photonics

MDPI

Article

Transformer-Assisted Service-Attribute-Aware Lightpath Provisioning in Multi-Granularity Optical Networks

Xinglong Zou, Zhilin Yuan, Vimal Bhatia and Yongcheng Li *

School of Electronic and Information Engineering, Soochow University, Suzhou 215006, China; 20255228037@stu.suda.edu.cn (X.Z.); 20224228048@stu.suda.edu.cn (Z.Y.); vbhatia@iiti.ac.in (V.B.)
* Correspondence: ycli@suda.edu.cn

Abstract

Multi-granularity optical networks (MGONs) provide flexible switching capabilities at the waveband, wavelength, and sub-wavelength levels, which are promising for accommodating heterogeneous service requests with different bandwidth, delay, and quality requirements. However, existing lightpath provisioning strategies usually rely on fixed routing or cost-oriented rules, making it difficult to match dynamic network resources with service-specific requirements. To address this problem, this paper proposes a service-aware framework for joint routing, modulation format, granularity, and wavelength assignment (RMGWA) in MGONs, which integrates a Transformer-assisted service-attribute-aware (TSAA) strategy with a waveplane-based RMGWA algorithm. In the proposed strategy, the Transformer module takes service attributes and present network states as inputs and generates service-specific link weights to guide candidate route selection. Then, the proposed waveplane-based RMGWA algorithm is applied to determine feasible resource assignments. Also, a mixed-integer linear programming (MILP) model is formulated as a benchmark for small-scale scenarios. Simulation results on n6s9 and USNET show that the proposed framework improves service satisfaction and spectrum utilization while maintaining low establishment cost. Moreover, it is observed that the proposed TSAA does not increase average completion time of services; instead, it supports faster service-aware routing decisions. These results demonstrate the effectiveness of integrating TSAA with the waveplane-based RMGWA algorithm for adaptive lightpath provisioning in dynamic MGON environments.

Keywords: multi-granularity optical networks; service-aware provisioning; RMGWA; Transformer

Check for updates

Received: 9 October 2026
Revised: 21 October 2026
Accepted: 21 October 2026
Published: 26 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

With the emergence of heterogeneous network services, including artificial intelligence-generated content (AIGC) model training, cloud computing, fifth-generation (5G) applications, network traffic has become increasingly diverse in terms of bandwidth, latency, reliability, and transmission-quality requirements [1]. For example, distributed AI training involves massive data transfers and therefore requires high-capacity connections, whereas interactive cloud services are particularly sensitive to latency, and mission-critical ultra-reliable low-latency communications (URLLC) imposes stringent reliability requirements. The conventional fixed-grid optical networks, however, provide limited flexibility in matching network resources to such diverse service requirements. Their fixed resource granularity and relatively rigid provisioning mechanisms may result in inefficient resource

utilization and make it difficult to simultaneously satisfy the differentiated requirements of heterogeneous services.

To address this issue, elastic optical networks (EONs) have been extensively studied to improve flexibility in spectrum allocation by dividing the optical spectrum into fine-grained frequency slots (FSs). Nevertheless, several limitations hinder their cost-effective deployment for heterogeneous service provisioning. Firstly, EON deployment typically requires flex-grid wavelength-selective switches (WSSs) and bandwidth-variable transceivers, increasing both upgrade costs and hardware complexity. Secondly, the dynamic allocation and release of variable-sized spectrum blocks may cause severe spectrum fragmentation, particularly under dynamic traffic conditions. Consequently, fine-grained spectrum allocation alone may not always provide a cost-effective solution that matches the diverse granularities of service demands.

In contrast, multi-granularity optical networks (MGONs) provide a promising solution by supporting sub-wavelength-, wavelength-, and waveband-level switching within a unified network. MGONs can be used to select an appropriate switching granularity according to the attributes of each service while remaining compatible with existing fixed-grid optical infrastructure. By avoiding exclusive reliance on fine-grained spectrum allocation, MGONs achieve a better balance among provisioning flexibility, hardware cost, switching overhead, and resource utilization.

However, the advantages of MGONs cannot be fully accomplished unless the switching granularity is properly matched to the attributes of each service request. For example, if a bandwidth-intensive service is provisioned with an overly fine switching granularity, multiple wavelength resources may be occupied separately, leading to resource fragmentation and reduced spectrum efficiency. Conversely, if a delay-sensitive service is aggregated into a coarse-granularity lightpath, additional waiting delay may be introduced.

Existing studies on MGONs have mainly focused on various aspects of network design and operation, including multi-granularity node architectures and lightpath provisioning strategies [2–5]. Nevertheless, efficient provisioning of heterogeneous services in MGONs remains challenging because the network resource cannot be allocated appropriately to match the need of different services. Therefore, a joint routing, switching-granularity, modulation-format, and wavelength-assignment algorithm and network design are required, which can recognize both service attributes and network resource states when provisioning each incoming service request.

To address this problem, we investigate service-attribute-aware lightpath provisioning in MGONs. The considered problem is divided into two closely related parts: service-attribute modeling and service-adaptive lightpath provisioning. In service-attribute modeling, each incoming request is characterized according to its bandwidth demand, delay sensitivity and transmission-quality requirements so as to determine its service type. Subsequently, in the lightpath provisioning stage, these service attributes are jointly considered with the network resource state to determine the route, switching granularity, modulation format, and wavelength assignment. Based on the two parts above, we formulate the service-aware RMGWA problem, which is studied to achieve efficient provisioning for heterogeneous services in MGONs.

Compared with conventional RWA schemes and cost-oriented MGON provisioning approaches, the proposed formulation jointly considers service type and network resource. Firstly, we develop a mixed-integer linear programming (MILP) model to obtain optimal solutions for small-scale network instances. Secondly, we propose an extended waveplane-based RMGWA algorithm, which is deployed in different strategies for dynamic lightpath provisioning in MGONs. Thirdly, we design a Transformer-assisted service-attribute-aware strategy, termed TSAA, to adapt lightpath provisioning decisions to service attributes and real-time network states.

https://doi.org/10.3390/photonics13111010

The remainder of this paper is organized as follows. Section 2 reviews related work on MGON architectures, lightpath provisioning, and service-attribute modeling. Section 3 introduces the MGON architecture and formulates the RMGWA problem. Section 4 presents the MILP model as an optimization benchmark. Section 5 describes the service modeling method. Section 6 presents the proposed TSAA strategy and waveplane-based RMGWA algorithm, with CP and QSP strategies for comparison. Section 7 presents the simulation results. Finally, Section 8 concludes the paper.

2. Related Work

2.1. Multi-Granularity Optical Switching Node Architecture

As the foundation of MGONs, the design of multi-granularity optical cross-connect (MG-OXC) node architectures directly determines how optical signals can be switched and groomed at different granularities. Accordingly, representative studies have explored diverse MG-OXC architectures. Wang et al. [6] provided a comprehensive review of advancements in MG-OXC architecture design, including two representative MG-OXC designs: the three-layer MG-OXC, consisting of FXC, BXC, and WXC layers, and the single-layer MG-OXC, which integrates these logical functions into a more compact structure. Their analysis clarified the trade-off between switching flexibility and implementation simplicity in MGON. Izmailov et al. [7] investigated hybrid optical devices that combine waveband (transparent optical) and wavelength (opaque OEO) switching, offering insights into multi-granularity switching architectures from a node dimension perspective. Zhang et al. [8] proposed a novel, cost-effective multi-granularity optical switching node with wavelength management and resource polling capabilities, which significantly reduces the number of optical switch ports and wavelength selective switches (WSSs). Huang et al. [2] introduced a flexible-grid multi-granularity optical node architecture supporting both fixed-grid and flex-grid technologies, enabling adaptive switching based on service signal requirements to meet diverse transmission demands. Zhang et al. [3] designed a simplified and efficient two-layer structure for a new multi-granularity space-division multiplexing (SDM) ROADM, achieving functions such as all-optical multi-granularity switching, traffic grooming, and service aggregation.

2.2. Lightpath Provisioning in MGON

In MGONs, the coexistence of waveband, wavelength, and sub-wavelength granularities makes lightpath provisioning more complex than the conventional RWA, since routing, traffic grooming, and granularity assignment needs to be jointly considered. Accordingly, a range of lightpath provisioning strategies has been developed for MGONs in the literature. To address this issue, Hou et al. [4] proposed an integrated grooming auxiliary graph-based routing algorithm, which supports waveband-level switching for wavelength demands and sub-wavelength traffic grooming, thereby reducing switching-port usage and deployment cost. Zhang et al. [9] further considered multi-granularity grooming for multicast traffic, where energy consumption, service quality, and grooming from low to high granularities are jointly taken into account to reduce blocking probability and energy consumption. Wang et al. [10] studied routing and multi-granularity assignment under dynamic traffic conditions, combining minimum-load routing with a first-fit strategy to select available fiber, waveband, or wavelength resources in a three-layer MG-OXC architecture.

Additionally, Dharmaweera et al. [11] combined traffic grooming with multipath routing in WDM networks and proposed analytical models for small-scale networks and heuristic algorithms for large-scale networks, significantly improving spectral efficiency and reducing resource overhead. Zhang et al. [12] introduced a three-layer auxiliary graph model tailored for dynamic traffic scenarios in MGONs. By adjusting the edge weights, this model enables flexible grooming strategies that improve spectral performance.

2.3. Service-Attribute Modeling in Optical Networks

Recent studies have increasingly incorporated application attributes, traffic dynamics, and service-level objectives into grooming and resource allocation decisions. For example, Tomkos et al. [13] proposed an SDN-based multilayer control framework for elastic WDM optical networks, where service requests are described by latency, jitter, bandwidth, reliability and security requirements. F. Metzger et al. [14] modeled IoT services according to periodic or event-driven traffic patterns, message intervals, traffic volume, maximum end-to-end delay, and loss rate, providing a basis for large-scale IoT traffic dimensioning. A. S. Santos et al. [15] studied application-aware service degradation in EONs by characterizing each request with service class, bandwidth demand, allowable degradation, and maximum delay, and mapping application-layer priorities to optical-layer lightpath classes.

