp. 1303–1317, Dec. 2019.

[28] H.-G. Kim, D.-Y. Lee, S.-Y. Jeong, H. Choi, J.-H. Yoo, and J. W.-K. Hong, "Machine learning-based method for prediction of virtual network function resource demands," in *Proc. IEEE Conf. Netw. Softwarization (NetSoft)*, Paris, France, Jun. 2019, pp. 405–413.

[29] M. Ghaznavi, A. Khan, N. Shahriar, K. Alsubhi, R. Ahmed, and R. Boutaba, "Elastic virtual network function placement," in *Proc. IEEE 4th Int. Conf. Cloud Netw. (CloudNet)*, Oct. 2015, pp. 1–6.

[30] V. Eramo, E. Miucci, M. Ammar, and F. G. Lavacca, "An approach for service function chain routing and virtual function network instance migration in network function virtualization architectures," *IEEE/ACM Trans. Netw.*, vol. 25, no. 4, pp. 2008–2025, Aug. 2017.

[31] D. Bega, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "DeepCog: Optimizing resource provisioning in network slicing with AI-based capacity forecasting," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 2, pp. 361–376, Feb. 2020.

[32] D. Bega, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "DeepCog: Cognitive network management in sliced 5G networks with deep learning," in *Proc. IEEE INFOCOM-IEEE Conf. Comput. Commun.*, Paris, France, Apr. 2019, pp. 280–288.

[33] D. Bega, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "α-OMC: Cost-aware deep learning for mobile network resource orchestration," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Paris, France, Apr./May 2019, pp. 423–428.

[34] B. Li, W. Lu, S. Liu, and Z. Zhu, "Deep-Learning-Assisted network orchestration for on-demand and cost-effective vNF service chaining in inter-DC elastic optical networks," *J. Opt. Commun. Netw.*, vol. 10, no. 10, p. D29, Oct. 2018.

[35] B. Li, W. Lu, S. Liu, and Z. Zhu, "Designing deep learning model for accurate vNF service chain pre-deployment in inter-DC EONs," in *Proc. Asia Commun. Photon. Conf. (ACP)*, Hangzhou, China, Oct. 2018, pp. 1–3.

[36] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 841–851, Jan. 2019.

[37] A. C. Riekstin, A. Langevin, T. Dandres, G. Gagnon, and M. Cheriet, "Time series-based GHG emissions prediction for smart homes," *IEEE Trans. Sustain. Comput.*, vol. 5, no. 1, pp. 134–146, Jan. 2020.

[38] H. Tang, D. Zhou, and D. Chen, "Dynamic network function instance scaling based on traffic forecasting and VNF placement in operator data centers," *IEEE Trans. Parallel Distrib. Syst.*, vol. 30, no. 3, pp. 530–543, Mar. 2019.

[39] F. Farahnakian, T. Pahikkala, P. Liljeberg, J. Plosila, N. T. Hieu, and H. Tenhunen, "Energy-aware VM consolidation in cloud data centers using utilization prediction model," *IEEE Trans. Cloud Comput.*, vol. 7, no. 2, pp. 524–536, Apr. 2019.

[40] J. Han, *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann, 2005.

[41] F. J. Ferrer-Troyano, J. S. Aguilar-Ruiz, and J. C. Riquelme, "Empirical evaluation of the difficulty of finding a good value of k for the nearest neighbor," in *Proc. Int. Conf. Comput. Sci.*, San Diego, CA, USA, Dec. 2003, pp. 766–773.

[42] L. Tang, X. He, P. Zhao, G. Zhao, Y. Zhou, and Q. Chen, "Virtual network function migration based on dynamic resource requirements prediction," *IEEE Access*, vol. 7, pp. 112348–112362, 2019.

[43] J. Pati, B. Kumar, D. Manjhi, and K. K. Shukla, "A comparison among ARIMA, BP-NN, and MOGA-NN for software clone evolution prediction," *IEEE Access*, vol. 5, pp. 11841–11851, 2017.

[44] Q. Yang, Y. Zhou, Y. Yu, J. Yuan, X. Xing, and S. Du, "Multi-step-ahead host load prediction using autoencoder and echo state networks in cloud computing," *J. Supercomput.*, vol. 71, no. 8, pp. 3037–3053, Aug. 2015.

[45] B. Song, Y. Yu, Y. Zhou, Z. Wang, and S. Du, "Host load prediction with long short-term memory in cloud computing," *J. Supercomput.*, vol. 74, no. 12, pp. 6554–6568, Dec. 2018.

