future internet

MDPI

Article

Proposal and Investigation of an Artificial Intelligence (AI)-Based Cloud Resource Allocation Algorithm in Network Function Virtualization Architectures †

Vincenzo Eramo [1,*], Francesco Giacinto Lavacca [2] and Tiziana Catena [1] and Paul Jaime Perez Salazar [1]

[1] Department of Information, Electronic, Telecommunication (DIET), "Sapienza" University of Rome—Via Eudossiana 18, 00184 Rome, Italy; tiziana.catena@uniroma1.it (T.C.); perezsalazar.1393125@studenti.uniroma1.it (P.J.P.S.)

[2] Fondazione Ugo Bordoni, Viale del Policlinico 147, 00161 Rome, Italy; fglavacca@fub.it

* Correspondence: Vincenzo.Eramo@uniroma1.it; Tel.: +39-06-44585372

† This paper is an extended version of paper published in Proceedings of the 22nd International Conference on Transparent Optical Network (ICTON), Bari, Italy, 19–23 July 2020.

Received: 28 September 2020; Accepted: 10 November 2020; Published: 13 November 2020

Abstract: The high time needed to reconfigure cloud resources in Network Function Virtualization network environments has led to the proposal of solutions in which a prediction based-resource allocation is performed. All of them are based on traffic or needed resource prediction with the minimization of symmetric loss functions like Mean Squared Error. When inevitable prediction errors are made, the prediction methodologies are not able to differently weigh positive and negative prediction errors that could impact the total network cost. In fact if the predicted traffic is higher than the real one then an over allocation cost, referred to as over-provisioning cost, will be paid by the network operator; conversely, in the opposite case, Quality of Service degradation cost, referred to as under-provisioning cost, will be due to compensate the users because of the resource under allocation. In this paper we propose and investigate a resource allocation strategy based on a Long Short Term Memory algorithm in which the training operation is based on the minimization of an asymmetric cost function that differently weighs the positive and negative prediction errors and the corresponding over-provisioning and under-provisioning costs. In a typical traffic and network scenario, the proposed solution allows for a cost saving by 30% with respect to the case of solution with symmetric cost function.

Keywords: network function virtualization; computing resources; machine learning; long short term memory

1. Introduction

Network Function Virtualization [1] technology allows the implementation of software middleboxes located in data centers, referred to as Network Function Virtual Infrastructure-Point of Presence (NFVI-PoP), and running on virtual machines. In these last few years the problem of resource reconfiguration in NFV environments has been widely investigated [2–6]. The studies focused on reactive techniques based on which the network is reconfigured as soon as traffic changes occur [7–10]. Everyone agrees that reactive techniques are ineffective in relation to the high variability of traffic and the high time of cloud resource allocation. For this reason proactive reconfiguration techniques have been proposed where the traffic or the amount of resources needed is predicted [11]. Traffic and

cloud resource prediction methodologies have been recently used in Network Function Virtualization environments for cloud and bandwidth resource allocation purposes. Both traditional and innovative prediction methodologies [12] have been proposed. For instance, Long Short Term Memory-based prediction techniques have been shown to be very effective in allocating resources. All of these techniques are based on the minimization of symmetric cost functions as the Mean Square Error (MSE) that equally weighs positive and negative prediction errors. However, the error sign can differently impact the cost increase due to prediction errors. For instance, when the Quality of Service degradation cost due to traffic loss is prevalent with respect to the cloud resource allocation cost, an algorithm is preferable that overestimates the offered traffic; conversely, the traffic underestimation is preferable in the opposite case when the cloud allocation cost is lower than the QoS degradation one.

To our best knowledge, only Bega et al. [13] proposes a solution for mobile network resource orchestration in which the different values of the over-provisioning and under-provisioning costs are taken into account. DeepCog is proposed, a framework for 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, which is independent of the lost traffic amount, when a QoS degradation occurs.

In this paper we 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 over-provisioning and under-provisioning costs. This objective is achieved by minimizing an asymmetric cost function characterized by a parameter that takes into account the over-provisioning and under-provisioning costs. The principle of the proposed solution can be applied to any prediction technique and in this work it is applied to one that predicts traffic with a Long Short Term Memory (LSTM) prediction methodology.

The main contributions of the manuscript are the following:

- The study and the investigation of a prediction-based resource allocation algorithm for NFV network environments;
- The study and the investigation of an LSTM-based traffic prediction algorithm with an asymmetric loss function that optimally predicts the traffic values according to the over-provisioning and under-provisioning costs;
- A performance comparison of the proposed solution to other ones proposed in literature and based on predictions with minimization of symmetric loss functions.

The paper is organized as follows. We describe the state-of-the-art, the problem statement and the prediction-based reconfiguration algorithm in Section 2. The traffic forecasting technique based on asymmetric traffic function is illustrated in Section 3. The numerical results, reported in Section 4, show the effectiveness of the proposed technique with respect to MSE-based traditional forecasting techniques in an NFV network environment.

2. Intelligence Artificial (AI)-Based Resource Allocation Algorithms in Dynamic Traffic Scenario

2.1. State-of-the-Art

In NFV networks, a network service is composed of a set of Network Functions (NFs) connected in a specific order. As these NFs are implemented as VNFs, the VNF Forwarding Graph (VNF-FG) provides the logical connectivity between them. In other words, a VNF-FG defines the possible sequences of VNFs that the packets traverse between two endpoints, to realize an end-to-end service. The variability of traffic and services required leads to the need to define algorithms for the reconfiguration of cloud and bandwidth resources. The typical variations that could happen are as follows [14]:

- The instantiation of new Virtual Network Function-Forwarding Graphs (VNF-FG);
- The extension or the reduction of VNF-FGs already instantiated with the addition or the removal of VNFs;

- The variation of required bandwidth of the current VNF-FGs.