In addition, H. Yang et al. [16] classified data-center applications into bandwidth-intensive and data-intensive services, and showed that optical topology configuration and traffic routing should be adapted to different traffic patterns. Z. Zhou et al. [17] represented optical fronthaul traffic as spatial–temporal traffic matrices and proposed a prediction model to support resource scheduling under diverse bandwidth, latency, and reliability requirements. G. Drainakis et al. [18] modeled data-center applications using RTT constraints, traffic priorities, and traffic-share scenarios, and designed application-aware WDM MAC protocols to prioritize latency-critical traffic.

The reviewed studies provide useful foundations for multi-granularity optical switching architectures, lightpath provisioning, and service-aware resource optimization. Architecture-oriented works mainly focus on the design and performance analysis of hierarchical or multi-granularity optical nodes, while provisioning studies usually address routing, wavelength assignment, grooming, or granularity selection under specific network constraints. In addition, recent works have introduced service-attribute modeling, traffic prediction, and intent-based control to improve the adaptability of optical network provisioning.

However, these research directions are still mostly investigated separately. Existing studies rarely integrate the joint interaction among heterogeneous service attributes, real-time multi-granularity resource states, routing decisions, switching granularity selection, modulation format selection, wavelength assignment, and service aggregation in a unified MGON provisioning framework.

3. Multi-Granularity Optical Network

3.1. Architecture of Multi-Granularity OXC Node

A node that integrates multiple switching technologies is referred to as a multi-granularity optical switching node. Supported by an MG-OXC, such a node enables service switching across multiple granularities and facilitates add/drop operations for lightpath services using different transmission granularities. By incorporating components such as demultiplexers and modulators within the MG-OXC, the node can achieve seamless conversions between waveband, wavelength, and sub-wavelength levels. According to the switching hierarchy, MG-OXC architectures can be classified into single-layer and multilayer structures.

This paper adopts a multi-granularity OXC architecture, as illustrated in Figure 1 [19], where compared with the single-layer design, the primary distinction lies in the interconnection of switching modules through dedicated interfaces. Specifically, the fiber cross-connect (FXC) and waveband cross-connect (BXC) are linked via waveband multiplexers/demultiplexers; the BXC and wavelength cross-connect (WXC) are connected through wavelength multiplexers/demultiplexers; and the WXC and sub-wavelength cross-connect (SXC) interface through modulators and demodulators.

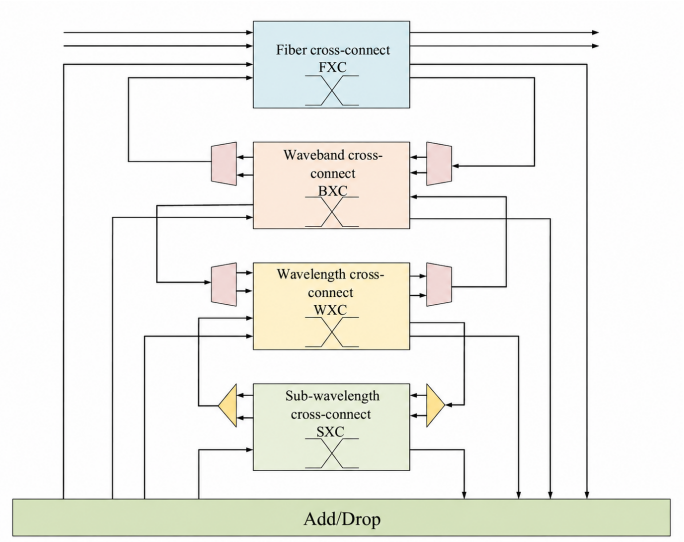


Figure 1. Multi-granularity OXC architecture.

A key characteristic of the multilayer MG-OXC is its hierarchical switching mechanism, which requires lower-granularity switching operations to traverse the higher-granularity switching modules. For instance, when two sub-wavelength channels in a fiber need to be switched at the sub-wavelength level, the process unfolds as follows: First, the signals are routed in the FXC to the fiber-to-band port and demultiplexed into individual wavebands. These wavebands are then forwarded to the BXC, where they are switched to the band-to-wavelength port and further demultiplexed into individual wavelengths. The selected wavelength is subsequently routed in the WXC to the wavelength-to-sub-wavelength port and modulated via an OFDM modulator into two sub-wavelength channels. Finally, the SXC performs the sub-wavelength switching on these channels.

3.2. Service-Aware RMGWA for Mixed Services in MGON

Unlike the traditional RWA problem in WDM optical networks, MGONs support multiple switching granularities, offering greater flexibility for lightpath establishment. Therefore, a critical challenge lies in how to select appropriate switching granularities, modulation formats and efficiently perform routing and wavelength assignment.

Figure 2 illustrates two representative RMGWA provisioning decisions for two services, r_1 and r_2 in a six-node MGON, requiring low latency and high bandwidth respectively. The network consists of MG-OXC nodes connected by fiber links, and each node supports waveband-, wavelength-, and sub-wavelength-level switching. The establishment cost is evaluated according to the number of transceiver pairs consumed by the occupied wavelengths. Specifically, each occupied wavelength consumes one transceiver pair on each hop between adjacent MG-OXC nodes.

Under the policy S1, shown in Figure 2a, both r_1 and r_2 are transmitted through BPSK-modulated sub-wavelength channels, with each service occupying one half of wavelength λ_1. This allocation prioritizes the minimization of lightpath establishment cost and consumes two transceiver pairs in total. Under policy S2, shown in Figure 2b, r_1 is still transmitted through a BPSK-modulated sub-wavelength channel $\lambda_1/2$, while the bandwidth-intensive service r_2 is assigned a 16-QAM-modulated waveband composed of $\lambda_1 + \lambda_2$. This configuration provides a more appropriate resource granularity for the heterogeneous service requirements but consumes four transceiver pairs in total.

This example demonstrates that the RMGWA procedure jointly determines the route, modulation format, switching granularity, and wavelength assignment. Different provisioning priorities may therefore lead to different resource configurations for the same network

topology. It also illustrates the trade-off between transceiver cost and service requirement satisfaction in MGONs.

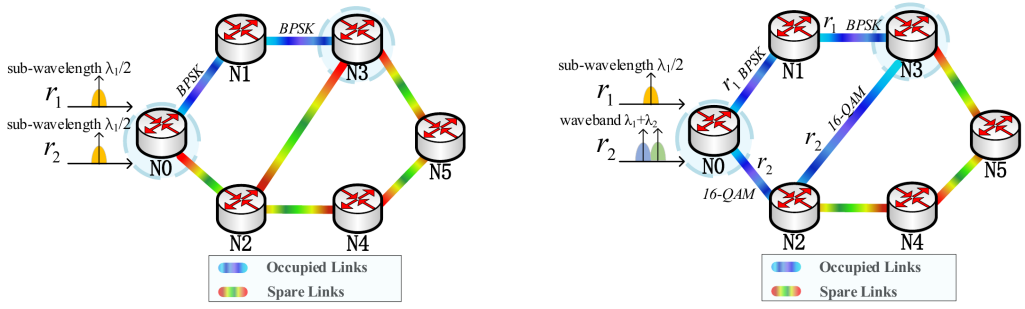


Figure 2. Examples of RMGWA provisioning in a six-node MGON: (**a**) policy S1; (**b**) policy S2.

3.3. Problem Statement

This paper considers a lightpath provisioning problem in MGONs, where service requests arrive sequentially and are provisioned according to their service attributes and the current network resource state. The network is modeled as a physical topology composed of MG-OXC nodes and fiber links. Each MG-OXC node supports waveband-, wavelength-, and sub-wavelength-level switching, and each fiber link provides a limited number of wavelengths. For each incoming service request, the source node, destination node, bandwidth demand, delay sensitivity, and transmission-quality requirement are known before provisioning.

Based on the above assumptions, we formulate an RMGWA problem in MGONs, where routing, modulation format, switching granularity, and wavelength assignment are jointly determined for each service request. The provisioning decision should satisfy resource capacity constraints, wavelength assignment constraints, and physical layer transmission constraints, while matching the allocated network resources to heterogeneous service attributes.

In order to provide a rigorous optimization benchmark, we further develop a MILP model with two joint objectives: minimizing the lightpath establishment cost and maximizing network resource utilization. However, as the scale of the network increases, the number of decision variables and constraints grows rapidly, making the MILP difficult to solve. Therefore, heuristic provisioning algorithms are further proposed to improve computational efficiency while maintaining service-aware resource allocation performance.

4. MILP Model for the Service-Aware RMGWA Problem

The service-attribute-aware lightpath provisioning problem is formulated as a MILP model. The model jointly determines routes, switching granularities, modulation formats, wavelength assignments, and possible service aggregation. The sets, parameters, and variables are summarized in Tables 1–3.

Table 1. Sets used in the MILP model.

Symbol	Description
N	Set of nodes in the MG-OXC network.
L	Set of physical links in the MG-OXC network.
E	Set of selectable switching granularities, including sub-wavelength, wavelength, and waveband.
M	Set of candidate modulation formats, including BPSK, QPSK, 8-QAM, and 16-QAM.
P	Set of pending service requests.
ρ	Set of service types, where $\rho \in \{0, 1, 2\}$ represents bandwidth-intensive, latency-sensitive, and QoS-sensitive services, respectively.
R	Set of candidate routes for services considered by the MILP model.

Table 2. Parameters used in the MILP model.

Symbol	Description
p	Service p to be transmitted in P.
s, d	Source and destination nodes of service p, respectively.
W	Maximum number of available wavelengths per channel.
BW_i	Spectrum bandwidth occupied by wavelength resource i, in GHz.
κ_p	Total number of transceivers used by all services $p \in P$ in the network.
$\delta^{p,r}$	Total number of hops for service p under route r.
C_p	Total cost of establishing an optical channel for service p, where $C_p = \kappa_p$.
$C_{p,k}$	Total cost of establishing an optical channel when services p and k are aggregated, where $C_{p,k} = \kappa_{p,k}$.
α	Tunable weighting coefficient for the trade-off between optical-channel establishment cost and network bandwidth-resource utilization.
∇	A tiny value.
BW_{total}	Total amount of available spectrum in the network, in GHz.
$N^{\text{pr}}_{\text{total}}$	Sum of the number of available transceivers across all links in the network.

