Profit-Aware Proactive Slicing Resource Provisioning with Traffic Uncertainty in Multi-Tenant FlexE-over-WDM Networks

Qize Guo*, Zhao Ming*, Hao Yu*, Yan Chen*, and Tarik Taleb†
*Faculty of Information Technology and Electrical Engineering, University of Oulu, Oulu, 90570 Finland
†Ruhr University Bochum, Bochum, 44801 Germany
{qize.guo, zhao.ming, hao.yu, yan.chen}@oulu.fi, tarik.taleb@rub.de

***Abstract*—Addressing the pressing requirement for dynamic and intelligent allocation of slicing resources, the dynamic provisioning of resources based on traffic predictions has emerged. Although this method favours proactive scheduling of network slices, more complexities are introduced by the prediction uncertainty. In addition, because multi-tenant networks are always changing in terms of technology and business model, profit-aware network slicing is becoming an important topic of study in the field of resource provision. This paper focuses on profit-aware slicing resource provisioning amid traffic uncertainty in multi-tenancy flexible Ethernet over wavelength division multiplexing networks. Specifically, we develop a profit model for multi-tenant network slicing, accounting for the impact of network prediction uncertainty, and formulate the problem as maximizing the profit of users primarily. To solve this problem, we propose a profit-aware resource provisioning approach that first checks if the slice requests are made by pruning algorithms and then determines the service relationship between slices and tenants by matching games. Simulation results demonstrate the superiority of the proposed algorithm over benchmarks in terms of user profit, total benefit, and denial ratio of service.**

***Index Terms*—Resource Provision, Multi-tenant, Network Slicing, and Prediction Uncertainty**

I. Introduction

The commercialization of 5G technology and the ongoing research on B5G and 6G have spawned a variety of emerging applications ranging from the consumer to the industrial sectors, such as the Internet of Vehicles, augmented reality/virtual reality, and smart manufacturing [1]–[3]. On the other hand, the rapid expansion of network application scenarios also leads to a significant increase in the number and diversity of network slices, increasing the importance of customised and intelligent management of these slices. Within the market innovation framework driven by the multi-tenancy model, profits gained through network slicing now significantly influence the resource provisioning of these slices [4], [5].

To deal with the dynamic traffic patterns of network slicing services, there is a growing interest in proactive network slicing mechanisms that enable dynamic network resource provisioning based on network slicing traffic predictions [6], [7]. However, the inherent uncertainty in network traffic prediction poses challenges. Strict adherence to prediction results when allocating network resources may lead to a decline in the quality of service or interruptions in network slicing [8]. Consequently, prediction-based resource allocation schemes generally incorporate over-provisioning to prevent slice degradation or blockages [9], [10]. Nevertheless, the profitability of slicing is still potentially impacted.

In multi-tenant networks, tenants lease network resources from network providers and rent them to users to provide slicing services. The primary drivers shaping network role behaviours are cost reduction and profit enhancement. Consequently, there is dedicated research aimed at formulating resource allocation strategies that maximize the profits of network roles [11]–[13]. However, the strategies with prediction uncertainty in proactive network slicing are not taken into account. However, existing studies on profit-based resource allocation have not considered the impact of prediction uncertainty with proactive network slicing.

Meanwhile, the current Flexible Ethernet (FlexE) over wavelength division multiplexing (WDM) transport network, employing FlexE and WDM/OTN, stands out as a leading technological solution for transport networks in the 5G era [14]. The development of a multi-tenancy mechanism is seamlessly achievable within the FlexE and WDM/OTN frameworks, thanks to the inherent support of FlexE for network slicing [15]. The distinctive technical characteristics of optical transport networks introduce variations in the handling and distribution of slices, setting them apart from radio access networks and core networks.

In this paper, we have analysed profit-aware proactive slicing resource provisioning mechanisms and explored the methodologies and optimisation approaches used in slicing resource provisioning based on the profit of users in multi-tenancy FlexE-over-WDM networks. Additionally, we evaluate the impact of prediction uncertainty on the profits of network roles. Overall, the contributions of this paper are summarised as follows:

This research work is partially supported by the European Union's Horizon 2020 Research and Innovation Program through the aerOS project under Grant No. 101069732; the Business Finland 6Bridge 6Core project under Grant No. 8410/31/2022; the European Union's HE research and innovation program HORIZON-JUSNS-2022 under the 6GSandbox project (Grant No. 101096328); and the Research Council of Finland 6G Flagship Programme under Grant No. 346208. This research was also partially conducted at ICTFICIAL Oy. The paper reflects only the authors' views, and the European Commission bears no responsibility for any utilization of the information contained herein.

