# Intent-Aware Reinforcement Learning Control for Point-to-Multipoint Light-Trees in Metro-Aggregation Networks

Polyzois Soumplis*†, Konstantinos Christodoulopoulos*‡, Konstantinos Yiannopoulos*§, Emmanouel Varvarigos*†
*Institute of Communication and Computer Systems (ICCS), 9 Iroon Polytechniou Str., 15773 Zografou, Greece
†School of Electrical and Computer Engineering, National Technical University of Athens, Zografou, Greece
‡Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Greece
§Department of Informatics and Telecommunications, University of the Peloponnese, Tripoli, Greece
soumplis@mail.ntua.gr

***Abstract*—Traffic in metro-access networks is becoming increasingly dynamic and diverse, driven by edge computing and the demands of bursty applications. Under these evolving conditions, optical aggregation must allocate resources while balancing latency, efficiency, and stability. Point-to-multipoint (P2MP) coherent transceivers with digital subcarrier multiplexing enable efficient aggregation by allowing one hub transceiver to serve multiple spokes via a light-tree, reducing transceiver count and energy consumption relative to point-to-point designs. However, configuring P2MP light-trees in dynamic settings is challenging, as resource allocation must satisfy optical feasibility constraints (e.g., OSNR thresholds), respect transceiver subcarrier capacity limits, and meet strict per-spoke performance objectives such as low latency, high spectral efficiency, or low-churn assignments. These objectives are often implicit, user-specific, and may shift with demand patterns and traffic cycles. In this work, we address this challenge with an intent-aware reinforcement learning controller that infers hidden service objectives online from historical interactions and allocates resources accordingly. Specifically, an encoder network learns a latent intent vector for each spoke that guides the policy by producing hub-selection preference weights and stability margins. The controller jointly optimizes the selection of the hub and subcarrier allocation to satisfy quality-of-transmission requirements while aligning with inferred intents. Evaluation on realistic metro topologies shows gains over intent-agnostic RL and static baselines up to 37% lower latency for latency-sensitive traffic, 25% higher spectral utilization, and 75% less reconfiguration overhead for stability-oriented services.**

***Index Terms*—Intent Based Networking, Reinforcement Learning, Point-to-multipoint communication**

## I. Introduction

Metro-aggregation networks form the bridge between access infrastructure and core backbone networks, carrying diverse traffic from residential users, enterprise clients, and emerging edge computing applications. These networks are evolving to support heterogeneous quality-of-service (QoS) requirements, including stringent latency bounds for real-time applications, high spectral efficiency for cost-sensitive bulk transport, and stable low-churn assignments for carrier-grade services [1].

This work was supported by the EU Horizon 2022 ALLEGRO project (Grant Agreement No. 101092766)

Point-to-multipoint coherent transceivers have recently emerged as an attractive solution for metro-aggregation environments. By enabling a single hub transceiver to serve multiple remote spokes over a shared optical light-tree using digital subcarrier multiplexing, P2MP systems can reduce the required transceiver count compared to conventional point-to-point (P2P) architectures. However, realizing this requires careful configuration decisions involving hub placement, spoke-to-hub subcarrier allocation, and when necessary, P2P links to forward aggregated traffic to backbone nodes [2]. These decisions must jointly satisfy optical signal-to-noise ratio (OSNR) constraints, respect per-transceiver subcarrier capacity limits, and adapt to time-varying demands [3].

As modern metro networks serve a heterogeneous set of users and services, they exhibit latent and time-varying preferences regarding latency, cost efficiency, spectral efficiency, or assignment stability. These are not explicitly provided to the controller and we refer to these preferences as intents [4]. The allocation of demands to light-trees (subcarriers and hub nodes) can be modeled as a generalized Bin Packing problem with variable item sizes, where the spectral resource consumption of a demand depends on the physical proximity to the selected hub (bin). However, re-solving this combinatorial optimization repeatedly under dynamic demands is computationally expensive, while operator-defined cost weights cannot accurately reflect per-spoke behavioral differences.

In addition, user intents are not explicitly provided to the network and must be inferred from past interactions, making traditional optimization-based control even less viable. Reinforcement learning (RL) provides an appealing alternative for sequential decision-making under dynamics and long-term effects, and has been widely explored for optical routing and resource allocation under QoT constraints [5].

In this work, we propose an intent-aware RL controller for dynamic P2MP light-tree configuration that explicitly addresses these limitations. We formulate the sequential decision problem as an infinite-horizon discounted Markov decision process (MDP) and develop an approximate dynamic programming (ADP) solution based on rollout policy improvement.