Table 3. Variables used in the MILP model.

Symbol	Description
r	Route allocated to service p from source node s to destination node d under the current strategy.
λ_i^p	Wavelength λ_i allocated to service p under the current strategy.
l	Link on the route of a service.
$\sigma^k_{\lambda_i}$	Binary variable: $\sigma^k_{\lambda_i} = 1$ if k services are aggregated and occupy wavelength λ_i, and $\sigma^k_{\lambda_i} = 0$ otherwise.
$\gamma^t_{e_j,\lambda_j}$	Real-valued variable indicating that t services are transmitted using switching granularity e_j and wavelength λ_j; its value depends on the occupied wavelength λ_j.
$\theta^p_{m,e}$	Number of service p that selects switching granularity e and modulation format m.
$\theta^{i,p,k}_{m,e}$	Number of switching-granularity e and modulation-format m selections when services p and k are aggregated for transmission.
μ^l_p	Binary variable: $\mu^l_p = 1$ if link l is on the route L, and $\mu^l_p = 0$ otherwise.
$\alpha^p_{m,e}$	Minimum OSNR value satisfying the transmission requirements of service p when modulation format m and switching granularity e are selected.
$\pi^i_{p,k}$	Binary variable: $\pi^i_{p,k} = 1$ if services p and k are selected for aggregated transmission, and $\pi^i_{p,k} = 0$ otherwise.
$x^{p,r}$	Binary variable: $x^{p,r} = 1$ if service p adopts route r, and $x^{p,r} = 0$ otherwise.
ϵ	Real-valued variable representing the wavelength-occupancy threshold at which the network is about to become congested.
ζ	Real-valued variable representing the minimum bandwidth-resource utilization of the network under ideal conditions.
$\text{OSNR}^l_{\text{rec}}$	Reciprocal of the OSNR for each link l on route $r^p_{s,d}$ allocated to service p.
$\text{OSNR}^m_{\text{rec}}$	Reciprocal of the minimum OSNR required to establish an optical channel for service p when modulation format m is selected.
$\text{OSNR}^{p,l}_{\text{rec}}$	Reciprocal of the sum of OSNR values on route link l allocated to service p.
$\text{OSNR}^{p,k,l}_{\text{rec}}$	Reciprocal of the sum of OSNR values on route link l allocated for aggregated transmission of services p and k.

The MILP objective in Equation (1) jointly considers lightpath establishment cost and network resource utilization:

$$\min \ \alpha\kappa - (1-\alpha)\frac{\sum_{i=1}^{n} BW_i}{BW_{\text{total}}}, \tag{1}$$

where α is a tunable trade-off coefficient, κ is the total number of transceiver pairs, and BW_i is the occupied bandwidth of wavelength i. The total number of transceiver pairs is expressed as

$$\kappa = \sum_{p \in P} \sum_{r \in R} \delta^{p,r} x^{p,r} + \sum_{\lambda_i \in W} \sum_{k \geq 2} (2-k) \sigma_{\lambda_i}^{k}, \tag{2}$$

Here, $x^{p,r}$ is a binary variable that equals 1 when service p selects route r and $\sigma_{\lambda_i}^{k}$ is a binary variable that equals 1 when k services are aggregated and occupy wavelength λ_i. Moreover, $\delta^{p,r}$ is the hop count of route r for service p. Equation (2) calculates the total number of transceiver pairs by accounting for route hop counts and the resource savings obtained when multiple services share a wavelength.

For any service, the wavelength-occupancy state is defined by Equation (3):

$$\gamma_{e_j,\lambda_i}^{t} = \begin{cases} \sum_i i, & i \geq 2, \\ 1, & i = 1. \end{cases} \tag{3}$$

For a single service, Equation (4) first requires a unique modulation-format and switching-granularity configuration. Based on this configuration, Equation (5) obtains the accumulated reciprocal OSNR across the selected links. Equation (6) then compares this value with the corresponding modulation-format threshold, while Equation (7) ensures that the selected configuration satisfies the service transmission-quality requirement:

$$\sum_{m \in M} \sum_{e \in E} \theta_{m,e}^{p} = 1. \tag{4}$$

$$\mathrm{OSNR}_{\mathrm{rec}}^{p,l} = \sum_{l} \mu_{p}^{l} \cdot \mathrm{OSNR}_{\mathrm{rec}}^{l}. \tag{5}$$

$$\mathrm{OSNR}_{\mathrm{rec}}^{p,l} - \mathrm{OSNR}_{\mathrm{rec}}^{m} \leq \nabla. \tag{6}$$

$$\mathrm{OSNR}_{\mathrm{rec}}^{p,l} \geq \alpha_{m,e}^{p}. \tag{7}$$

For aggregated services, Equation (8) selects a common modulation format and switching granularity. Equation (9) calculates the accumulated reciprocal OSNR on the shared lightpath, which is constrained by the modulation-format threshold in Equation (10). Equations (11) and (12) then ensure the transmission quality of services p and k, respectively:

$$\sum_{m \in M} \sum_{e \in E} \theta_{m,e}^{i,p,k} = 1. \tag{8}$$

$$\mathrm{OSNR}_{\mathrm{rec}}^{p,k,l} = \sum_{l \in L} \mu_{p}^{l} \cdot \mathrm{OSNR}_{\mathrm{rec}}^{l} + \sum_{l \in L} \mu_{k}^{l} \cdot \mathrm{OSNR}_{\mathrm{rec}}^{l}. \tag{9}$$

$$\mathrm{OSNR}_{\mathrm{rec}}^{p,k,L} - \mathrm{OSNR}_{\mathrm{rec}}^{m} \leq \nabla. \tag{10}$$

$$\mathrm{OSNR}_{\mathrm{rec}}^{p,L} \geq \alpha_{m,e}^{p}. \tag{11}$$

$$\mathrm{OSNR}_{\mathrm{rec}}^{k,L} \geq \alpha_{m,e}^{k}. \tag{12}$$

Minimal network costs used. Equations (13)–(15) are used to guarantee that each service is aggregated at most once and that the numbers of occupied transceiver pairs and wavelengths remain within the available limits:

$$\sum_{i} \pi_{p,k_1}^{i} + \sum_{i} \pi_{p,k_2}^{i} \leq 1. \tag{13}$$

$$\sum_{p\in P}\sum_{r\in R}\delta^{p,r}x^{p,r}+\sum_{\lambda_i\in W}\sum_{k\geq 2}(2-k)\sigma_{\lambda_i}^{k}\leq N_{\text{total}}^{\text{pr}}. \tag{14}$$

$$\left[\gamma_{e_j,\lambda_i}^{t}\right]\leq W. \tag{15}$$

Maximum link bandwidth used. Equation (16) prevents wavelength-resource conflicts between different services on the same physical link:

$$\sum_{i}\lambda_i^{p}\sum_{j}\lambda_j^{k}\leq 1. \tag{16}$$

To prevent excessive concentration of wavelength resources, Equation (17) limits the wavelength occupancy ratio on each physical link to the prescribed congestion threshold:

$$\frac{\gamma_{e_j,\lambda_i}^{t}}{W}\leq \varepsilon. \tag{17}$$

Meanwhile, Equation (18) requires the bandwidth-resource utilization to meet the desired target:

$$\frac{\sum_{i=1}^{n}BW_i}{BW_{\text{total}}}\geq \zeta. \tag{18}$$

Finally, Equation (19) ensures that the total occupied spectrum does not exceed the spectrum available across the network.

$$\sum_{i=1}^{n}BW_i\leq BW_{\text{total}}. \tag{19}$$

Because several of the above resource and transmission constraints contain products of binary and bounded integer variables, they are nonlinear in their original form. For the product $y=x_1x_2$, where $x_1\in\{0,1\}$ and $0\leq x_2\leq u$, the standard linear reformulation is given by Equations (20)–(23) [20]:

$$y\leq u\cdot x_1. \tag{20}$$

$$y\leq x_2. \tag{21}$$

$$y\geq x_2-u\cdot(1-x_1). \tag{22}$$

$$y\geq 0. \tag{23}$$

Together, Equations (20)–(23) ensure that y is equal to x_1x_2 while preserving the linear form of the MILP model.

The dominant numbers of variables and constraints both scale as $\mathcal{O}(|\mathbf{N}|\cdot|\mathbf{M}|\cdot|\mathbf{P}|\cdot|\mathbf{E}|)$.

5. Service-Attribute Modeling for MGON

To support service-aware RMGWA in MGONs, each incoming service request is first described according to its observable transmission requirements. Inspired by application-aware service modeling studies [13–18], this paper considers service-side attributes including source node, destination node, bandwidth demand, holding time, delay requirement, and transmission-quality requirement. These attributes do not directly provide the service type. Instead, they are used as the input information for service-attribute modeling, through which the dominant provisioning requirement of each service is identified.

The proposed service-attribute modeling architecture is shown in Figure 3. When a service request arrives, the control plane first collects its service attributes and constructs a structured service-attribute vector. The attributes are normalized and evaluated through three score functions. The dominant service requirement is then identified according to the

maximum score, yielding the service type ρ_p, where $\rho_p = 0$, $\rho_p = 1$, and $\rho_p = 2$ represent bandwidth-intensive, delay-sensitive, and QoS-sensitive services, respectively.