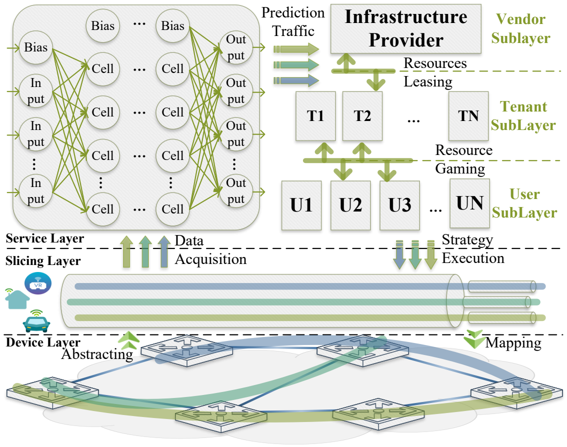

Fig. 1. A general network architecture of slicing resource provisioning in Multi-Tenant FlexE-over-WDM Networks.

- We propose an architecture for slicing resource provisioning and develop a profit-aware model with taking into account the impact of network prediction uncertainty in proactive dynamic slicing in FlexE-over-WDM network.
- We formulate the problem as maximizing the profit of users and propose a profit-aware resource provisioning framework to solve this problem.
- Simulation results demonstrate that our proposed scheme outperforms other baselines in improving user profit and total benefit and reducing the denial ratio of service.

II. System Model and Problem Formulation

In this section, we first introduce the considered system model and then formulate the optimization problem.

A. System Model

As illustrated in Fig. 1, the network infrastructure comprises devices in the device layer. Tenants rent resources from network providers to create network slices tailored to the diverse services they offer to users. To optimize network resource utilization, a dynamic slicing adjustment strategy based on predictions is employed. The neural network gathers and predicts the traffic of network slices, generating corresponding resource provisioning strategies for multi-tenancy. Users can choose network slicing from different tenants based on revenue assessments. At each time interval t, resource provisioning dynamically adjusts based on prediction results and the tenant-user relationship. If the proactive provisioned resources cannot meet the actual service requirements, tenants and users may incur additional costs. These costs include additional resources and dispatch costs to expand slice capacity or the cost of service dispatch. This paper focuses on resource provisioning among different network roles based on slicing profits.

We define some operational symbols, with $*(t)$ representing the value of * at time interval t, and use $o(x,y) = \lceil x/y \rceil$ to represent the upward rounding of x/y. The infrastructure of the FlexE-over-WDM network can be denoted as $\mathbf{G} = (\mathbf{V}, \mathbf{E})$. FlexE switches are distributed across network nodes $\mathbf{V}$:$\{V_n, n \in [1, N_n]\}$ to facilitate inbound and outbound traffic for network slicing. The links $\mathbf{E}:\{E_{mn}, \forall m, n \in \mathbf{V}\}$ are realized by wavelength of WDM, the bandwidth is $\mathbf{B}:\{B_{mn}, \forall m, n \in \mathbf{V}\}$, and the transport delay is denoted by $\boldsymbol{\tau}:\{\tau_{m,n}, \forall m, n \in \mathbf{V}\}$. Simultaneously, the number of slots that can be scheduled per FlexE switch are denoted by $\mathbb{N}^f : \left\{\mathbb{N}_n^f, \forall n \in [1, N_n]\right\}$. Each slot has a bandwidth of b_f. The wavelengths of the links are represented by $\mathbb{N}^\lambda : \left\{\mathbb{N}_{mn}^\lambda, \forall m, n \in \mathbf{V}\right\}$, with each wavelength having a bandwidth of b_λ.

In the context of multi-tenancy, the tenants are denoted by $\mathbf{T}$:$\{T_k, k \in [1, N_t]\}$, and the corresponding tenant network of T_k is represented as $\mathbf{G}^k = \left(\mathbf{V}^k, \mathbf{E}^k\right)$. The number of FlexE slots and wavelengths for T_k are represented by N_k^f and $N_{mn}^{\lambda,k}$, $\forall m, n \in \mathbf{V}$. The services from users are denoted by f_i:$\{s_i, d_i, b_i, \tau_i, R_i\}, i \in [1, N]$, with elements ordered as the source node, destination node, bandwidth requirement, latency requirement, and the set of transport paths of f_i. The FlexE granularity offered by the network provider to the tenants is a single slot, while the wavelength granularity is a single wavelength. The granularity offered by the tenant to the user is based on bandwidth, denoted as δ_b.