The reconfiguration of NFV networks is based on various techniques of which the most important are those of:

- Increasing and decreasing the cloud resources (CPU, memory, disk, etc.) assigned to Virtual Network Function Instances (VNFI) that support VNFs; it is possible to apply horizontal and vertical scaling techniques; the former are based on the increase (scaling in) and decrease (scaling out) of the number of Virtual Machines assigned to each VNFI; the latter are based on the increase (scaling up) or decrease (scaling down) of the cloud resources assigned to the single Virtual Machine in which the VNFI are executing the VNFs.
- Migrating VNFIs to other servers or even other NFVI-PoPs with the application of the above-mentioned scaling techniques.

Reactive reconfiguration procedures are not adequate due to the high time required to reallocate cloud resources [11]. For this reason algorithms based on predictions have recently been proposed. There are two categories of solutions: the first one based on traffic prediction [15], the second one based on the prediction of resources to be allocated [12].

Some state-of-the-art solutions are reported in Table 1. They are compared in terms of prediction type (traffic or resource-based), prediction methodology and loss function characteristic to be minimized. Among the solutions based on traffic prediction, Li et al. [15] proposes a Deep Learning (DL) framework based on Long/Short Term Memory recurrent neural networks [16] to predict the VNF-FGs requests in an NFV network with NFVI-PoPs interconnected by an Elastic Optical Network; the arrival and hold-on times, the bandwidth, the originating and terminating nodes and the type of the SFCs are predicted. Among the solutions based on the prediction of the resources to be allocated, Farahnakian et al. [17] proposes regressive algorithms for estimating memory and processing consumption in cloud data centers; the proposed solutions are based on Linear Regression [18] and K-Nearest Neighbor Regression (K-NNR) [19] methods that notoriously determine the prediction by minimizing symmetric error functions.

Unfortunately, there are random components that are not predictable and that lead 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 over-provisioning cost; in the opposite case less resources will be allocated and this will lead to a QoS degradation and to an under-provisioning cost characterized by the compensation due to the user. To our knowledge only in [13] traffic prediction is performed taking into account the over-provisioning and under-provisioning cost. The authors use convolutional neural networks for the prediction, model only the datacenters and consider under-provisioning costs independent of traffic loss. Conversely we propose a solution in which: (i) both NFVI and transport infrastructures are modeled; (ii) the prediction is performed by using an LSTM recurrent neural network with an asymmetric loss function where the cost is dependent on the lost traffic amount.

Table 1. Comparison between our proposal and the main related works.

Work	**Prediction Type**	**Prediction Methodology**	**Minimized Loss Function**
Schneider et al. [12]	Resource	Linear Regression, Support Vector Machine	Symmetric
Li et al. [15]	Traffic	LSTM	Symmetric
Farahnakian et al. [17]	Resource	Regression Linear, K-Nearest Neighbor Regression	Symmetric
Bega et al. [13]	Traffic	Convolutional Neural Networks	Asymmetric
Our Proposal	Traffic	LSTM	Asymmetric

A preliminary result on the advantages of the traffic prediction with an asymmetric loss function has been investigated in [20] when the prediction is based on Seasonal Auto Regressive Integrated Moving Average (ARIMA) processes. In this manuscript we extend the proposed solution to the case of LSTM-based predictions. The following contributions are added in this manuscript:

- An innovative prediction algorithm based on an LSTM recurrent neural network with an asymmetric cusp loss function is proposed;
- The performance of electrical networks is investigated; conversely, resource allocation for Optical NFV networks is investigated in [20];
- Extensive numerical results are reported in which the operational costs of an NFV network with resource allocation based on symmetric and asymmetric LSTM are evaluated respectively;
- New results are presented with respect to [20] in which the resource allocation is not only performed after on the prediction step but the new approach allows for a multi-step prediction and resource allocation.

2.2. Under-Provisioning and Over-Provisioning Costs in Prediction-Based NFV Reconfiguration Algorithms

We show a simple scenario of one VNFI activated in the NFVI-POP of Figure 1a. Processing resources, represented by black rectangles, are allocated to the VNFI. In a dynamic traffic scenario, the cloud resources have to be reallocated to the VNFI according to the current traffic conditions. We report the cloud resource reconfiguration in Figure 1b in the case of a traffic increase. For handling this increase the cloud resources allocated to the VNFI are increased by applying a vertical scaling technique that leads to increase the processing capacity of an amount represented with a grey rectangle in Figure 1b. Reactive reconfiguration approaches are not suited in NFV environments especially due to the high time needed to reconfigure the cloud resources [15]. 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 consequent increase in operational network cost.

Over provisioning occurs when the predicted traffic is higher than the real one; in this case more cloud resources than needed are allocated; an example of over provisioning is illustrated in Figure 1c where the additional cloud 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 Figure 1d 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 service provider due to the compensation cost to be paid to the user for the lost traffic.

The proposed resource allocation procedures are based on two algorithms:

- A reconfiguration algorithm: it uses the predicted traffic values to reconfigure bandwidth and cloud resource, migrate VNFI, etc.;
- A traffic prediction algorithm: it uses LSTM-based advanced prediction mechanisms to predict the traffic values.

Briefly we discuss how the NFV architecture proposed by European Telecommunication Standards Institute (ETSI) [21,22] may be extended to support the proposed resource allocation procedure. The main extensions are the following:

- The Operation Support System/Business Support System (OSS/BSS) may receive the measured real traffic values from the monitored network devices; then it may perform the prediction algorithm to determine the predicted traffic values;
- The Network Function Virtualization Orchestrator (NFVO) receives the predicted traffic values and by applying the reconfiguration algorithm can decide on reallocating bandwidth and cloud

resources and/or migrating VNFIs; Virtual Manager Infrastructures and network controllers are used to actuate the reconfigurations decided by the NFVO.