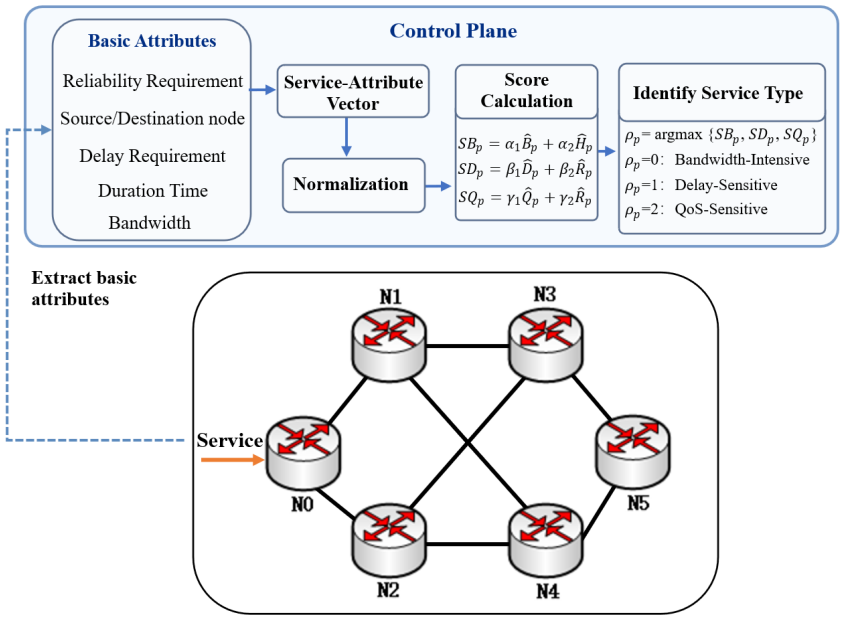


Figure 3. Service-attribute modeling architecture.

The service attribute modeling procedure is summarized as follows. First, the control plane receives an incoming service request and extracts its basic attributes, including reliability requirement, bandwidth demand, duration time, delay requirement, and transmission-quality requirement. Second, these attributes are normalized to eliminate differences in their numerical scales. Third, the normalized attributes are evaluated using three specific score functions. Finally, the service type ρ_p is determined by selecting the category with the highest score. The identified type, together with the original service attributes, is then provided to the subsequent TSAA and RMGWA procedures to adaptively determine the route, modulation format, switching granularity, and wavelength assignment.

For each incoming service request p, its observable service-attribute vector is defined as

$$\mathbf{S}_p = [B_p, H_p, D_p, Q_p, R_p], \tag{24}$$

where B_p is the bandwidth demand, H_p is the duration time, D_p is the maximum tolerable delay, Q_p is the transmission-quality requirement, and R_p is the reliability requirement. Since these attributes have different units and scales, they are normalized before classification. Specifically, the normalized bandwidth demand, holding time, delay sensitivity, transmission-quality requirement, and reliability requirement are defined as

$$\hat{B}_p = B_p / B_{\max}. \tag{25}$$

$$\hat{H}_p = H_p / H_{\max}. \tag{26}$$

$$\hat{D}_p = (D_{\max} - D_p) / (D_{\max} - D_{\min}). \tag{27}$$

$$\hat{Q}_p = Q_p / Q_{\max}. \tag{28}$$

$$\hat{R}_p = R_p / R_{\max}. \tag{29}$$

Based on the application-aware service modeling idea, three service-type scores are then calculated as

$$S_{B,p} = \alpha_1 \hat{B}_p + \alpha_2 \hat{H}_p. \tag{30}$$

$$S_{D,p} = \beta_1 \hat{D}_p + \beta_2 \hat{R}_p. \tag{31}$$

$$S_{Q,p} = \gamma_1 \hat{Q}_p + \gamma_2 \hat{R}_p. \tag{32}$$

where $S_{B,p}$, $S_{D,p}$, and $S_{Q,p}$ represent the bandwidth-intensive, delay-sensitive, and transmission-quality-sensitive scores, respectively. Since the attributes are normalized as above, the simulations assign equal weights to the two terms in each score function, setting $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = \gamma_1 = \gamma_2 = 0.5$ so that neither attribute is favored within its score. The same coefficient values are used for n6s9 and USNET. The service type of request p is finally identified as

$$\rho_p = \arg\max\{S_{B,p}, S_{D,p}, S_{Q,p}\}. \tag{33}$$

Accordingly, $\rho_p = 0$ indicates a bandwidth-intensive service, $\rho_p = 1$ indicates a delay-sensitive service, and $\rho_p = 2$ indicates a QoS-sensitive service. Based on the above scoring rules, each service request is assigned a service type ρ_p, which provides a quantified description of the dominant service requirement and is subsequently combined with the original service attributes and current network state for service-specific link-weight prediction.

6. Heuristic Algorithm

The MILP model proves effective in addressing optimization problems of smaller scales. However, its computational complexity increases significantly as the network scale expands. To tackle this challenge and enhance allocation efficiency, we introduce heuristic strategies and algorithms for handling optimization problems of larger magnitudes. Our devised strategies focus on adapting switching modes for each lightpath service, presenting three distinct strategies: Cost-Priority (CP), Quality-of-Service Priority (QSP) which are for comparison, and the proposed Transformer-assisted service-attribute-aware strategy (TSAA).

6.1. Switching Granularity Adaptation Strategies

Cost-Priority (CP) Strategy. This strategy gives priority to provisioning combinations with lower lightpath establishment costs. For each service request p, CP strategy examines feasible modulation formats $m \in M$, switching granularities $e \in E$, and eligible aggregation choices. The cost of an individual service is represented by C_p, while the cost of aggregating services p and k is represented by $C_{p,k}$. The algorithm ranks the candidate combinations in ascending order of C_p or $C_{p,k}$, and then selects the lowest cost feasible combination. The strategy therefore does not explicitly account for service type or differentiated QoS requirements.

Quality-of-Service Priority (QSP) Strategy. This strategy gives priority to feasible provisioning combinations according to the dominant requirement of service p [21]. It selects the route, modulation format, and switching granularity that provide the highest transmission quality for the identified service type, without considering lightpath establishment cost or service aggregation between node pairs. Consequently, QSP may occupy excessive wavelength resources and cause unbalanced link loads because network resource utilization is not explicitly optimized.

6.2. Transformer-Assisted Service-Attribute-Aware (TSAA) Strategy

The Cost-Priority strategy, from the perspective of cost minimization, effectively reduces the cost of optical channel establishment but struggles to satisfy the transmission requirements of all services. In contrast, the Quality-of-Service Priority strategy focuses

on service quality while neglecting both the lightpath establishment and the efficient coordination of network resource.

To address the limitations of both strategies, this chapter proposes a TSAA strategy. Different from CP and QSP, TSAA jointly considers the identified service type, service attributes, and network resource states. For each incoming request, the service-attribute modeling module first identifies the dominant service requirement and outputs service type. Then, TSAA uses it with the present state of the network to generate specific link weights so that candidate routes can be selected according to both service requirements and resource availability.

In the provisioning process, bandwidth-intensive services are guided towards routes with sufficient residual capacity and better aggregation opportunities; delay-sensitive services are guided towards routes with lower hop count and lower transmission delay; and QoS-sensitive services are guided towards routes and modulation formats that better satisfy physical-layer constraints. Meanwhile, TSAA also avoids heavily occupied links to improve resource-utilization efficiency. Therefore, TSAA does not directly replace the RMGWA procedure; instead, it provides adaptive routing guidance, while the subsequent waveplane-based RMGWA algorithm completes feasible modulation-format, switching-granularity, and wavelength assignment.

6.3. Transformer-Assisted Link-Weight Prediction

In the proposed TSAA strategy, the Transformer module is used as an adaptive link-weight predictor to support service-aware routing and granularity assignment. For each incoming service request, the Transformer takes the service type, service attributes and the current network state as inputs to generate adaptive link weights $\hat{w}$ that guide candidate routing decisions. The TSAA process is shown in Figure 4.

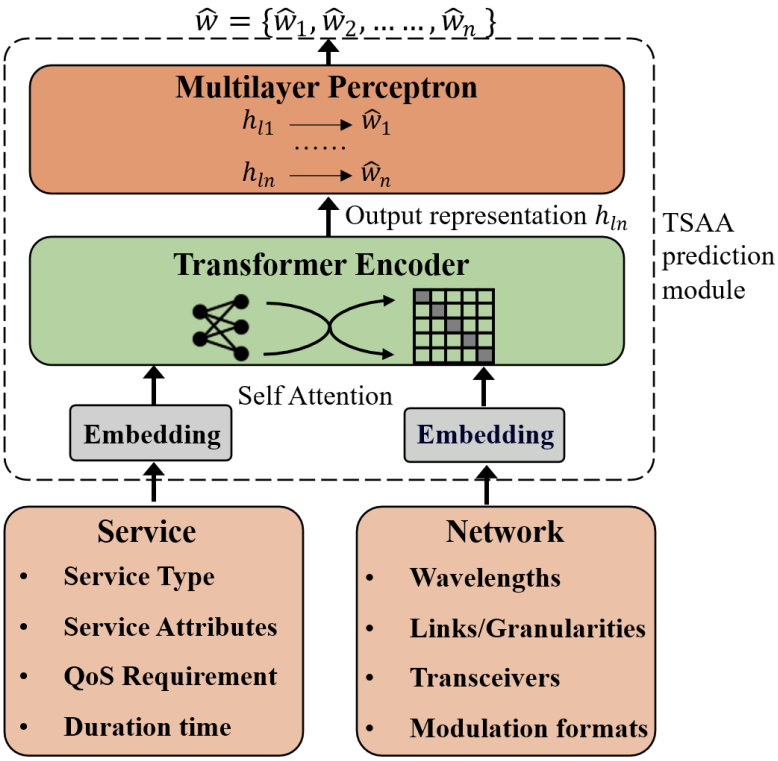


Figure 4. TSAA prediction procedure.

For service request, the service vector is constructed based on the identified service type and service attributes. For each physical link, the link-state vector is constructed from the current resource occupation variables and physical layer parameters.

The service vectors and the link-state vectors are first mapped into a unified hidden space through embedding layers and then fed into a Transformer encoder. Through the self-attention mechanism, the Transformer captures the dependency between the service attributes and the multi-granularity resource states of different links. The output representation corresponding to each link is passed through a multilayer perceptron to obtain an adaptive link weight $\hat{w}$.