To handle the combinatorial complexity of joint spoke-to-hub assignment under shared port capacity constraints, we employ a greedy heuristic guided by the learned value function and intent-driven preference scores. Extensive numerical experiments on realistic metro topologies demonstrate that our intent-aware controller consistently outperforms both intent-agnostic RL baselines and static optimization heuristics across multiple performance dimensions. The rest of the paper is organized as follows: Section II reviews related work; Section III presents the system model; Section IV formulates the problem as an MDP; Section V describes the proposed ADP/RL solution; Section VI presents the experimental evaluation; and Section VII concludes the paper.

## II. Related Work

Point-to-multipoint (P2MP) coherent transmission leveraging digital subcarrier multiplexing has been investigated for metro-aggregation networks with hub-and-spoke traffic patterns. In [6], [7], the authors quantified the architectural and techno-economic benefits of replacing point-to-point (P2P) transceivers with P2MP coherent transceivers, resulting in reduced transceiver count and lower aggregation complexity in realistic metro settings. The authors in [8] studied multi-period planning and capacity dimensioning, typically using ILP formulations that minimize transceiver cost under forecast traffic profiles. In prior work we showed deeper P2MP root placement unlocks multiplexing gains [2] and studied QoS-aware and risk-aware online reconfiguration [9], [10]. Recent contributions have also extended the use of P2MP coherent pluggables to more general WDM/elastic optical architectures, including wavelength-switched optical networks with hub-and-spoke traffic [11]. In all these works, the optimisation objectives are primarily cost- and capacity-oriented, and the decision variables are hub placement, spoke assignment and transceiver type, solved via ILPs or heuristic approaches.

Intent-based networking (IBN) and zero-touch network and service management (ZSM) are widely viewed as the foundation paradigms for automating complex decision processes. The authors in [12] stress the role of AI-driven closed-loop management in mapping high-level service policies to actionable configurations and maintaining consistency across multiple administrative and technological domains. In this direction, intent-driven autonomous network and service management frameworks have been proposed, where intents are fed into hierarchical closed loops that use analytics and machine learning to adjust configurations and maintain SLA compliance [13]. More recent work on intent-based networking architectures discusses how to derive software frameworks and data models that support intent ingestion, validation and translation into lower-level policies [14]. Intent-driven orchestration has also been studied in the context of security slice management and ZSM-compliant closed-loop control [15].

Machine learning and, in particular, reinforcement learning have been extensively investigated for routing and resource allocation in optical networks. Machine-learning-based approaches have been considered for routing, wavelength and spectrum assignment, and virtual topology design, highlighting the potential of learning-based methods to approach ILP performance with significantly reduced runtime [5]. Specifically, RL has been applied to dynamic resource allocation problems in optical transport, including routing and spectrum assignment, spectrum defragmentation and lightpath reuse, often in flex-grid or multi-band scenarios [16]. RL has also been considered in routing in all-optical networks with physical impairments, demonstrating how learning can cope with QoT constraints in a dynamic setting [17].

Prior work has also explored learning latent context variables from interaction history and using them to condition the control policy. In meta-RL, PEARL infers probabilistic task context variables online from past transitions and uses them as an additional state input to enable fast adaptation and structured exploration [18]. Similarly, Bayes-adaptive deep RL methods such as VariBAD learn latent belief representations over unknown task parameters via a variational encoder and condition decisions on this inferred latent state [19]. In contrast to these settings, our latent vector captures per-spoke service intent and directly shapes P2MP hub and subcarrier allocation decisions under optical feasibility constraints. While these works address dynamic control and QoT-awareness, they typically optimise a single operator-defined objective (e.g., blocking probability, spectrum utilisation, or path length) and treat all demands as homogeneous. User-specific or service-specific preferences are either encoded implicitly through fixed reward weights or ignored entirely. Moreover, to the best of our knowledge, there is no prior work that specifically targets P2MP coherent aggregation with a latent intent model that influences hub assignment and subcarrier allocation decisions.

## III. System Model

### A. Metro-aggregation architecture

We consider a metro-aggregation architecture represented by a directed graph $G = (\mathcal{V}, \mathcal{E})$, as illustrated in Fig. 1. The node set is partitioned into access nodes $\mathcal{A}$, metro-aggregation nodes $\mathcal{M}$, and backbone nodes $\mathcal{F}$, with $\mathcal{V} = \mathcal{A} \cup \mathcal{M} \cup \mathcal{F}$. Access terminations are organized into horseshoes/rings. Each spoke $s \in \mathcal{S}$ is attached to an access node and routes bidirectional traffic towards the backbone. Metro nodes $\mathcal{M}$ provide candidate locations for P2MP hubs, while backbone nodes $\mathcal{F}$ represent upstream core PoPs. Network operation evolves over discrete epochs $t \in \{0, 1, 2, \ldots\}$, where spoke $s$ generates demand $L_{s,t}$ and the controller may update the P2MP light-tree configuration.