Regarding traffic prediction, we use x_i and q_i to represent the actual and predicted traffic of f_i. To assess the usability of predictions during resource provisioning, each slice's prediction is labeled to indicate credibility, denoted as $\mathbf{A}:\{A_i\}$, where $A_i = 1$ means the predicted traffic of f_i is credible. Conditions under which slice predictions are credible or not have been discussed in previous studies [6]. We define $T_a^i \in [0, 1]$ as resource over-provisioning factor to realize the level of resource over-provisioning based on prediction results. If the predicted result is credible, the resource provisioning strategy is based on the predictions; otherwise, it is based on b_i, as shown in (1). Additionally, we use Δb_i to represent the difference in provisioned resources for f_i after and before emergency resource adjustment.

$$b_i' = \begin{cases} o\left(\max\left(\frac{q_i(t)}{1-T_a^i}, b_i\left(t\right)\right), \delta_b\right), & A_i\left(t-1\right) = 1 \\ b_i'' = o\left(b_i\left(t\right), \delta_b\right), & A_i\left(t-1\right) = 0 \end{cases}. \quad (1)$$

Regarding the price at resource provisioning, we use $\left\{C_{cost}^k, C_r^k, C_m^k, C_{rf}^k, C_{r\lambda}^k, C_e^k, C_{el}^k\right\}$ to represent the cost for T_k. The elements are ordered as the total cost to serve the services, the resource rental cost from the network provider,and the service management cost of tenant services, the scheduling cost of FlexE slots, the scheduling cost of wavelength, and use C_v^i to represent the additional dispatch cost of f_i during service time. Furthermore, let $\mathbf{p}$=$\{p_f, p_\lambda, p_m, p_t, p_u\}$ represent the unit price of resources, with elements ordered as the price of FlexE slots, the wavelength, the management price of the slices, and the unit price of bandwidth from tenants and users. Let $\boldsymbol{\eta}$=$\{\eta_\lambda, \eta_f, \eta_i, \eta_d\}$ represent the price factor of wavelength and FlexE slot costs for emergency adjustments of tenants, the price factor of bandwidth costs for emergency adjustments of f_i, and the price factor of bandwidth costs for service dispatch. Moreover, let $\mathbf{U}$=$\left\{U_t^k, U_u^i\right\}, \forall k \in [1, N_t], i \in [1, N]$ denotes the profit, with its elements representing the profit of T_k and f_i, respectively. We propose two conditions that are assumed for realism. First, due to service competition of T_k, profit margins can only be maintained at $\mathcal{K}$. The actual transport latency of f_i is denoted as $\tau_i{}'$. The value of the services is influenced by τ_i' with the factor γ_i, i.e.,

$$\gamma_i = \begin{cases} \exp\left(\frac{\tau_i - \tau_i'}{\tau_i}\right), & \text{if} \quad \tau_i' \leq \tau_i, \\ 0, & \text{otherwise.} \end{cases}, \tag{2}$$

which indicates the delay has exponential impacts on the value of slices. We define the slice management cost of T_i as

$$C_m^i = p_m \left(\frac{\sum_k I_{ki}}{N_c^i}\right)^2, \tag{3}$$

where N_c^i represents the tenant's management capacity. and I_{ki} indicates the service relationship. that $I_{ki} = 1$ when f_i is served by tenant T_j, and $I_{ki} = 0$ otherwise.

This cost model indicates that as more service slices are added, the overhead will grow quadratically. We define the value of slice f_i as

$$v_i = \gamma_i p_u \ln\left(1 + x_i\right). \tag{4}$$

Additionally, we use Θ to indicate whether the resources are sufficient for new slices, with $\Theta = 1$ indicating insufficient resources for any requested slice, and use $\Omega : \{\Omega_{ij}\}$ to indicate whether resources can be extended when the actual traffic requirement of f_i exceeds the original provisioned amount at the beginning of the time interval.

Before delving into dynamic resource provisioning, we first formulate the model as follows: For tenants, the number of FlexE slots on node V_m and wavelengths on link E_{mn} leased to T_k can be calculated as

$$\begin{aligned} N_{k,m}^f &= o\left(\sum_{f_i \in \mathbf{F}} \sum_{V_m \in \{s_i, d_i\}} b_i' I_{ki}, b_f\right), \\ N_{k,mn}^\lambda &= o\left(\sum_{f_i \in \mathbf{F}} \sum_{E_{mn} \in R_i} b_i' I_{ki}, b_\lambda\right). \end{aligned} \tag{5}$$

FlexE slots and wavelengths change dynamically with dynamic resource provisioning. We use ΔN_k^f and ΔN_k^λ to indicate the difference in slots and wavelengths for T_k during the adjacent time intervals. And use δN_k^f and δN_k^λ to denote their difference at emergency resource adjustments.