Based on the predicted link weights, a service-specific weighted topology is constructed. The *K shortest path* algorithm is then applied to generate candidate routes from source node to destination node. For each candidate route, TSAA evaluates feasible combinations of modulation format, switching granularity, and wavelength. The selected provisioning decision should satisfy the OSNR constraints, resource capacity constraints, and wavelength constraints defined in the MILP model. To rank different feasible candidates, TSAA uses the following joint score function:

$$J = \beta_1 C_p + \beta_2 U_R + \beta_3 \delta^{p,r} + \beta_4 V_p. \tag{34}$$

where C_p is the lightpath establishment cost defined in the MILP model, U_R represents the resource occupation level of route r, $\delta^{p,r}$ denotes the number of hops of service p on route r, and V_p denotes the granularity-conversion penalty. If service aggregation is adopted, $C_{p,k}$ is used instead of C_p. The coefficients β_1–β_4 are adjusted according to service type ρ. For bandwidth-intensive services, TSAA gives more weight to resource utilization; for delay-sensitive services, it gives more weight to hop count and aggregation delay; for QoS-sensitive services, it assigns a higher penalty to unnecessary granularity conversions and unsuitable modulation selections. Among all feasible candidates, the one with the minimum joint score is selected.

The Transformer module is trained offline using service traces generated under different traffic loads and network states. Each training sample consists of a service request, the corresponding network snapshot, and the target provisioning decision obtained from the MILP model in small-scale scenarios or from the best-performing heuristic decision in large-scale scenarios. The training objective is expressed as

$$\mathcal{L} = \mathcal{L}_R + \xi_b \mathcal{L}_{\text{block}} + \xi_c \mathcal{L}_{\text{cost}} + \xi_v \mathcal{L}_{\text{vio}}. \tag{35}$$

where $\mathcal{L}_R$ guides the predicted link weights toward the target route, $\mathcal{L}_{\text{block}}$ penalizes blocked service requests, $\mathcal{L}_{\text{cost}}$ penalizes high-cost provisioning decisions, and $\mathcal{L}_{\text{vio}}$ penalizes violations of OSNR, resource capacity, and network resource conflict constraints. The loss coefficients were selected through sensitivity tests on n6s9 and USNET using matched traffic traces. Results showed that $(\xi_b, \xi_c, \xi_v) = (2.0, 0.25, 2.0)$ provided a better balance between service satisfaction and transceiver cost. The same coefficient values were therefore used for both topologies. After training, the Transformer is used online to rapidly infer adaptive link weights, while the subsequent waveplane-based RMGWA procedure guarantees feasible resource assignment.

6.4. TSAA Algorithm Pseudocode

The TSAA pseudocode describes the provisioning process. TSAA first combines the identified service type and service attributes with the present network state to generate adaptive link weights using $\Pi(\theta)$. For each candidate route, the waveplane-based RMGWA procedure evaluates all feasible provisioning configurations and selects the best one. Once a configuration is provisioned, the network state is updated, while resources are released when the service expires. The pseudocode of the TSAA strategy is shown as follows (Algorithm 1):

Algorithm 1: TSAA Strategy

Input: a network topology $G = (N, L)$, traffic demand $p(s, d, \lambda_{sd}, t_{ptr}, \rho) \in P$, M, E, $L_{i,j}^{p}$, Transformer weight function $\Pi(\theta)$.

Set $C_p = \infty$, $r_{s,d}^{p} =$ null, $m_p =$ null, $e_p =$ null, $w_p =$ null, Block = null.
  For each $p(s, d, \lambda_{sd}, t_{ptr}, \rho)$ in P:
    $E_p =$ map to Granularity(BW) // Map the granularity to a standard vector format.
    $F_p =$ Extract Features(G, p) // Extract the features of network topology, service attributes, and transmission requirements.
    $W_L = \Pi(G, E_p, F_p)$ // The Transformer performs inference based on the extracted granularity features, network topology features, and service features, and outputs the global weights for each link.
    $G_p =$ Prune_Graph$(G, E, L_{i,j}^{p})$ // Network topology pruning, thereby maximizing network utilization.
    $R_c =$ K_shortest_paths(G_p, s, d, W_L, M, K) // Obtain candidate routing paths.
    For each path R_k in $\boldsymbol{R}_c$:
      If w_k is not None and Spectrum$_k \geq$ Spectrum$_p$:
        Evaluate feasible modulation format m_k, switching granularity e_k, and wavelength w_k.
        $m_p = m_k$, $e_p = e_k$, $w_p = w_k$.
        $r_{s,d}^{p} = R_k$, $C_p = C_k$.
      Else If:
        $r_{s,d}^{p} =$ null, Block = Block + 1.
        Break.
        Config$_n = (p, L_{i,j}^{p}, m_p, e_p, w_p)$ // The Transformer pre-configures the nodes along the route.
      End If.
    End For.
    If $t \geq t_{ptr}$:
      $m_p =$ null, $e_p =$ null, $w_p =$ null, $r_{s,d}^{p} =$ null // Release the resource of the expired service.
    End If.
    If $\frac{\sum_{i=1}^{n} BW_i}{BW_{\text{total}}} \geq \epsilon$ or Block $\geq X$:
      Update network state and resource occupation.
    End If.
  End For.
End.

The computational complexity of the above strategy is at the level of $\mathcal{O}(|\mathbf{P}| \cdot |\mathbf{E}| \cdot |\mathbf{M}|)$, where $|\mathbf{P}|$ denotes the total number of service requests, $|\mathbf{E}|$ represents the number of available switching granularities, and $|\mathbf{M}|$ indicates the number of available modulation formats.

6.5. Waveplane-Based RMGWA Algorithm

We develop a waveplane-based RMGWA algorithm by extending the conventional waveplane heuristic [22]. After CP, QSP, or TSAA generates candidate provisioning preferences, the proposed algorithm performs granularity-aware resource search and completes feasible resource assignment. Unlike the conventional wavelength-plane representation, it jointly models wavelength, waveband, and sub-wavelength resources according to the selected switching granularity.

The WP algorithm integrates the routing and wavelength assignment steps, effectively reducing the blocking probability of traditional RWA algorithms and improving network resource utilization. The core idea is to construct multiple virtual topologies (referred to as wavelength planes) based on the physical network topology. In each wavelength plane, a virtual link corresponds to a specific wavelength on a physical fiber link, and each node in the plane directly maps to a node in the physical topology. In multi-granularity optical networks, the concept of wavelength planes is further extended to support three levels

of switching granularity: (a) Wavelength Plane: Each virtual link represents a specific wavelength on a fiber link. (b) Waveband Plane: Each virtual link represents a specific waveband, composed of multiple consecutive wavelengths. (c) Sub-wavelength Plane: Each virtual link represents a specific sub-wavelength channel on a fiber link.

When constructing wavelength planes, it is essential to first determine the wavelength usage status of each link. If a particular wavelength is already occupied on a given link, that link is removed from the corresponding wavelength plane. Figure 5 presents an example of wavelength planes in a network. Taking Wavelength Plane 1 as an example, suppose wavelength λ_1 on link 2–4 in the network topology is occupied. Consequently, this link is removed from Wavelength Plane 1, forming a new virtual topology. Following this method, if the maximum number of available wavelengths per link is W, then W wavelength planes can be constructed. If each waveband contains B wavelengths, W/B waveband planes can be created. Furthermore, if each wavelength can be subdivided into S sub-wavelengths, a total of $W \times S$ sub-wavelength planes can be constructed. As shown in Figure 5, based on the usage status of wavelengths λ_0, λ_1, and λ_2 in the network, three wavelength planes, one waveband plane (with $B = 3$), and six sub-wavelength planes (with $S = 2$) can be constructed.

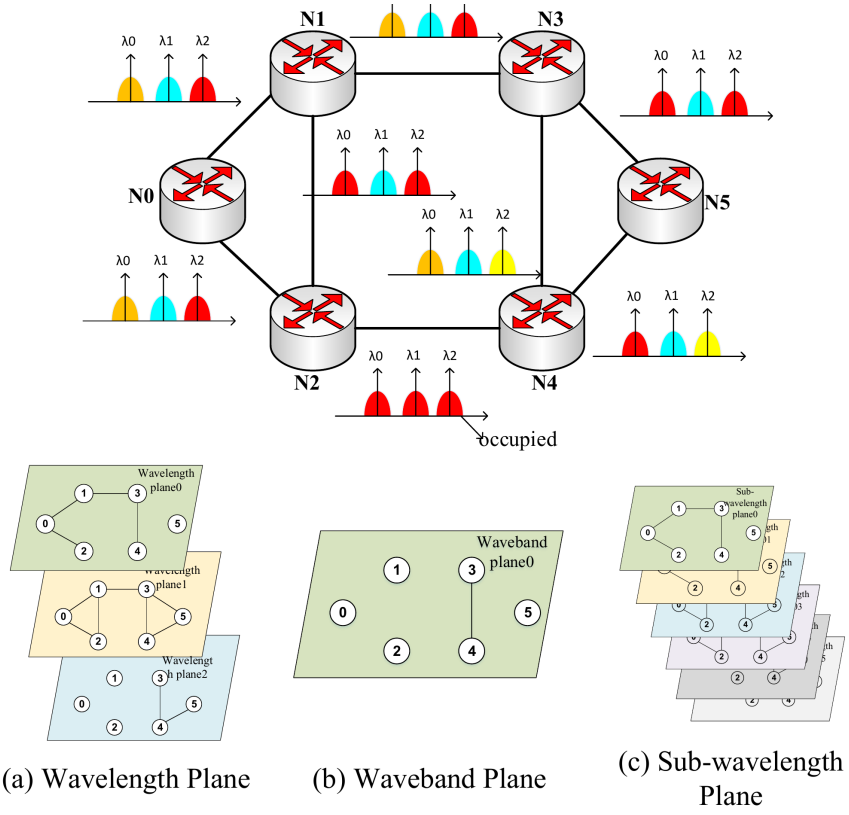


Figure 5. Example of waveplane.

Based on the WP thus constructed, we implement the RMGWA algorithm for provisioning lightpaths in the MGON. The pseudocode of the algorithm is as follows (Algorithm 2):

The time complexity of the algorithm is determined by the selected granularity adaptation strategy. The time complexity of CP, QSP or the TSAA strategy are equal as $\mathcal{O}(|\mathbf{P}| \cdot |\mathbf{E}| \cdot |\mathbf{M}|)$.

Algorithm 2: Waveplane-based RMGWA

Input: $G_p = (N, L)$, $p(s, d, \lambda_{sd}, t_{pg}, t_{ptr}) \in P$.