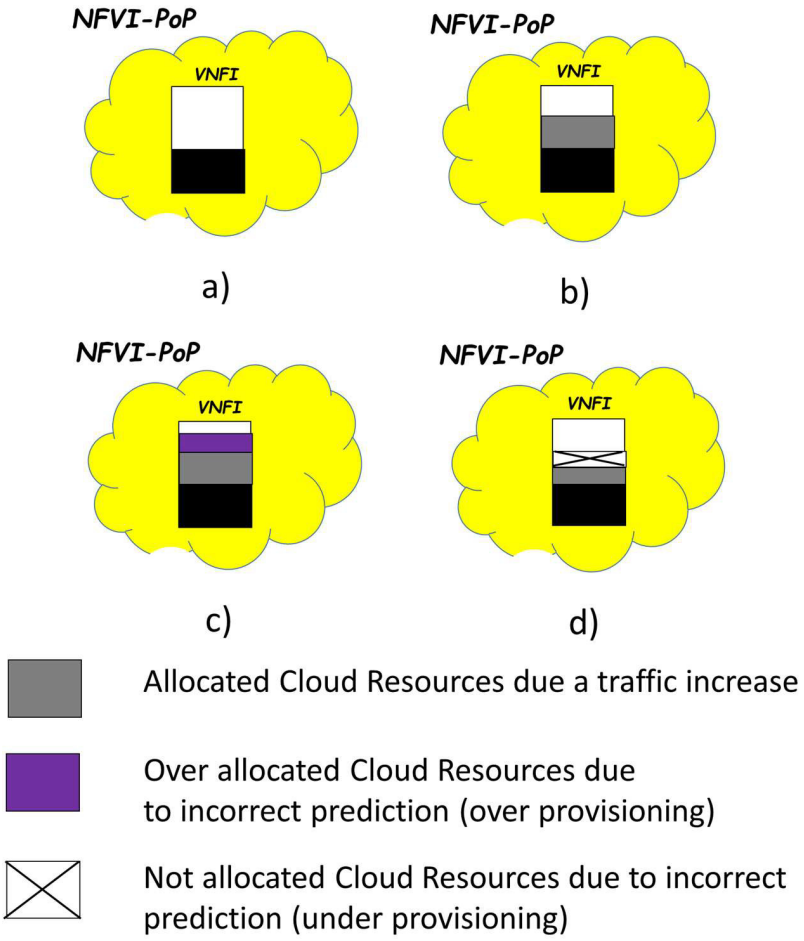

Figure 1. Cloud resource allocation for one Virtual Network Function Instance (VNFI) in a Network Function Virtual Infrastructure-Point of Presence (NFVI-PoP) (**a**); resource reconfiguration when a traffic increase occurs (**b**); resources over (**c**) and under (**d**) provisioning when a traffic prediction error is made.

2.3. Prediction-Based NFV Reconfiguration Algorithm

We illustrate a very general NFV reconfiguration algorithm for the case where reconfiguration is necessary due to VNF-FG traffic variations. The sets, parameters and variables are defined in Tables 2 and 3. In this paper we assume that the VNF-FG are linear graphs in which each link is characterized by the same bandwidth value. Because we also assume that N VNF-FGs are considered and the VNF-FG bandwidth can vary over time then we denote with $b_j(i)$ $(i = 1, \cdots, N; j = 1, \cdots)$ the bandwidth of i-th VNF-FG in j-th Time Interval (TI). The TI duration is denoted with T_s.

To introduce the algorithm we introduce the VNFI graph, the nodes of which correspond to the instanced VNFIs while the edges correspond to virtual links interconnecting the VNFIs. It should be noted that the VNFI graph corresponds to a VNF-FG when VNFIs are not shared between VNF-FGs. On the contrary, in the case of VNFI sharing, it provides information on the total set of VNFIs instantiated to support VNF-FGs and their interconnection.

Table 2. Sets and Parameters.

Sets and Parameters	Definition
N	Number of VNF-FGs
$b_j(i)$	Bandwidth of the i-th VNF-FG in the j-th TI
V	Set of nodes of the VNFI graph
L	Set of links of the VNFI graph
$\bar{V}$	Set of nodes of the physical graph
$\bar{L}$	Set of links of the physical graph
$f_v^{(j)}$	Processing capacity required by the node $v \in V$ in the j-th TI
$f_e^{(j)}$	Bandwidth of the link $e \in L$ in the j-th TI
C_{RA}	Cloud Resource Allocation Cost
C_{QoS}	QoS degradation cost
h	Prediction step
$b_{n,h}$	Vector of the h bandwidth values $b_{(n+j)}$ $(j = 1, \cdots, h)$ for the generic VNG-FG

Table 3. Variables.

$\widehat{b}_{n+j}(i)$	Predicted bandwidth of the i-th VNF-FG in $(n+j)$-th TI
$\widehat{f}_v^{(n+j)}$	Predicted Processing capacity required by the node $v \in V$ in $(n+j)$-th TI
$\widehat{f}_e^{(n+j)}$	Predicted bandwidth of the link $e \in L$ in $(n+j)$-th TI
$\widehat{b}_{n,h}$	Vector of the h predicted bandwidth values $\widehat{b}_{n+j}(i)$ $(j = 1, \cdots, h)$ for the generic VNG-FG
e_{n+j}	Bandwidth prediction error for a generic VNF-FG in $(n+j)$-th TI

The NFV reconfiguration 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 of any VNFI is executed; (ii) the cloud (processing) resources to be assigned to the VNFIs; (iii) in which network paths any logical link has to be routed. 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 [8] 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 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 capacity $f_v^{(j)}$ and the link bandwidth $f_e^{(j)}$ are given by the sum of the bandwidths of VNF-FGs that share the node $v \in V$ and the link $e \in L$ respectively. Hence the processing capacities and the link bandwidths are depending on the offered VNF-FG bandwidths and for this reason they are not a-priori known. We report a reconfiguration solution based on the prediction of the offered VNF-FG bandwidths. The algorithm can be easily extended to the case in which the values of $f_v^{(j)}$ $(j = 1, 2, \cdots; v \in V)$ and $f_e^{(j)}$ $(j = 1, 2, \cdots; e \in L)$ are directly predicted.