The rental resource cost of T_k can be obtained by

$$C_r^k = C_{rf}^k + C_{r\lambda}^k, \tag{6}$$

where C_{rf}^k and $C_{r\lambda}^k$ can be derived as

$$C_{rf}^k = \begin{cases} p_f \sum_{V_m \in \mathbf{V}^k} N_{k,m}^f, & \text{if} \quad t = 0, \\ p_f \sum_{V_m \in \mathbf{V}^k} \Delta N_{k,m}^f, & \text{otherwise.} \end{cases}, \tag{7}$$

and

$$C_{r\lambda}^k = p_\lambda \sum_{E_{mn} \in \mathbf{E}^k} N_{k,mn}^\lambda. \tag{8}$$

Simultaneously, C_e^k and C_{el}^k under emergency adjustment can be calculated as

$$\begin{aligned} C_e^k &= \eta_f \delta N_k^f + \eta_\lambda \delta N_k^\lambda, \\ C_{el}^i &= \eta_d \left(b_i'' - b_i'\right). \end{aligned} \tag{9}$$

Subsequently, the profit of f_i can be calculated by

$$U_u^i = v_i - p_s^i b_i', \tag{10}$$

and the profit of T_k can be obtained as

$$U_t^k = C_s^k - C_{cost}^k = \sum p_s^i b_i' I_{ki} - C_{cost}^k. \tag{11}$$

In the context of dynamic resource provisioning, slices exist in credible and not credible states. Concurrently, there are three possibilities for f_i, denoted by α_i, during resource provisioning at time duration t: fulfillment, scalability, and non-scalability, denoted by $\mathscr{F}$, $\mathscr{S}$, and $\mathscr{N}$, respectively.

i) Fulfillment: The actual traffic of f_i does not exceed b_i', indicating that the resources need no further modifications during the t-th time interval.

ii) Scalability: If the actual traffic of f_i exceeds b_i', and the provisioned resources can be extended in an emergency, the tenant and user must expand resources at an additional cost. Alternatively, they may choose to forgo the extension of resources, resulting in $\alpha_i = \mathscr{N}$, a decision influenced by profit considerations.

iii) Non-scalability: If the actual traffic of f_i exceeds b_i', and the provisioned resources cannot be extended due to insufficient resources or user decisions, the user of f_i must bear the loss of dispatch.

The C_{cost}^k of T_k and the profit of f_i, i.e., U_u^i need to be updated in dynamic resource provisioning as

$$\begin{aligned} C_{cost}^k = \sum_{\alpha_i = \mathscr{F}} \left(C_r^k + C_m^k\right) + \sum_{\alpha_i = \mathscr{S}} \left(C_r^k + C_m^k + C_e^k\right) \\ + \sum_{\alpha_i = \mathscr{N}} \left(C_r^k + C_m^k + C_{el}^k\right) \end{aligned} \tag{12}$$

and

$$U_u^i = \begin{cases} v_i - p_s^i b_i', & \alpha_i = \mathscr{F} \\ v_i - p_s^i b_i'' - \eta_i \Delta b_i, & \alpha_i = \mathscr{S} \\ v_i - p_s^i b_i' - \eta_d b_i, & \alpha_i = \mathscr{N} \end{cases}. \tag{13}$$

B. Problem Formulation

We aim to maximize the profit of the users by optimizing the service relations of tenants and users ($\mathbf{I}$ and R_i) while satisfying the constrains of the model. Thus, and the optimization problem can be formulated as

$$\max_{\mathbf{I}, R_i} \sum_{f_i \in \mathbf{F}} U_u^i, \tag{14a}$$

$$s.t. \sum_{T_k \in \mathbf{T}} I_{ki} \leq 1, \quad \forall f_i \in \mathbf{F}, \tag{14b}$$

$$\sum_{T_k \in \mathbf{T}} N_{k,m}^f \leq \mathbb{N}^f, \quad \forall V_m \in \mathbf{V}, \tag{14c}$$

$$\sum_{T_k \in \mathbf{T}} N_{k,mn}^\lambda \leq \mathbb{N}^\lambda, \quad \forall E_{mn} \in \mathbf{E}, \tag{14d}$$

$$U_t^k = \mathcal{K} C_{cost}^k, \quad \forall T_k \in \mathbf{T}, \tag{14e}$$

Here, constraint (14b) ensures one service can only be served by one tenant at one time; constraints (14c) and (14d) ensure that the provisioned resources of FlexE slots and wavelengths cannot exceed the capacity of the network, respectively; constraint (14e) ensures the profitability of the tenants.

III. Profit-Aware Resource Provisioning

In this section, we introduce a profit-aware resource provisioning framework for multi-tenancy FlexE-over-WDM networks, aiming to maximize the total profit of all users.

A. Profit-Aware Resource Provisioning Analysis

As the state $\mathscr{N}$ is unpredictable. we first consider the case where $\alpha_i = \mathscr{F}$ and $\alpha_i = \mathscr{S}$. The probability of α_i to be $\mathscr{F}$ can be obtained based on the historical statistics of the prediction, denoted by π_i. Then, U_u^i can be redefined as

$$\bar{U}_u^i = \pi_i \left(v_i - p_s^i b_i'\right) + \left(1 - \pi_i\right) \left(v_i - p_s^i b_i'' - \eta_i \Delta b_i\right), \quad (15)$$

Simultaneously, b_i' can be expressed as

$$\bar{b}_i' = \pi_i b_i' + \left(1 - \pi_i\right) b_i'', \quad (16)$$

Considering the predicted traffic, the scenario where $\alpha_i = \mathscr{N}$ is unexpected. Further discussion on the criteria for such a situation will be addressed later.