If $\text{Flag}_p == 0$ then:
  Obtain the list of all available routes for service request p, denoted as $\boldsymbol{RL}_p$.
  Based on OSNR calculation, determine the set of available modulation formats M_p on the shortest route r for service p.
  Select the switching granularity e_p^* and modulation format m_p^* for service p from candidate provisioning combinations which is formed by CP, QSP or the TSAA strategy.
  Update $\boldsymbol{RL}_p$ by removing all routes that do not support the selected modulation format m_p^*.
  Generate the corresponding waveplane list $\boldsymbol{\Theta}_w$ based on the selected switching granularity e_p^*, as illustrated in Figure 5.
  For each $\omega \in \boldsymbol{\Theta}_w$:
    If $R \in \boldsymbol{RL}_p$ and $R \in \omega$ then:
      Allocate bandwidth to service p on route R and establish the lightpath.
      $\text{Flag}_p = 1$.
    End If.
  End For.
End If.
End.

7. Simulation Setup and Result Analysis

7.1. Test Conditions

To evaluate the performance of the proposed strategies of switching mode adaptation for lightpath services, we consider two test networks, i.e., the 6-node, 9-link (n6s9) network, and the 24-node, 43-link (USNET) network, as shown in Figure 6. The n6s9 network serves as the testing ground for evaluating the outcomes of the model and strategies, while the USNET network is employed to appraise the results obtained from various strategies. For a fair comparison, all strategies are evaluated under the same service-arrival sequences, topology settings, network resource and transceiver-cost parameters. The CP, QSP, and TSAA strategies are all implemented within the same waveplane-based RMGWA framework.

Table 4 summarizes four typical service models: (1) The On–Off Model [23] represents high-bandwidth traffic such as data center services. Here, λ is the probability of transitioning from "on" to "off", μ is the reverse, h is the transmission rate, and θ is the exponential parameter (set to 1 in this paper). (2) The Fixed-Period Model [24] represents low-latency traffic such as 5G fronthaul. σ is the data generated per period, and τ is the period length. (3) The Bernoulli Distribution Model [14] represents high-QoS traffic such as wireless services. With probability p, the service generates σ units of data per unit time. (4) The Compound Poisson Model [25] represents general-purpose traffic (e.g., LTE-A), which has no strict latency or QoT requirements. Here, λ is the packet arrival rate, and $1/\sigma$ is the average packet size.

We assume that the sub-wavelength and wavelength account for a fixed grid at 12.5 GHz and 50 GHz, respectively. Furthermore, the waveband consists of two 50 GHz wavelengths, occupying a fixed 100 GHz grid, which can address interoperability issues of flexible grid network over legacy fixed grid networks. The total number of available wavelengths in each fiber link is 15. The bandwidth required by each service is uniformly distributed in the range of [20, 800] Gb/s to construct a balanced mixed-service workload. The parameters of the arrival curve are shown in Table 4.

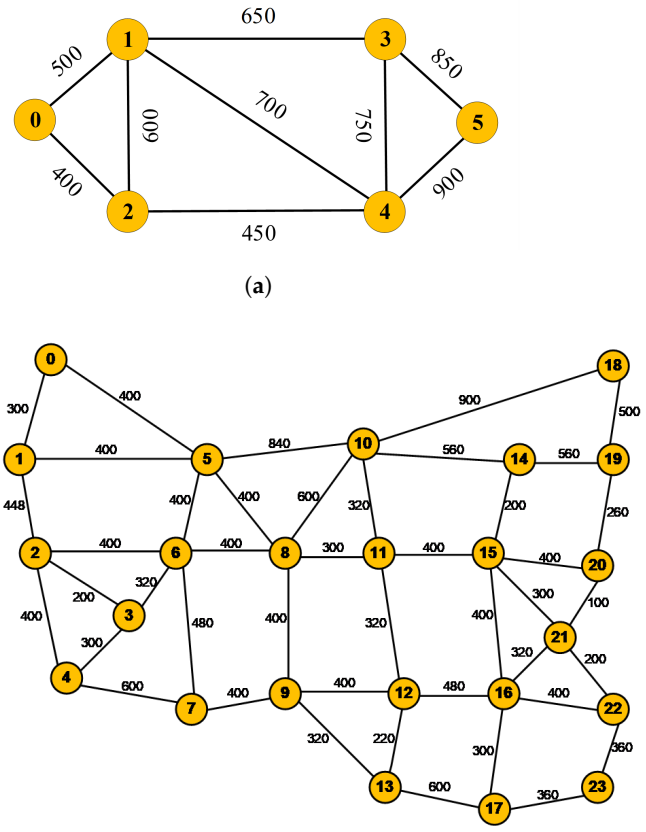


Figure 6. Test networks: (**a**) six-node, nine-link n6s9; (**b**) 24-node, 43-link USNET.

Table 4. Arrival curves and parameter values of four typical service models.

Service Model	Arrival Curve $\alpha(t)$	Parameter	Value
On–Off Model	$\frac{t}{2\theta}(\theta h - \lambda - \mu) + \frac{t}{2\theta}\sqrt{(\theta h + \lambda - \mu)^2 + 4\lambda\mu}$	λ, μ	0.5, 0.5
		h	7×10^9 bit/s
Fixed-Period Model	$\sigma\lfloor t/\tau \rfloor$	τ	0.6 ms
		σ	0.8 Mbits
Bernoulli Distribution Model	$\frac{t}{\theta}\log(1 - p + pe^{\theta\sigma})$	p	0.2
		σ	2×10^9 bit/s
Compound Poisson Model	$\lambda t/(\sigma - \theta)$	λ	2.5×10^4 per/s
		σ	1.25×10^{-4}

In this work, we consider four candidate modulation formats (including BPSK, QPSK, 8-QAM, and 16-QAM) for lightpath establishment. The spectrum efficiency and OSNR thresholds for these modulation formats are shown in Table 5.

Table 5. Modulation formats, spectrum efficiencies, and OSNR thresholds [26].

Modulation Format	Spectrum Efficiency	OSNR Threshold (dB)
BPSK	1	9.0
QPSK	2	12.0
8-QAM	3	16.0
16-QAM	4	18.6

Table 6 summarizes the computational resource consumption of the Transformer module. For each incoming request, TSAA generates link weights through a single forward pass, and the effect of this online computation on provisioning time is reflected in the average completion time results discussed below.

Table 6. Computational resource consumption of the TSAA control module.

Network Topology	Transformer Parameters	Forward Computational Cost/Service	FP32 Parameter Memory
n6s9	73,153	2.834 MFLOPs	0.292 MiB
USNET	75,457	15.681 MFLOPs	0.301 MiB

The MILP model was solved using AMPL with Gurobi 12.0.1 [27], and the heuristic strategies were implemented in Python 3.10.19.

The TSAA link-weight predictor was trained using dynamically generated service requests following a Poisson arrival process. The training load was progressively increased from 250 to 450 Erlang by adjusting the Poisson arrival rate, and each load level was maintained for 1000 time slots, resulting in approximately 87,500 service requests for each topology. The proportions of bandwidth-intensive, delay-sensitive, and QoS-sensitive services were set to 0.4, 0.3, and 0.3, respectively, to construct an approximately balanced synthetic workload with sufficient representation of each service class. The Transformer training configurations are shown in Table 7.

Table 7. TSAA training configuration.

Configuration	Value	Mode
Batch Size	64	Offline
Encoder Layer	2	Offline
Attention Heads	4	Offline
Learning Rate	1×10^{-4}	Offline
Hidden-Dimension	64	Offline
Feed-Forward Dimension	128	Offline

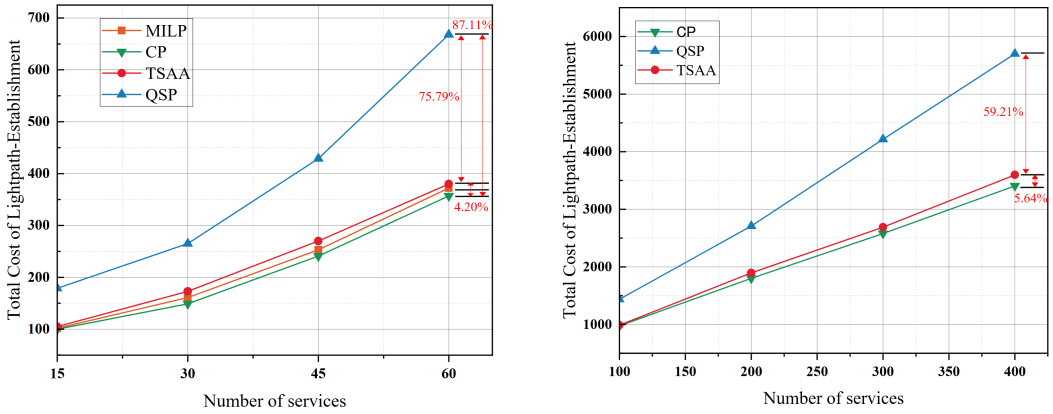
7.2. Lightpath Establishment Cost

This section compares the total cost incurred during lightpath establishment under different strategies. The results for the n6s9 and USNET networks are presented in Figure 7, where the x-axis denotes the number of lightpath requests and the y-axis represents the total cost. The labels used in the figure are as follows: (1) "MILP" refers to the mixed-integer linear programming optimization model; (2) "CP" denotes the Cost-Priority strategy; (3) "QSP" stands for the Quality-of-Service Priority strategy; and (4) "TSAA" represents the Transformer-assisted service-attribute-aware strategy. For fair comparison, all strategies are implemented based on the waveplane-based RMGWA algorithm. The results from Figure 7 reveal several key findings:

Firstly, as the number of lightpath requests increases, the total cost rises accordingly. This is due to each additional service requiring dedicated transceivers and wavelengths to establish its lightpath, leading to an overall increase in the cost. This trend is consistently observed in both network scenarios.

Secondly, among the three strategies, the CP strategy yields the lowest cost, as it always selects the minimum-cost configuration without considering service-specific transmission requirements. In contrast, the QSP and TSAA strategies optimize lightpath allocation based on individual service demands, resulting in relatively higher costs. Specifically, the cost under the QSP strategy is significantly higher than that of the CP strategy, with increases of 75.79% and 59.21% in the n6s9 and USNET networks, respectively. This demonstrates that prioritizing service quality often leads to substantial cost overhead.