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 $\bar{G} = (\bar{V}, \bar{L})$, the VNF-FG bandwidths

$b_j(i)$ $(i = 1, \cdots, N, j = 1, \cdots, n)$ known up to TI n and the VNFI graph $G = (V, L)$. Next a multi-step ahead prediction of the VNF-FG bandwidth is performed in step 2 by predicting the next h VNF-FG bandwidth values $\widehat{b}_{n+j}(i)$ $(i = 1, \cdots, N, j = 1, \cdots, h)$. 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 $n+1, \cdots, n+h$. 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 $\Gamma_{n+j}(\bar{G}, G)$ $(j = 1, \cdots, h)$ to be applied in the TIs $n+1, \cdots, n+h$. We apply the reconfiguration algorithms proposed in [8] referred to as Least Cloud Resource and Bandwidth (LCBC) and Deployment Costs Aware (DCA). The new embeddings are evaluated from the current embedding $\Gamma_c(\bar{G}, G)$, which is the one applied in TI n. Finally the framework returns the evaluated embeddings $\Gamma_{n+j}(\bar{G}, G)$ $(j = 1, \cdots, h)$.

Algorithm 1 PREDICTION-BASED NFV RECONFIGURATION ALGORITHM

1: **Input:**
Cloud Infrastructure and Network Graph $\bar{G} = (\bar{V}, \bar{L})$
VNF-FG bandwidths: $b_j(i)$ $(i = 1, \cdots, N, j = 1, \cdots, n)$
VNFI graph: $G = (V, L)$
Current Embedding $\Gamma_c(\bar{G}, G)$
/*Multi-step ahead VNF-FG bandwidth Prediction*/
2: **Predict** the VNF-FG bandwidths $\widehat{b}_{n+j}(i)$ $(i = 1, \cdots, N, j = 1, \cdots, h)$ with an 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 $n+1, \cdots, n+h$
4: **Reconfigure** the bandwidth and the cloud resources by applying the algorithms LCBC/DCA [8] and evaluating the embeddings $\Gamma_{n+j}(\bar{G}, G)$ $(j = 1, \cdots, h)$ in the ITs $n+1, \cdots, n+h$
5: **Output:** $\Gamma_{n+j}(\bar{G}, G)$ $(j = 1, \cdots, h)$

3. LSTM Prediction Algorithm

The L unfolded stages version of the LSTM prediction framework is illustrated in Figure 2 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 $(j = n - L + 1, \cdots, n)$;
- The feed forward network layer: it evaluates from the output of the last LSTM layer the h steps ahead of the predicted bandwidth values $\hat{b}_{n+j}$ $(j = 1, \cdots, h)$ stored in the vector $\hat{\mathbf{b}}_{n,h}$.

The VNF-FG bandwidth predictions are performed by the LSTM layer, which has as inputs the VNF-FG bandwidth values b_j $(j = n - L + 1, \cdots, n)$. The output $\mathbf{h}_n$ is processed by a feed forward neural network, which provides an evaluation of the vector $\hat{\mathbf{b}}_{n,h}$ of predicted VNF-FG bandwidth values.

In the LSTM layer the state variable $\mathbf{s}_j$ $(j = n - L + 1, \cdots, n)$ is also updated. In the LSTM Cell Block LCB_j, the state variable $\mathbf{s}_j$ in the j-TI depends on the following variables: (i) the VNF-FG bandwidth value b_j; (ii) the output $\mathbf{h}_{j-1}$ in the $(j-1)$-th TI; (iii) the state variable $\mathbf{s}_{j-1}$ in the $(j-1)$-st TI.

The operation mode of a single LCB is well known because LSTM has been applied in many fields (handwriting recognition, speech recognition, power consumption prediction, etc.)[16]. However, the training of the LSTM recurrent neural network is performed by minimizing the symmetric loss function.

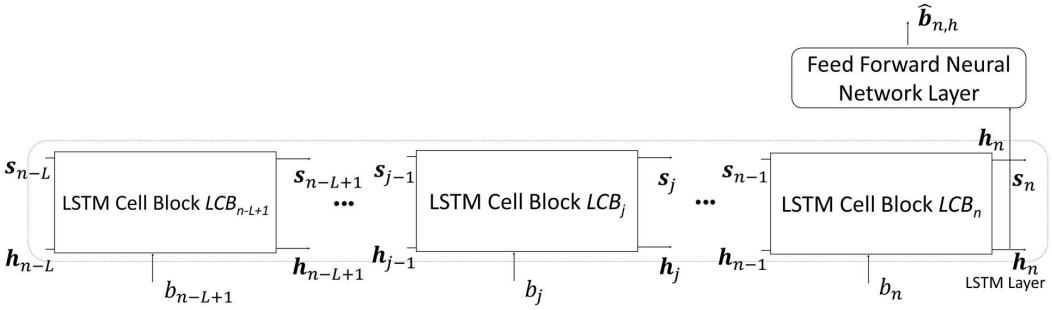

Figure 2. Long Short Term Memory (LSTM) Prediction Framework.