In light of this, we establish the value of Ω_i based on the anticipated bandwidth of the service dispatch. Given that the network offers FlexE slots and wavelengths, two conditions must be met by f_i to set $\Omega_i = 0$.

i) Slots: At node V_k, the total number of FlexE slots occupied by all serviced slices, including both entering and exiting slices, is greater than the total number of FlexE slots provided by the node, as shown in (17).

$$\sum_{T_k \in \mathbf{T}} \sum_{f_i \in \mathbf{F}} \sum_{V_j \in \{s_i, d_i\}} o\left(\bar{b}', \delta_b\right) \geq \mathbb{N}^f, \forall V_n \in \mathbf{V} \quad (17)$$

ii) Wavelengths: The wavelengths occupied by all tenants exceed the capacity of the network. To provide a mathematical expression for this condition, we define $\mathbf{G}_{mn}'$ as the pruning of $\mathbf{G}^k$ by E_{ij}^k. This can be achieved through the following steps: a) Find the k-shortest path (KSP) for each slice with source and destination nodes not being V_m and V_n (or V_n and V_m) from $\mathbf{G}^k$. b) Calculate the least-cost paths for each slice separately using a matching game. c) Add to the path the desired flows $\bar{b}_i'$ of the remaining slices of the network, excluding network slices other than $V_m \rightarrow V_n$ or $V_n \rightarrow V_m$. d) Prune $\mathbf{G}^k$, with the network retaining its existing capacity and only the KSP between V_n and V_m. By following these steps, we obtain $\mathbf{G}_{mn}'$.

We define $Flow(\mathbf{G}_{mn}', V_m, V_n)$ as the maximum flow between V_m and V_n in $\mathbf{G}_{mn}'$. The corresponding condition can be expressed as shown in (18).

$$Flow\left(\mathbf{G}_{mn}', V_m, V_n\right) \leq \sum_{f_k \in \mathbf{F}} \sum_{\substack{(s_k = V_m, d_k = V_n) \\ (s_k = V_n, d_k = V_m)}} b_k' \quad (18)$$

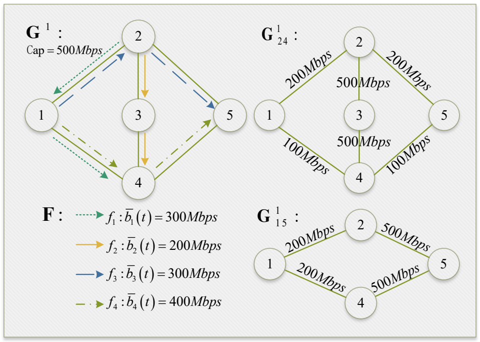

Fig. 2. The DRL-based INC-enhanced task offloading framework in MEC networks.

For better understanding, consider the example shown in Fig. 2. The tenant network $\mathbf{G}^1$ has 5 nodes with a link capacity of $500Mbps$. There are 4 credible slices (f_1 to f_4) with expected bandwidths: $\bar{b}_1 = 300$ Mbps, $\bar{b}_2 = 200$ Mbps, $\bar{b}_3 = 300$ Mbps, and $\bar{b}_4 = 500$ Mbps. Assume that the optimal cost paths obtained according to the matching game are $f_1 : [E_{21}, E_{14}]$, $f_2 : [E_{23}, E_{34}]$, $f_3 : [E_{12}, E_{25}]$, $f_4 : [E_{14}, E_{45}]$, respectively. Following the steps outlined earlier, we obtain $\mathbf{G}_{15}^1$ and $\mathbf{G}_{24}^1$. In $\mathbf{G}_{15}^1$, link E_{12} is allocated to f_1, leaving a remaining bandwidth of $200Mbps$ on this link. Similar capacities for the other links can be determined. There are two eligible paths between V_1 and V_5, resulting in a maximum stream of $400Mbps$. Since this exceeds the sum of $\bar{b}_3'$ and $\bar{b}_4'$, $\Omega_3 = \Omega_4 = 1$. The costing of f_3 and f_4 needs to take into account the cost of dispatch for failing to meet the requirements of their extensions. Similarly, in $\mathbf{G}_{24}^1$, the maximum stream is $700Mbps$, which is greater than the sum of $\bar{b}_1'$ and $\bar{b}_2'$. Thus, $\Omega_1 = \Omega_2 = 0$. When calculating the costs of f_1 and f_2, the costs of their extensions are calculated on the basis that no dispatch will be incurred.