On the other hand, the TSAA strategy achieves a more balanced trade-off between cost and service quality. While its cost is slightly higher than that of the CP strategy, the increase is only 4.20% and 5.64% in the n6s9 and USNET networks respectively, far below the increases seen in the QSP strategy. Moreover, the cost curve of the TSAA strategy closely aligns with that of the MILP model, demonstrating that the proposed adaptive link-weight mechanism improves provisioning decisions without causing excessive establishment cost.



(**a**) n6s9. (**b**) USNET.

Figure 7. Total cost of lightpath establishment versus number of services under different strategies.

7.3. Service Satisfaction Ratio

In this section, the performance of the proposed scheme is evaluated through the satisfaction ratio, which is defined by Equation (36).

$$S_R = \frac{N_{\text{satisfied}}}{N_{\text{total}}} \times 100\%, \tag{36}$$

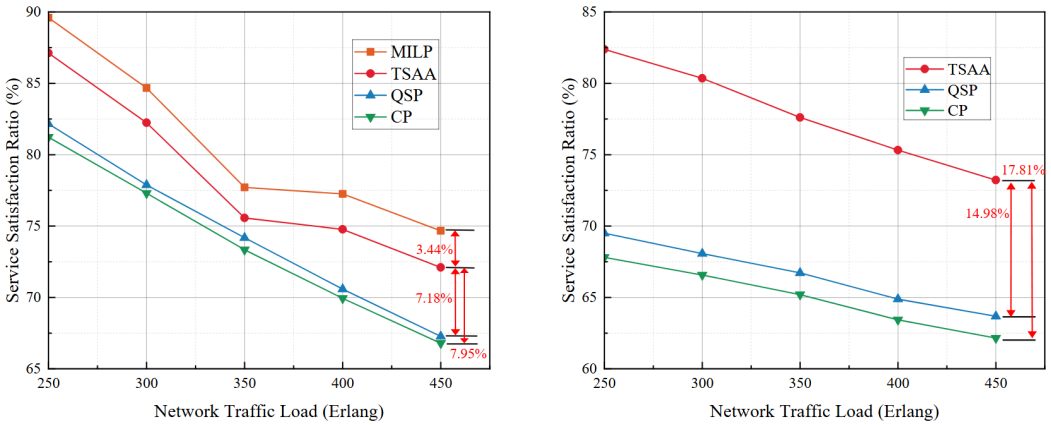
where $N_{\text{satisfied}}$ denotes the number of service requests that are successfully established and meet their corresponding service requirements, and N_{total} denotes the total number of generated service requests. A higher service satisfaction ratio indicates that more services can be provisioned while satisfying their required transmission conditions.

Figure 8 compares the service satisfaction ratio under different traffic loads. In the n6s9 network, the MILP result is included as a benchmark, whereas it is not reported for USNET because the global optimization model becomes computationally prohibitive in the larger topology.

As the traffic load increases from 250 to 450 Erlang, the satisfaction ratio of all schemes gradually decreases, mainly because more service requests compete for limited wavelength resources, switching granularities, and feasible transmission paths. Nevertheless, TSAA consistently maintains a higher satisfaction ratio than the heuristic baselines. In the n6s9 network, although MILP achieves the best result due to its global optimization capability, TSAA closely approaches the MILP benchmark, with only a 3.44% gap at 450 Erlang. More importantly, TSAA outperforms QSP and CP by 7.18% and 7.95%, respectively. This improvement indicates that TSAA can dynamically adjust link weights according to both service requirements and real-time resource states, thereby guiding candidate-route generation toward paths with higher provisioning feasibility.

Moreover, in the larger USNET network, the advantage of TSAA becomes more pronounced. At 450 Erlang, TSAA improves the satisfaction ratio by 14.98% over QSP and 17.81% over CP, which shows that the proposed strategy is particularly effective when the topology contains more routing alternatives and the resource distribution becomes more complex. Therefore, the results demonstrate that TSAA not only improves the probability

of successful lightpath establishment, but also enhances the ability of MGONs to satisfy heterogeneous service requirements under dynamic traffic conditions.



(**a**) n6s9. (**b**) USNET.

Figure 8. Service satisfaction ratio versus offered traffic load under different strategies.

7.4. Spectrum Utilization

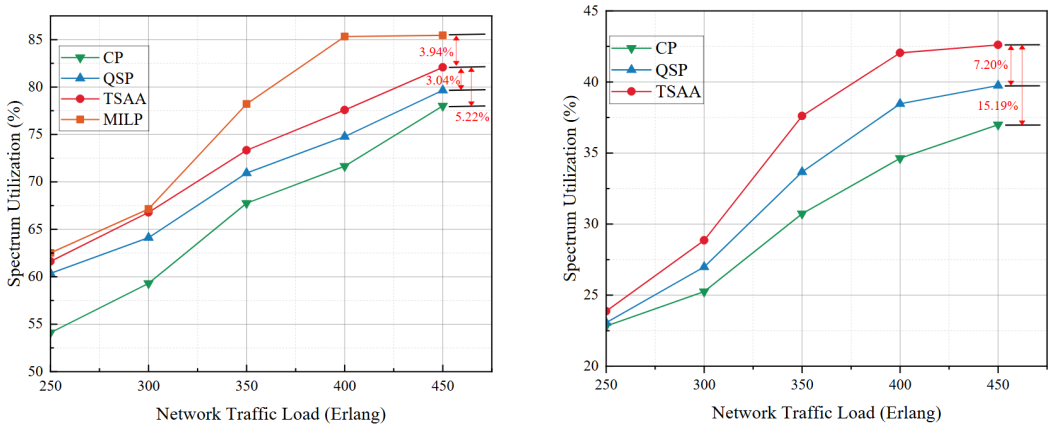
Figure 9 illustrates spectrum utilization, which is determined by Equation (18) under different strategies in the n6s9 and USNET networks as a function of the offered traffic load. A higher spectrum utilization indicates that the algorithm can more efficiently allocate services to appropriate spectral resources, thereby avoiding blocking and ensuring efficient network operation. This demonstrates the algorithm's stronger capability to perceive and schedule network spectral resources.

As the traffic load increases from 250 to 450 Erlang, spectrum utilization increases for all strategies because more services occupy the available wavelength resources. In the n6s9 topology, MILP achieves the highest utilization due to its global optimization capability, while TSAA consistently outperforms CP and QSP.

However, TSAA still remains close to the MILP benchmark, with only a 3.94% gap at 450 Erlang, and consistently outperforms the others. In the n6s9 topology, TSAA strategy improves spectrum utilization by up to 5.22% and 3.04% relative to CP strategy and QSP strategy respectively.

Spectrum utilization in the USNET network also increases with traffic load, but due to its larger spectral resource pool, the values are lower than those in the n6s9 network under the same load conditions. Despite this, TSAA maintains its leading edge. On average across the entire load range, TSAA strategy achieves a spectrum utilization that is 15.19% higher than that of CP strategy and 7.20% higher than that of QSP strategy. This outcome underscores TSAA's scalability and its ability to deliver superior spectral efficiency even in large-scale topologies where resources are more abundant.

The underlying reason lies in TSAA's adaptive selection of modulation formats and switching granularities according to service attributes, which avoids both over-provisioning and under-utilization by efficiently packing services into the available spectrum. In contrast, the CP strategy, driven solely by cost minimization, often leaves spectrum underused, while the QSP strategy, in its pursuit of strict quality guarantees, tends to reserve excessive resources. QSP's performance falls between that of CP and TSAA because its quality-first orientation inherently trades off spectral efficiency. These observations collectively confirm that the TSAA strategy strikes an effective balance between resource efficiency and service quality.

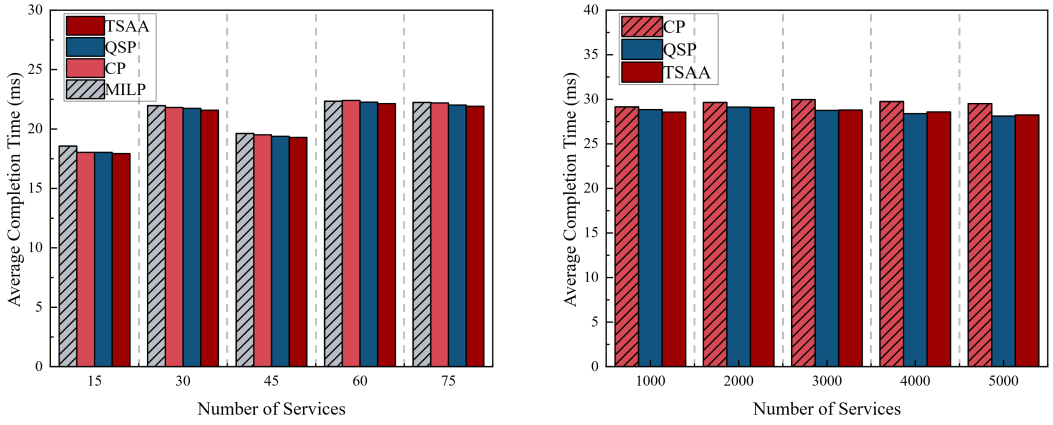


(**a**) n6s9. (**b**) USNET.

Figure 9. Spectrum utilization versus offered traffic load under different strategies.

7.5. Average Completion Time

We further compare the different strategies in terms of the average completion time, which is defined as the average elapsed time from the generation of a service request to the completion of its lightpath provisioning. It includes candidate route generation, route selection, and resource assignment for all strategies, with feature preparation and Transformer inference additionally included for TSAA. For the MILP benchmark, the reported average completion time excludes the solution time. Figure 10 presents the results where TSAA maintains a stable and competitive performance in both the n6s9 and USNET networks. Although TSAA introduces a Transformer-assisted control module to predict specific link weights, the results show that this additional module does not increase average completion time. In the n6s9 network, the average completion time of TSAA remains close to the MILP benchmark under different service loads, indicating that the Transformer-assisted routing guidance can approach the optimized routing behavior without causing additional service delay. In the larger USNET network, TSAA also keeps average completion time at a comparable level to CP and QSP as the number of services increases. This is because the Transformer module rapidly maps service attributes and network states into adaptive link weights, which accelerates candidate route selection and avoids repeated path evaluation. Therefore, TSAA improves the intelligence of service-aware routing while maintaining efficient service provisioning, demonstrating that the introduced control module does not impose extra delay on the lightpath establishment process.