At the same time all of the prediction-based resource allocation algorithms in NFV environments aims at exactly forecasting either the traffic [15] or the resources [12] to be allocated. They are based on the minimization of symmetric cost functions of the errors $e_{n+j} = b_{n+j} - \hat{b}_{n+j}$ $(j = 1, \cdots, h)$ where b_{n+j} is the real VNF-FG bandwidth value in the $(n+j)$-st TI. Examples of these functions are the Mean Squared Error (MSE) or the Mean Absolute Error (MAE). The choice of symmetric cost functions leads to equally weigh 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 a cusp linear loss function as represented in Figure 3 where the slopes are dependent on the resources allocation cost C_{RA} and QoS degradation cost C_{QoS} both defined in \$ per Gbit. As reported in Figure 3 the training process minimizes the Asymmetric Mean Absolute Error $AMAE_{n,h}$ expressed by:

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

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

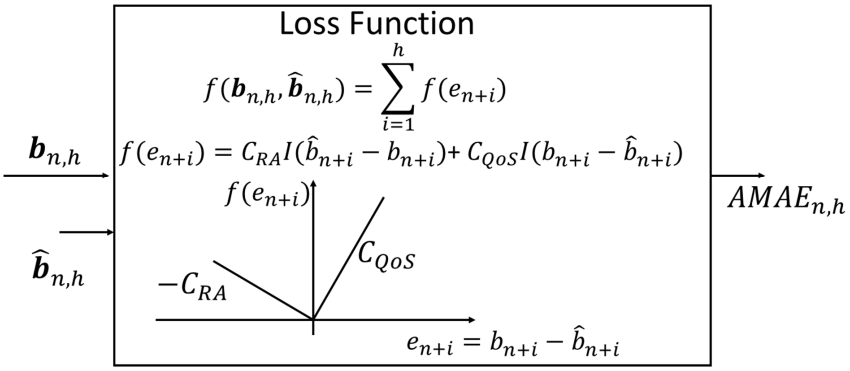

Figure 3. Definition of the asymmetric loss function.

4. Numerical Results

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

We provide some results in the case of the Deutsche Telekom (DT) network reported in Figure 4. The network is composed of 14 switches and 24 links. The main input parameters, their description and the values range are reported in Table 4.

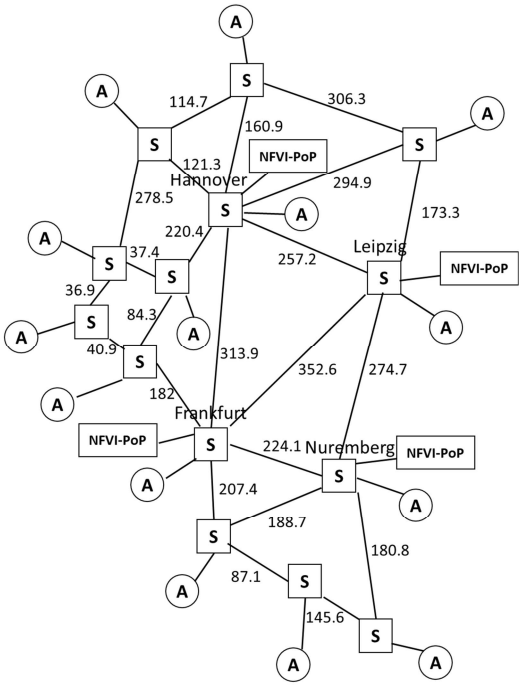

Figure 4. Deutsche Telekom network topology. The distances are expressed in km.

Table 4. Input parameters.

B_L	Link Bandwidth	30 Gbps
B_{NP}	Number of NFVI-PoPs	4
$N_{\bar{V}}$	Number of cores in any NFVI-PoP	48
c_{av}^{core}	Average Core Cost	1 \$/h
ς	Core cost unbalancing factor	1.4
T_s	Time duration of a TI	1 h
C_{RA}	Cloud Resource Allocation cost (\$ to be paid for one traffic Gbit)	0.025 \$/Gbit
C_{QoS}	QoS degradation cost (\$ to be paid for one lost traffic Gbit)	0.00025–2.5 \$/Gbit

The cloud resources are placed in $N_{NP} = |\bar{V}_{NP}| = 4$ NFVI-PoPs located in the cities of Hannover, Leipzig, Frankfurt and Nuremberg. Each NFVI-PoP is equipped with $N_{\bar{v}} = 48$ cores. The core costs of the NFVI-PoPs are randomly chosen among the values $c_{NP_i}^{core} = \varsigma^i C_0$ $(i = 0, \cdots, N_{NP} - 1)$ [8,23] where C_0 is a normalization cost and the parameter ς characterizes the unbalancing of the core costs in the different NFVI-PoPs. In particular the cost unbalancing is high, as the parameter ς is higher. For $\varsigma = 1$ we achieve equal costs and the balanced case. We will carry out the analysis when the average core cost c_{av}^{core} is fixed to 1 \$/h. The knowledge of c_{av}^{core} leads to the normalization value $C_0 = N_{NP} * c_{av}^{core} \frac{1-\varsigma}{1-\varsigma^{N_{NP}}}$. Next we provide the results in the case of $\varsigma = 1.4$.

We assume the link bandwidth B_L equals 30 Gbps.

We consider four SFs: Firewall (FW), Intrusion Detection System (IDS), Network Address Translator (NAT) and Proxy. VNFIs can be instantiated in the NFVI-PoPs to support these SFs. They are supported by software modules characterized by the maximum processing capacities 900, 600, 900 and 600 Mbps with the number of allocated cores equal to respectively 4, 8, 2 and 4. We consider linear VNF-FG of the same type and composed by one FW, one IDS, one NAT and one Proxy.