B. Profit-aware Slicing Resource Provisioning Algorithm

Based on the analysis, we propose the Profit-Aware Slicing Resource Provisioning Algorithm with Multi-Tenancy Gaming (PS-MTG), consisting of two sub-algorithms. The Slicing Request Pre-Check Algorithm is primarily used to verify whether the slicing satisfies the conditions mentioned earlier, updating Ω for the slices. Lines 3-8 are used to check the condition of the slots, while Lines 9-21 primarily address the wavelength condition, beginning with the pruning of the tenant network. In the pruning process, Lines 9-15 realise steps a and b. Line 17 is used to calculate the pruning graph, and Lines 18-21 are employed to check whether the wavelength condition holds.

After the pre-check, we obtain the set of slice requests $\mathbf{F}'$ that can be served by tenants. We then design the matching algorithm between slices and tenants, as shown in Algorithm 2. The loop, starting from Line 2 and continuing until the network is congested or all slices are served, is initiated. Lines 3-12 are primarily for match gaming, determining the served tenant for f_k. During the games, user f_i receives quotations from tenants and selects the most profitable one to fulfill the service. Line 7 initially removes f_i from the best relations

Algorithm 1 Pre-Check on Credible-Predicted Slice Requests.

```
Input: G, F, T, $A_i$, $x_i, q_i, \pi_i \forall f_i \in$ F.
 1: Initialize: The set of slices with services $\mathbf{F}' = \mathbf{F}$.
 2: Sorted F by slice value.
 3: for $V_i \in \mathbf{V}$ do
 4:     if (17) holds then
 5:         if $s_k = V_i || d_k = V_i$ then
 6:             Update the cost $C^i_{rf}$ of $f_k$.
 7:     else
 8:         Set $\mathbf{F}' \leftarrow \mathbf{F}' - f_i$, and refuse service $f_k$.
 9: for $f_k \in \mathbf{F}$ do
10:     Calculate KSP for $f_k$ and save them to $R_i$.
11:     if $R_i = \emptyset$ then
12:         Refuse Service $f_i$.
13:     else
14:         Update $C^i_s$.
15:     Categorize slices by source and destination.
16: for each endpoints pair $(s_k = V_m, d_k = V_n)$ do
17:     Calculate $\mathbf{G}^k_{mn}$.
18:     if (18) holds then
19:         Update the cost $C^i_{r\lambda}$ of $f_k$.
20:     else
21:         Set $\mathbf{F}' \leftarrow \mathbf{F}' - f_i$, and refuse service $f_k$.
```

Algorithm 2 Slice-Tenant Matching for Credible Prediction.

```
Input: G, $\mathbf{F}'$, T, $x_i, q_i, \pi_i \forall f_i \in \mathbf{F}'$.
 1: Initialize: Indication of network congestion $\Theta = 0$.
 2: while $|\mathbf{F}| > 0 || \Omega_k = 1, \forall f_k \in \mathbf{F}$ do
 3:     for $f_k \in \mathbf{F}'$ do
 4:         Calculate the expected profit $U^k_u$ of $f_k$.
 5:     Creating match relations between slices and tenants
        through matching games (make I).
 6:     Set $\mathbf{F}' \leftarrow \mathbf{F}' - f_i$, and update $\mathbf{G}^m$.
 7:     if $\Theta(t) = 0$ (The network is not congested) then
 8:         for $f_i \in \mathbf{F}'$ do
 9:             if $\mathbf{G}'_i$ congested after serve $f_i$ then
10:                 Set $\Theta(t) = 1$.
11:     if $\Theta(t) = 1$ then
12:         Run Algorithm. 1 with $\mathbf{F} = \mathbf{F}'$.
13: Update the cost, profit and path of $\mathbf{F}'$..
```

in the matching game. If the tenant network $\mathbf{G}^n$ becomes congested after serving f_i (i.e., $\Theta(t) = 1$), indicating the need for resource provisioning changes, Algorithm 1 is executed to update all tenants with the latest resource information and costs (Line12). Once the match is concluded, the final costs, profits, and paths of $\mathbf{F}'$ are updated (Line 13).

IV. Simulation Results

In this section, we evaluate our proposed algorithm and compare it with other baselines.

A. Setup and Baselines

We evaluate the proposed profit-aware resource provisioning algorithm in 14 nodes NSF Network, as Fig. 3 shows. Meanwhile, we present the values of the parameters in the model, shown in Table I. In the simulation, 200 services are created by intercepting and scaling real data, with the sliced data traffic restricted to the range of 0 to 10 Gbps. The delay for each slice is stipulated to be randomly selected within the range of 40 ms to 100 ms, at 10 ms intervals. The comparison algorithms employed in the simulation include the **FULL** algorithm, which performed resource provisioning without utilizing the proactive network slicing mechanism and adheres to a fixed bandwidth allocation, and the **FIX** algorithm, which is based on the proactive slicing mechanism for over-provisioning. We use AI-based (GRU-base) prediction from our previous studies [6] to make predictions and decide the credibility.