(**a**) n6s9. (**b**) USNET.

Figure 10. Average completion time versus number of services under different strategies.

8. Conclusions

This paper investigated service-aware lightpath provisioning in multi-granularity optical networks, where heterogeneous services have different bandwidth, delay, and quality requirements. To address the mismatch between service attributes and network resource allocation, we developed an integrated framework for RMGWA in MGONs. TSAA predicts adaptive link weights from service attributes and network states to guide candidate-route selection, while the proposed waveplane-based RMGWA algorithm searches wavelength, waveband, and sub-wavelength planes to complete feasible resource allocation under multi-granularity constraints. Simulation results on n6s9 and USNET show that the framework improves service satisfaction and spectrum utilization while maintaining low lightpath establishment cost. Moreover, the Transformer-assisted module provides adaptive routing guidance without increasing average completion time. The proposed framework therefore offers an effective and scalable solution for intelligent lightpath provisioning in MGONs.

Author Contributions: Conceptualization, Y.L.; methodology, Y.L. and X.Z.; software, X.Z.; validation, X.Z. and Z.Y.; formal analysis, X.Z. and Z.Y.; investigation, X.Z.; writing—original draft preparation, X.Z. and Z.Y.; writing—review and editing, V.B. and Y.L.; supervision, Y.L.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was jointly supported by the National Key R&D Program of China under Grant 2023YFB2905505 and the National Natural Science Foundation of China (NSFC) under Grant 62271338.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and code supporting the reported results are available from the corresponding author upon reasonable request.

Acknowledgments: The authors thank the reviewers and colleagues who provided constructive feedback on this work.

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

References

1. Zervas, G.S.; De Leenheer, M.; Sadeghioon, L.; Klonidis, D.; Qin, Y.; Nejabati, R.; Simeonidou, D.; Develder, C.; Dhoedt, B.; Demeester, P. Multi-Granular Optical Cross-Connect: Design, Analysis, and Demonstration. *J. Opt. Commun. Netw.* **2009**, *1*, 69–84. [CrossRef]
2. Huang, Z.; Zhu, B.; Zhu, M.; Jiang, M.; Song, X.; Zhao, J.; Wang, Z.; Hu, F. A novel multi-granularity flexible-grid switching optical-node architecture. *China Commun.* **2023**, *20*, 209–217. [CrossRef]
3. Zhang, J.; Zhu, B.; Yan, X.; Chen, C.; Jiang, M.; Hu, F. A novel multi-granularity two-layer SDM ROADM architecture. *Opt. Commun.* **2021**, *479*, 126473. [CrossRef]
4. Hou, W.; Guo, L.; Wang, X.; Cao, J.; Wu, J.; Li, Y. A New Routing Algorithm Based on Integrated Grooming Auxiliary Graph in Multi-granularity Optical Networks. In *Proceedings of the INC, IMS and IDC, Seoul, Republic of Korea, 25–27 August 2009*; IEEE: Piscataway, NJ, USA, 2009; pp. 462–467.
5. Liu, W.; Yang, H.; Li, J.; Shi, Y.; Li, Z.; Zhao, Y. A Protection Scheme Based on Multi-Granularity Connection Resource Sharing in Optical Networks. In *Proceedings of the ACP/POEM, Wuhan, China, 4–7 November 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–4.
6. Wang, Y.; Cao, X. Multi-Granular Optical Switching: A Classified Overview for the Past and Future. *IEEE Commun. Surv. Tutor.* **2012**, *14*, 698–713. [CrossRef]
7. Izmailov, R.; Kolarov, A.; Fan, R.; Araki, S. Hierarchical optical switching: A node-level analysis. In *Proceedings of the HPSR, Kobe, Japan, 26–29 May 2002*; IEEE: Piscataway, NJ, USA, 2002; pp. 309–313.
8. Zhang, G.; Xiong, Q.; Shen, S.; Ye, Y. Novel multi-granularity optical switching node with wavelength management pool resources. In *Proceedings of the ACP, Shanghai, China, 2–6 November 2009*; IEEE: Piscataway, NJ, USA, 2009; pp. 1–5.
9. Zhang, S.; Wang, X.; Huang, M. A Multi-granularity Grooming Scheme for One-to-Many Multicast Traffic. In *Proceedings of the DCABES, Xianning, China, 24–27 November 2014*; IEEE: Piscataway, NJ, USA, 2014; pp. 215–219.

10. Wang, J.; Zhang, P. Based on MG-OXC Routing and Multi-Granularity Assignment Algorithm. In *Proceedings of the CSRSWTC, Chengdu, China, 21–25 July 2013*; IEEE: Piscataway, NJ, USA, 2013; pp. 420–422.
11. Dharmaweera, M.N.; Zhao, J.; Yan, L.; Karlsson, M.; Agrell, E. Traffic-grooming- and multipath-routing-enabled impairment-aware elastic optical networks. *J. Opt. Commun. Netw.* **2016**, *8*, 58–70. [CrossRef]
12. Zhang, J.; Ji, Y.; Song, M.; Zhao, Y.; Yu, X.; Zhang, J.; Mukherjee, B. Dynamic Traffic Grooming in Sliceable Bandwidth-Variable Transponder-Enabled Elastic Optical Networks. *J. Light. Technol.* **2015**, *33*, 183–191. [CrossRef]
13. Tomkos, I.; Rožić, Ć.; Savi, M.; Sköldström, P.; Lopez, V.; Chamania, M.; Siracusa, D.; Matrakidis, C.; Klonidis, D.; Gerstel, O. Application Aware Multilayer Control and Optimization of Elastic WDM Switched Optical Networks. In *Proceedings of the OFC, San Diego, CA, USA, 11–15 March 2018*; Optica Publishing Group: Washington, DC, USA, 2018; pp. 1–3.
14. Metzger, F.; Hossfeld, T.; Bauer, A.; Kounev, S.; Heegaard, P.E. Modeling of Aggregated IoT Traffic and Its Application to an IoT Cloud. *Proc. IEEE* **2019**, *107*, 679–694. [CrossRef]
15. Santos, A.S.; de Santi, J.; Figueiredo, G.B.; Mukherjee, B. Application-Aware Service Degradation in Elastic Optical Networks. *IEEE Trans. Netw. Serv. Manag.* **2022**, *19*, 949–961. [CrossRef]
16. Yang, H.; Zhu, Z. Application-aware Configuration of All-optical Interconnects in Hyper-FleX-LION. In *Proceedings of the OGC, Shenzhen, China, 6–11 December 2022*; IEEE: Piscataway, NJ, USA, 2022; pp. 214–217.
17. Zhou, Z.; Zhang, H.; Zhu, W.; Fu, Z.; Gu, R. Spatial-Temporal Prediction Model for Optical Fronthaul Network Traffic Based on ASTFN. In *Proceedings of the ICOCN, Qufu, China, 31 July–3 August 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–3.
18. Drainakis, G.; Baziana, P.; Bogris, A. ALAS-DCN: Application-Aware Low-Latency Scalable Data Center Network Architecture for Intra-Rack and Inter-Rack Optical Packet Switching. *J. Opt. Commun. Netw.* **2026**, *18*, 409–423. [CrossRef]
19. Du, Y.; Sun, Y.; Gao, Y.; Cao, X. Layered Multi-Granularity CDC-ROADM Node for Quantum-Classical SDM-WDM Optical Networks. In *Proceedings of the 2025 9th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC), Beijing, China, 7–9 November 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 286–290.
20. Forrester, R.J.; Waddell, L.A. Linearization Strategies for Binary Quadratic and Higher-Order Polynomial Programs. In *Encyclopedia of Optimization*; Pardalos, P.M., Prokopyev, O.A., Eds.; Springer: Cham, Switzerland, 2023.
21. Yuan, Z.; Zhou, H.; Li, Y.; Zhang, J.; Shen, G. Strategies of Switching Granularity Selection for Lightpath Services in a Multi-Granularity Optical Network. In *Proceedings of the ACP, Wuhan, China, 4–7 November 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–3.
22. Shen, G.; Wu, C.; Dong, J. A Waveplane-Based Regenerator Placement Approach for Minimizing Numbers of Regenerators and Wavelengths. In *Proceedings of the ACP, Beijing, China, 12–15 November 2013*; Optica Publishing Group: Washington, DC, USA, 2013; p. AF4F.5.
23. Marvi, M.; Aijaz, A.; Khurram, M. Integrating Stochastic Geometry and ON/OFF Traffic Models: Toward Spatio-Temporal Analysis of Wireless Networks With Heterogeneous Services. *IEEE Trans. Netw. Sci. Eng.* **2022**, *9*, 1668–1679. [CrossRef]
24. Itou, S.; Uchiyama, K.; Shioda, S. Fixed-Period Packet Sampling and its Application to Flow Rate Estimation. In *Proceedings of the ICC, Glasgow, UK, 24–28 June 2007*; IEEE: Piscataway, NJ, USA, 2007; pp. 279–286.
25. Yang, H.H.; Arafa, A.; Quek, T.Q.S.; Poor, H.V. Locally Adaptive Scheduling Policy for Optimizing Information Freshness in Wireless Networks. In *Proceedings of the GLOBECOM, Waikoloa, HI, USA, 9–13 December 2019*; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6.
26. Shen, G.; Li, Y.; Zhao, H. Quasi-CWDM optical network: Cost effective and spectrum efficient architecture for future optical networks. In *Proceedings of the ONDM, Pisa, Italy, 11–14 May 2015*; IEEE: Piscataway, NJ, USA, 2015; pp. 1–6.
27. Gurobi Optimization, LLC. *Gurobi Optimizer Reference Manual*; Gurobi Optimization, LLC: Beaverton, OR, USA, 2026. Available online: https://www.gurobi.com (accessed on 18 September 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.