The choice of the core costs and processing capacities leads to a cloud resource allocation cost C_{RA} = 0.025 \$/Gb for the VNF-FG considered.

We assume that one VNF-FG is established for each tuple of access nodes reported in Figure 4. As VNF-FG bandwidth values, we consider the real traffic values measured at hourly intervals and reported in [24]. These values are used to forecast the future traffic values according to the procedure illustrated in Section 3. We evaluate the operation cost of the NFV network reported in Figure 4 when the resource allocation is based on predicted rather than real VNF-FG bandwidth values. In particular we have evaluated the cost for the period from 21 June 2004 to 25 June 2004 [24] by predicting the VNF-FG bandwidth values requested between all of the tuples of access nodes of Figure 4. Because real traffic is not available for the Deutsche Telecom network, we have used the ones available in [24] for other networks with similar size. The predicted values are evaluated by applying the proposed traffic forecasting algorithm and from the knowledge of the real requested VNF-FG bandwidth values from 31 May 2004 to 20 June 2004 [24]. 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 [16]: (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 to 20; that is, the LSTM training process is executed 20 times to find the best model to perform forecasting.

We assume the cloud and bandwidth resource allocation is performed by executing two algorithms that have been proposed by the authors in a previous paper [8]. It has been shown how these algorithms, referred to as Least Cloud Resource and Bandwidth (LCBC) and Deployment Costs Aware (DCA) [8], perform well in allocating resources for NFV networks and allows for an operation cost optimization when real traffic data are known.

We report the cost in Figure 5 when the parameter w, defined as the ratio of the Resource Allocation cost C_{RA} to the QoS degradation C_{QoS}, is varied from 0.0001 to 1 and the resource allocation is performed in the cases in which the MSE and Asymmetric (ASYM) LSTM prediction techniques are used. Notice that because C_{RA} is fixed to 0.025 \$/Gb, the variation of w is obtained by varying C_{QoS} from 250 to 0.025 \$/Gb. In particular we study a case of interest in which the QoS degradation cost C_{QoS} is higher than or equal to the cloud resource allocation cost C_{RA}. The results of Figure 5 have been achieved in the case of prediction step h equal to 1, 12 and 24.

From the results reported in Figure 5 we can make the following remarks:

- 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 equal only for $w = 1$, that is, when the over-provisioning and under-provisioning costs are equal; as a matter of example, the total costs of the MSE and ASYM solutions for $w = 0.04$ and $h = 12$ are 134 \$ and 96 \$ with 28% cost advantage of our proposed asymmetric LSTM prediction solution;
- The better performance in total cost of the asymmetric prediction solution for w lower than 1 is due to the fact that it reduces the resource under-provisioning periods and consequently, the costs due to the QoS degradation.

To justify the results we also report in Figure 6 the predicted VNF-FG bandwidth values from 21 June 2004 to 25 June 2004 [24] for the traffic offered between two access nodes of the Deutsche Telekom network. In particular we compare the real, MSE and ASYM LSTM predicted traffic for values of w equal to 0.1 and prediction steps equal to 1, 12 and 24 in Figure 6a–c respectively. From these figures 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 the time series values because of the higher QoS degradation costs; conversely the ASYM LSTM predictions allows for a correct operation mode by overestimating the predicted values.

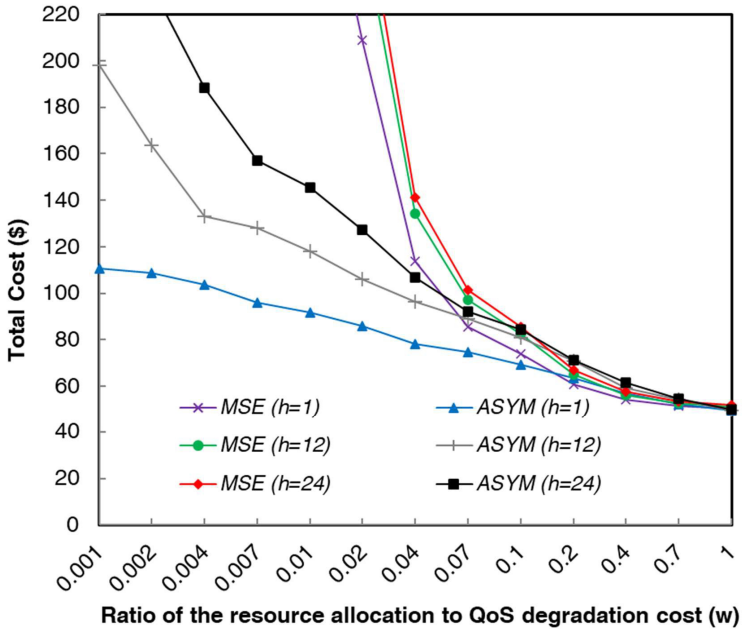

Figure 5. Cost in allocating resources for the NFV network of Figure 4 when w varies from 0.0001 to 1. The total cost is reported when the allocation algorithms use the MSE and ASYM LSTM predicted SFC bandwidth values and prediction steps h equal to 1, 12 and 24.