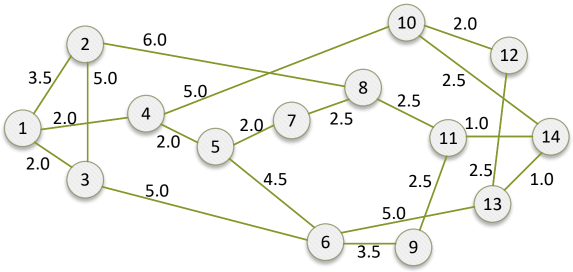

Fig. 3. The 14-node NSF network.

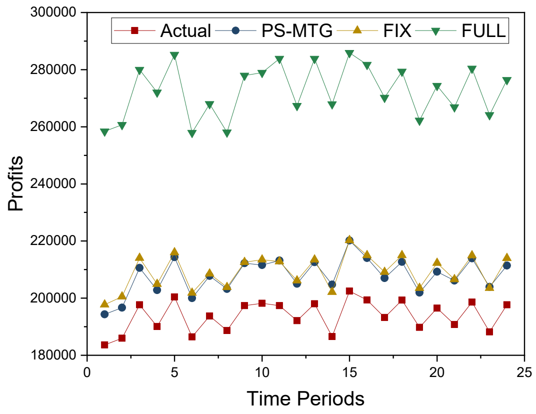

Fig. 4. The benefit of users.

B. Evaluation Results

TABLE I
THE VALUE OF PARAMETERS IN SIMULATION

Param	Value	Param	Value	Param	Value
b_f	5 Gbps	b_λ	10 Gbps	$\mathbb{N}^f$	50
η_λ	5	η_f	5	η_i	5
p^i_f	20	p^i_λ	30	p^i_m	50
δ_b	100 Mbps	T^i_a	10%	N^i_c	30
$\mathbb{N}^\lambda$	20	η_d	10	p^i_s	5

We evaluate four metrics: actual user revenue; expected revenue with PS-MTG; expected revenue with proactive allocation using the FIX algorithm but without employing the PS-MTG algorithm to scale sliced resources; and expected revenue with the FULL algorithm assuming allocation based

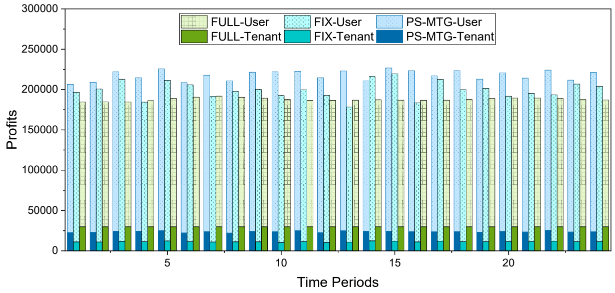

Fig. 5. Total benefits of different algorithms.

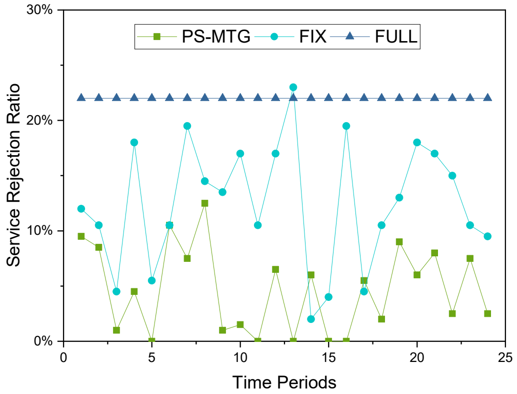

Fig. 6. Rejection ratio of different algorithms.

on the total slice bandwidth. In simulation, FIX and FULL algorithm are use the same parameters with PS-MTG algorithm.

The simulation spans 24 consecutive time intervals, illustrating user benefits in Fig. 4. Algorithms incorporating an proactive network slicing mechanism demonstrate more precise benefit assessments. The overall accuracy using the PS-MTG algorithm is 7.26%, and for the FIX algorithm, it is 7.95%, with a slight advantage in overall user benefit accuracy for the PS-MTG algorithm.

The total gains of the considered algorithms are shown in Fig. 5. We can observe that PS-MTG consistently achieves substantial gains overall. In comparison to the FIX algorithm, the PS-MTG algorithm lags slightly by 2.48% at time period 14, but outperforms the FIX algorithm in all other time periods, resulting in an overall improvement of 9.13%. Conversely, the FULL algorithm exhibits the poorest performance.