[46] H. M. Nguyen, G. Kalra, and D. Kim, "Host load prediction in cloud computing using long short-term memory Encoder–Decoder," *J. Supercomput.*, vol. 75, pp. 7592–7605, Aug. 2019.

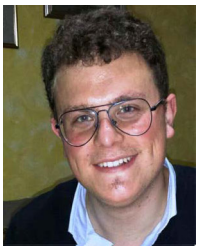
[47] S. Rahman, T. Ahmed, M. Huynh, M. Tornatore, and B. Mukherjee, "Auto-scaling VNFs using machine learning to improve QoS and reduce cost," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kansas City, MO, USA, May 2018, pp. 1–6.

[48] V. Eramo, T. Catena, F. G. Lavacca, and F. di Giorgio, "Study and investigation of SARIMA-based traffic prediction models for the resource allocation in NFV networks with elastic optical interconnection," in *Proc. 22nd Int. Conf. Transparent Opt. Netw. (ICTON)*, Bari, Italy, Jul. 2020, pp. 1–4.

[49] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*. Cham, Switzerland: Springer, 2016.

[50] M. Niglio, "Multi-step forecasts from threshold ARMA models using asymmetric loss functions," *Stat. Methods Appl.*, vol. 16, no. 3, pp. 395–410, Oct. 2007.
[51] SND-Lib. *Dynamic Traffic Section*. Accessed: Oct. 10, 2020. [Online]. Available: http://sndlib.zib.de/home.action
[52] ETSI Industry Specification Group (ISG) NFV. (Jan. 2015). *ETSI Group Specifications on Network Function Virtualization*. [Online]. Available: http://docbox.etsi.org/ISG/NFV/Open/Published/

VINCENZO ERAMO (Member, IEEE) received the Laurea degree in electronics engineering and the Ph.D. degree in information and communications engineering from the University of Roma La Sapienza, in 1995 and 2001, respectively. From June 1996 to December 1996, he was a Researcher with the Scuola Superiore Reiss Romoli. In 1997, he joined the Fondazione Ugo Bordoni as a Researcher with the Telecommunication Network Planning Group. He was an Assistant Professor and an Aggregate Professor with the INFOCOM Department, University of Rome La Sapienza, from November 2002 to October 2005 and November 2006 to June 2010, respectively. He is currently an Associate Professor with the Department of Engineering of Information, Electronics and Telecommunications. His research activities have been carried out in the framework of national and international projects. In particular, he was a Scientific Coordinator for the University of Roma La Sapienza of E-PhotoONe+ and BONE, two Networks of Excellence focusing on the study of Optical Networks and financed by European Commissions (FP6 and FP7), from 2006 to 2007 and 2008 to 2011, respectively. He was a winner of the Exemplary Editor Award 2016 and 2017 of IEEE COMMUNICATIONS LETTERS. He was an Associate Editor of IEEE TRANSACTIONS ON COMPUTERS from July 2011 to June 2015. He has been an Associate Editor of IEEE COMMUNICATIONS LETTERS since September 2014.



FRANCESCO GIACINTO LAVACCA received the Laurea (M.Sc.) degree *(cum laude)* in electronic engineering and the Ph.D. degree in information technology from the Sapienza University of Rome, Italy, in 2013 and 2017, respectively. He is currently a Postdoctoral Researcher with the Department of Information, Electronic and Telecommunication engineering (DIET), Sapienza University of Rome. Since 2016, he has been a Visiting Researcher with the College of Computing, Georgia Institute of Technology, Atlanta, GA, USA. He has been with Fondazione Ugo Bordoni since November 2018. He was involved in the framework of national and international projects, like Advanced Avionic Architecture (AAA) and Nano Micro Lanch Vehicle (NMLV) with Italian Space Agency (ASI) and European Space Agency (ESA), respectively. His current research interests include all-optical networks and switching architectures, 5G networks, network function virtualization, and time-triggered and deterministic Ethernet.

TIZIANA CATENA received the Laurea (M.Sc.) degree *(cum laude)* in electronic engineering from the Sapienza University of Rome, Italy, in 2018, where she is currently pursuing the Ph.D. degree in information and communication technology (ICT). Her current research interest includes network function virtualization.

FLAVIO DI GIORGIO received the Laurea (M.Sc.) degree *(cum laude)* in electronic engineering from the Sapienza University of Rome, Italy, in 2020. He is currently involved in research activities on network function virtualization with Sapienza University of Rome.