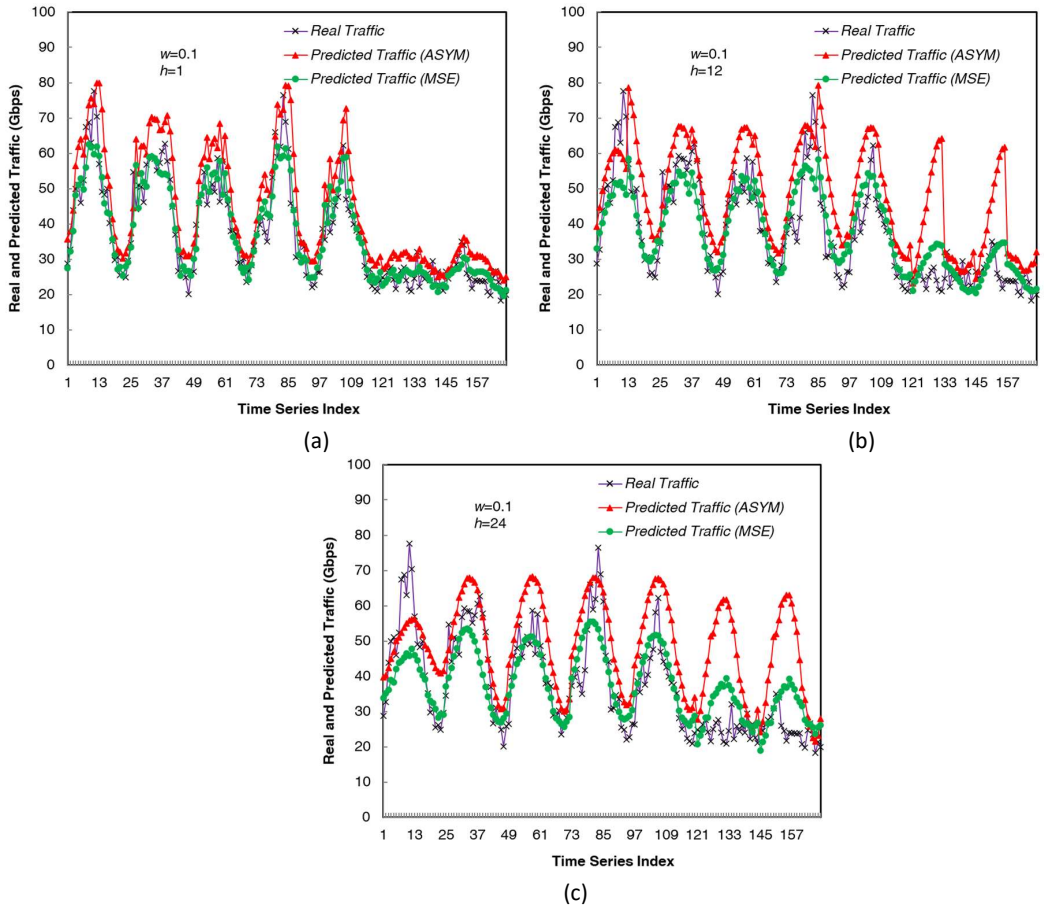

Figure 6. Comparison of the Real, Mean Square Error (MSE) and Asymmetric (ASYM) LSTM prediction values for w = 0.1 and prediction step h equal to 1 (**a**), 12 (**b**) and 24 (**c**).

These results are confirmed in Figure 7 where we report the total cost as a function of the prediction step h for values w varying from 0.01 to 1.

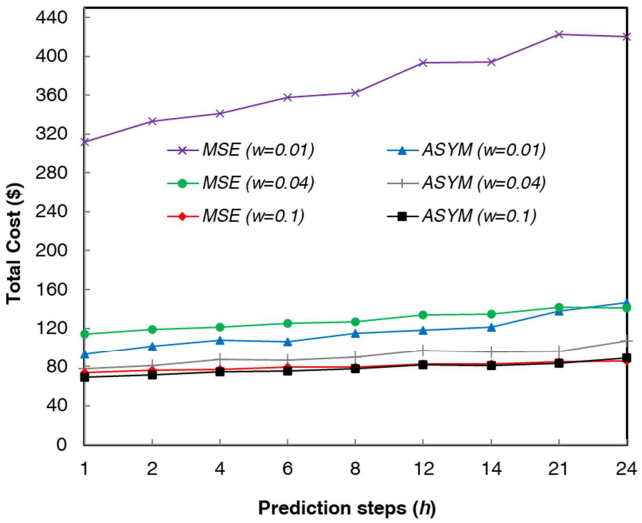

Figure 7. Total cost in allocating resources for the NFV network of Figure 4 as a function of the prediction step h and when w varies from 0.01 to 1. The results are reported for the MSE and ASYM LSTM prediction solutions.

5. Conclusions

We have developed a traffic forecasting algorithm for the allocation of resources in NFV environments that can differently weigh the over-provisioning and under-provisioning costs. The proposed solution is inherited from the classical LSTM prediction algorithm and it is based on minimizing an asymmetric cost function of the prediction error. The use of the prediction technique proposed in an NFV network scenario with the interconnection of four NFVI-PoPs has led to cost advantages by 40% compared to prediction techniques based on minimizing the symmetric cost functions of the prediction error.

Author Contributions: Methodology, V.E. and T.C.; Software, F.G.L., T.C. and P.J.P.S.; Writing—original draft, V.E.; Writing—review & editing, V.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Barakabitze, A.A.; Ahmad, A.; Mijumbi, R.; Hines, A. 5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges. *Comput. Netw.* **2020**, *167*, 1–40. [CrossRef]
2. Pei, J.; Hong, P.; Pan, M.; Liu, J.; Zhou, J. Optimal VNF Placement via Deep Reinforcement Learning in SDN/NFV-Enabled Networks. *IEEE J. Sel. Areas Commun.* **2020**, *38*, 263–278. [CrossRef]
3. Farkiani, B.; Bakhshi, B.; MirHassani, S.A. A Fast Near-Optimal Approach for Energy-Aware SFC Deployment. *IEEE Trans. Netw. Serv. Manag.* **2019,** *16*, 1360–1373. [CrossRef]
4. Yu, Y.; Bu, X.; Yang, K.; Nguyen, H.K.; Han, Z. Network Function Virtualization Resource Allocation Based on Joint Benders Decomposition and ADMM. *IEEE Trans. Veh. Technol.* **2020**, *17*, 622–636. [CrossRef]
5. Eramo, V.; Lavacca, F.G. Computing and Bandwidth Resource Allocation in Multi-Provider NFV Environment. *IEEE Commun. Lett.* **2018**, *22*, 2060–2063. [CrossRef]
6. Karimzadeh-Farshbafan, M.; Shah-Mansouri, V.; Niyato, D. Reliability Aware Service Placement Using a Viterbi-Based Algorithm. *IEEE Trans. Netw. Serv. Manag.* **2020**, *17*, 1706–1708. [CrossRef]