Additionally, we also compare the refusal rate of sliced services of the considered schemes, as shown in Fig. 6. In order to demonstrate the effectiveness of the algorithm, we overload services, which leads to a higher rejection ratio. Notably, PS-MTG has the lowest service rejection ratio, while the FULL algorithm exhibits a consistently high rejection ratio. The FIX algorithm, on the other hand, displays a fluctuating service rejection ratio, reaching up to 23%.

V. Conclusion

This paper focuses on slice resource provisioning in a multi-tenant FlexE-over-WDM network that maximizes profits while taking prediction uncertainty into account. Our approach involves modeling the multi-tenancy FlexE-over-WDM network, taking into account the profit generated by different roles. Specifically, we address the uncertainty in traffic prediction, analyze its impact, and formulate the problem of maximizing slice profits. To address this challenge, we introduce a profit-aware resource provisioning algorithm, comprising a precheck algorithm and a matching gaming algorithm. Simulation results demonstrate that the proposed algorithm outperforms benchmark methods in terms of user profit, total benefit, and service refusal ratio.

References

[1] C.-X. Wang, X. You, X. Gao, X. Zhu, Z. Li, C. Zhang, H. Wang, Y. Huang, Y. Chen, H. Haas, J. S. Thompson, E. G. Larsson, M. D. Renzo, W. Tong, P. Zhu, X. Shen, H. V. Poor, and L. Hanzo, "On the road to 6g: Visions, requirements, key technologies, and testbeds," *IEEE Commun. Surv. Tutorials*, vol. 25, no. 2, pp. 905–974, Feb. 2023.

[2] W. Jiang, B. Han, M. A. Habibi, and H. D. Schotten, "The road towards 6g: A comprehensive survey," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 334–366, Feb. 2021.

[3] Z. Ming, X. Li, C. Sun, Q. Fan, X. Wang, and V. C. Leung, "Sleeping cell detection for resiliency enhancements in 5g/b5g mobile edge-cloud computing networks," *ACM Trans. Sens. Netw.*, vol. 18, no. 3, pp. 1–30, Apr. 2022.

[4] M. Saqib, H. Elbiaze, and R. Glitho, "A profit-aware adaptive approach for in-network traffic classification," in *Proc. IEEE International Conference on Communications (ICC)*, May 2023, pp. 3351–3356.

[5] Q.-T. Luu, S. Kerboeuf, and M. Kieffer, "Uncertainty-aware resource provisioning for network slicing," *IEEE Trans. Netw. Serv. Manag.*, vol. 18, no. 1, pp. 79–93, Mar. 2021.

[6] Q. Guo, R. Gu, Z. Wang, T. Zhao, Y. Ji, J. Kong, R. Gour, and J. P. Jue, "Proactive dynamic network slicing with deep learning based short-term traffic prediction for 5g transport network," in *Proc. Optical Fiber Communication Conference (OFC)*, Mar. 2019, p. W3J.3.

[7] 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.

[8] Q. Guo, R. Gu, H. Yu, T. Taleb, and Y. Ji, "Probabilistic-assured resource provisioning with customizable hybrid isolation for vertical industrial slicing," *IEEE Transactions on Network and Service Management*, vol. 20, no. 2, pp. 1660–1675, 2023.

[9] T. Panayiotou, M. Michalopoulou, and G. Ellinas, "Survey on machine learning for traffic-driven service provisioning in optical networks," *IEEE Commun. Surv. Tutorials*, vol. 25, no. 2, pp. 1412–1443, Feb. 2023.

[10] S. Kashyap and A. Singh, "Prediction-based scheduling techniques for cloud data center's workload: a systematic review," *Cluster Comput.*, pp. 1–27, May 2023.

[11] M. Datar, E. Altman, and H. L. Cadre, "Strategic resource pricing and allocation in a 5g network slicing stackelberg game," *IEEE Transactions on Network and Service Management*, vol. 20, no. 1, pp. 502–520, 2023.

[12] J. Zheng, A. Banchs, and G. de Veciana, "Constrained network slicing games: Achieving service guarantees and network efficiency," *IEEE/ACM Transactions on Networking*, pp. 1–16, 2023.

[13] R. Ou, G. Sun, D. Ayepah-Mensah, G. O. Boateng, and G. Liu, "Two-tier resource allocation for multitenant network slicing: A federated deep reinforcement learning approach," *IEEE Internet of Things Journal*, vol. 10, no. 22, pp. 20 174–20 187, 2023.

[14] S. Huang, B. Guo, and Y. Liu, "5g-oriented optical underlay network slicing technology and challenges," *IEEE Commun. Mag.*, vol. 58, no. 2, pp. 13–19, Feb. 2020.

[15] R. Vilalta, R. Martínez, R. Casellas, R. Muñoz, Y. Lee, L. Fei, P. Tang, and V. López, "Network slicing using dynamic flex ethernet over transport networks," in *Proc. European Conference on Optical Communication (ECOC)*, Sept. 2017, pp. 1–3.