Let $\mathcal{D} \subseteq \mathcal{M}$ denote candidate hub nodes. Each hub $d \in \mathcal{D}$ has ports $\mathcal{P}_d$, where each port $p \in \mathcal{P}_d$ hosts at most one P2MP transceiver of type $\tau \in \mathcal{T}_h$ with $W_\tau$ subcarrier capacity and symbol rate $R_{\mathrm{SC}}$ (Gbaud).

Let $\mathcal{Q} = \{m_1, m_2, \ldots, m_{|\mathcal{Q}|}\}$ denote supported modulation formats ordered by increasing spectral efficiency. Each format $m \in \mathcal{Q}$ has spectral efficiency $b_m$ (bits/symbol) and required OSNR threshold $O_m$ (dB). The achievable bit rate per subcarrier is:

$$C_{\mathrm{SC}}(m) = b_m \cdot R_{\mathrm{SC}}. \tag{1}$$

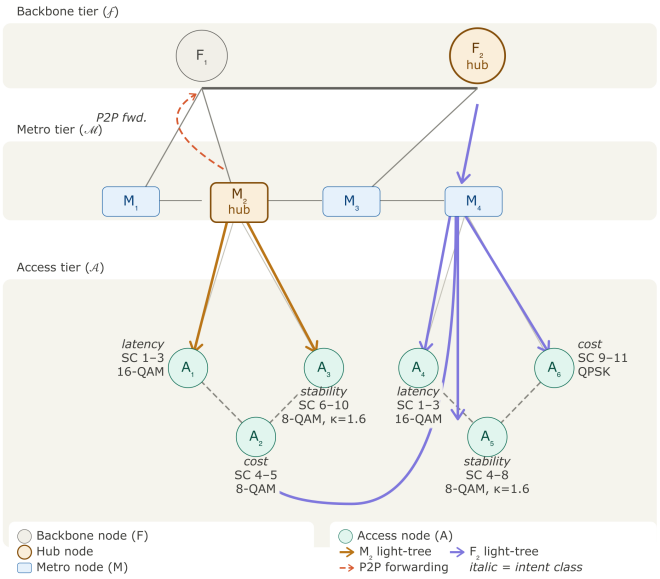


Fig. 1. Three-tier metro-aggregation architecture with P2MP hub placement at multiple levels.

Each spoke $s$ is attached to a horseshoe terminating at metro node $\hat{d}(s) \in \mathcal{M}$. The spoke experiences access OSNR $O_s^{\mathrm{HS}}$ at $\hat{d}(s)$, depending on position, fiber quality, and path length. Additional degradation along the metro segment yields $O_{\hat{d}(s),d}$. The end-to-end OSNR at hub $d$ is:

$$\mathrm{OSNR}_{s,d}^{-1} = \left(O_s^{\mathrm{HS}}\right)^{-1} + \left(O_{\hat{d}(s),d}\right)^{-1}. \tag{2}$$

The OSNR-feasible modulation formats for spoke-hub pair $(s,d)$ are:

$$\mathcal{Q}_{s,d}^{\mathrm{feas}} = \{m \in \mathcal{Q} : \mathrm{OSNR}_{s,d} \geq O_m\}. \tag{3}$$

We select the highest feasible modulation format:

$$m_{s,d}^{\star} = \arg \max_{m \in \mathcal{Q}_{s,d}^{\mathrm{feas}}} b_m. \tag{4}$$

If $\mathcal{Q}_{s,d}^{\mathrm{feas}} = \emptyset$, hub $d$ cannot serve spoke $s$. The *OSNR-feasible hub set* is:

$$\mathcal{D}_s^{\mathrm{feas}} = \left\{d \in \mathcal{D} : \mathcal{Q}_{s,d}^{\mathrm{feas}} \neq \emptyset\right\}. \tag{5}$$

Serving demand $L_{s,t}$ at hub $d$ with modulation $m_{s,d}^{\star}$ requires at least:

$$n_{s,d,t}^{\min} = \left\lceil \frac{L_{s,t}}{b_{m_{s,d}^{\star}} \cdot R_{\mathrm{SC}}} \right\rceil \tag{6}$$

subcarriers. More distant hubs typically achieve lower OSNR, requiring lower modulation formats and higher subcarrier consumption. Total subcarriers on port $p$ at hub $d$ cannot exceed transceiver capacity:

$$\sum_{s \in \mathcal{S}_{d,p,t}} n_{s,d,p,t} \leq W_\tau, \qquad \forall d \in \mathcal{D},\ \forall p \in \mathcal{P}_d, \tag{7}$$

where $\mathcal{S}_{d,p,t}$ denotes spokes assigned to port $p$ of hub $d$ at epoch $t$.