7. Eramo, V.; Miucci, E.; Ammar, M.; Lavacca, F.G. An Approach for Service function Chain Routing and Virtual Function Network Instance Migration in Network Function Virtualization Architectures. *IEEE ACM Trnsa. Netw.* **2017**, *25*, 2008–2025. [CrossRef]
8. Eramo, V.; Lavacca, F.G. Optimizing the Cloud Resources, Bandwidth and Deployment Costs in Multi-Providers Network Function Virtualization Environment. *IEEE Access* **2017**, *7*, 46898–46916. [CrossRef]
9. Karimzadeh-Farshbafan, M.; Shah-Mansouri, V.; Niyato, D. A Dynamic Reliability-Aware Service Placement for Network Function Virtualization (NFV). *IEEE J. Sel. Areas Commun.* **2020**, *38*, 318–333. [CrossRef]
10. Ma, W.; Beltran, J.; Pan, D.; Pissinou, N. Placing Traffic-Changing and Partially-Ordered NFV Middleboxes via SDN. *IEEE Trans. Netw. Serv. Manag.* **2019**, *16*, 1303–1317. [CrossRef]
11. Kim, H.N.; Lee, D.; Jeong, S.; Choix, H.; Yoo, J.; Hong, J.W. Machine Learning-Based Method for Prediction of Virtual Network Function Resource Demands. In Proceedings of the 2019 IEEE Conference on Network Softwarization (NetSoft), Paris, France, 24–28 June 2019.
12. Schneider, S.; Satheeschandrany, N.P.; Peuster, M.; Karl, H. Machine Learning-Based Method for Prediction of Virtual Network Function Resource Demands. In Proceedings of the 2020 IEEE Conference on Network Softwarization (NetSoft), Ghent, Belgium, 29 June–3 July 2020.
13. Bega, D.; Gramaglia, M.; Fiore, M.; Banchs, A.; Costa-Perez, X. DeepCog: Optimizing Resource Provisioning in Network Slicing with AI-Based Capacity Forecasting. *IEEE J. Sel. Areas Commun.* **2020**, *38*, 361–376. [CrossRef]
14. Soualah, O.; Mechtri, M.; Ghribi, C.; Zeghlache, D. Online and batch algorithms for VNFs placement and chaining. *Comput. Netw.* **2019**, *158*, 98–113. [CrossRef]
15. Li, B.; Lu, W.; Liu, S.; Zhu, Z. Deep-Learning-Assisted Network Orchestration for On-Demand and Cost-Effective vNF Service Chaining in Inter-DC Elastic Optical Networks. *IEEE J. Opt. Commun. Netw.* **2018**, *10*, D29–D41. [CrossRef]
16. Kong, W.; Dong, Z.Y.; Jia, Y.; Hill, D.J.; Xu, Y.; Zhang, Y. Short-Term Residential Load Forecasting Based on LSTM Recurrent Neural Network. *IEEE Trans. Smart Grid* **2019**, *10*, 841–851. [CrossRef]
17. Farahnakian, F.; Pahikkala, T.; Liljeberg, P.; Plosila, J.; Hieu, N.T.; Tenhunen, H. Energy-Aware VM Consolidation in Cloud Data Centers Using Utilization Prediction Model. *IEEE Trans. Cloud Comput.* **2019**, *7*, 524–536. [CrossRef]
18. Han, J. *Data Mining: Concepts and Techniques*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2005.
19. Ferrer-Troyano, F.J.; Aguilar-Ruiz, J.S.; Riquelme, J.C. Empirical Evaluation of the Difficulty of Finding a Good Value of k for the Nearest Neighbor. In Proceedings of the 2013 International Conference Computing Science 2003, San Diego, CA, USA, 10–12 December 2003.
20. Eramo, V.; Catena, T.; Lavacca, F.G.; Giorgio, F.D. Study and Investigation of SARIMA-based Traffic Prediction Models for the Resource Allocation in NFV networks with Elastic Optical Interconnection. In Proceedings of the 2020 22nd International Conference on Transparent Optical Networks (ICTON), Bari, Italy, 19–23 July 2020.
21. Yilma, G.M.; Yousaf, Z.F.; Sciancalepore, V.; Costa-Perez, X. Benchmarking open source NFV MANO systems: OSM and ONAP. *Comput. Commun.* **2020**, *161*, 86–98. [CrossRef]
22. Trakadas, P.; Karkazis, P.; Leligou, N.; Zahariadis, T.; Vicens, F.; Zurita, A.; Alemany, P.; Soenen, T.; Parada, C.; Bonnet, J.; et al. Comparison of Management and Orchestration Solutions for the 5G Era. *J. Sens. Actuator Netw.* **2020**, *9*, 4. [CrossRef]
23. Eramo, V.; Lavacca, F.G. Proposal and Investigation of a Reconfiguration Cost Aware Policy for Resource Allocation in Multi-Provider NFV Infrastructures Interconnected by Elastic Optical Networks. *J. Light. Technol.* **2019**, *37*, 4098–4114. [CrossRef]
24. SND-Lib. Available online: http://sndlib.zib.de/home.action (accessed on 11 November 2020).