### B. Latent Intent Vectors and Service Differentiation

Different spokes exhibit heterogeneous service priorities typically not explicitly declared to the network controller. We model these implicit preferences through a latent intent vector.

For each spoke $s$, we define $z_{s,t} \in \mathbb{R}^k$ at epoch $t$, inferred from interaction history $\mathcal{H}_{s,t}$ via a learned encoder:

$$z_{s,t} = g_\phi(\mathcal{H}_{s,t}), \tag{8}$$

where $g_\phi$ is a two-layer ReLU MLP (hidden dims 64 and 32, output dim $k = 8$) trained end-to-end with the rollout policy via Adam (lr=$10^{-3}$, batch size 32), and $\mathcal{H}_{s,t}$ contains the last $T_h = 24$ hours of normalized demands, assigned hubs, latencies, and reconfiguration events.

Intent influences decisions through the following two mechanisms. First, to accommodate demand fluctuations and QoT variations without frequent reconfiguration, the controller allocates additional subcarriers beyond the strict minimum:

$$\kappa_{s,t} = \kappa(z_{s,t}), \qquad \kappa_{s,t} \geq 1, \tag{9}$$

where $\kappa(\cdot) : \mathbb{R}^k \to [1, \infty)$ maps intent to an overprovision factor. Stability-oriented spokes receive larger $\kappa_{s,t}$. The actual allocation becomes:

$$n_{s,d,p,t} = \left\lceil \kappa_{s,t} \cdot n_{s,d,t}^{\min} \right\rceil. \tag{10}$$

Secondly, among OSNR-feasible hubs $d \in \mathcal{D}_{s,t}^{\mathrm{feas}}$, we define an attribute vector:

$$\psi_{s,d,t} = [\ell_{s,d},\ u_{d,t},\ c_{s,d,t},\ h_d], \tag{11}$$

where $\ell_{s,d}$ is the physical latency (propagation delay or hop count) from spoke $s$ to hub $d$, $u_{d,t}$ is the current utilization of hub $d$, $c_{s,d,t}$ indicates whether assigning to hub $d$ would require reconfiguration (binary: 1 if $d \neq d_{s,t-1}$, 0 otherwise), and $h_d$ indicates whether hub $d$ has direct backbone connectivity (binary variable equal to 1 if $d \in \mathcal{F}$, 0 otherwise).

Intent determines preference weights $w(z_{s,t}) \in \mathbb{R}^4$, yielding a hub preference score:

$$\Phi_{s,d,t}(z_{s,t}) = \langle w(z_{s,t}), \psi_{s,d,t} \rangle. \tag{12}$$

This model enables spoke-specific trade-offs: latency-sensitive spokes (e.g., edge computing traffic) prefer nearby hubs with high weight on $\mathrm{lat}_{s,d}$; cost-sensitive spokes (e.g., backbone-bound traffic) prefer core hubs with direct connectivity, placing high weight on $\mathrm{core}_d$ to avoid expensive P2P reach-extension ports; stability-oriented spokes (e.g., carrier-grade services) prefer to remain at their current hub, placing high weight on $\mathrm{churn}_{s,d,t}$, and additionally receive larger subcarrier headroom $\kappa_{s,t}$ to buffer against demand fluctuations.

Latency-sensitive spokes weight $\ell_{s,d}$ heavily; cost-sensitive spokes weight $h_d$ to prefer backbone-connected hubs; stability-oriented spokes weight $c_{s,d,t}$ and receive larger $\kappa_{s,t}$ to buffer fluctuations.

## IV. MDP Formulation

We formulate the dynamic P2MP light-tree configuration problem as an infinite-horizon discounted Markov decision process (MDP). The MDP is characterized by state space $\mathcal{X}$, action space $\mathcal{A}$, transition kernel $\mathsf{P}$, single-stage cost function $g$, and discount factor $\gamma \in (0, 1)$.

### A. State Space

The system state at epoch $t$ captures current demands, inferred service intents, and the previous assignment configuration:

$$X_t = \left(\{L_{s,t}\}_{s\in\mathcal{S}}, \{z_{s,t}\}_{s\in\mathcal{S}}, A_{t-1}\right), \tag{13}$$

where $L_{s,t}$ is the demand of spoke $s$, $z_{s,t} \in \mathbb{R}^k$ is the latent intent vector inferred via the encoder $g_\phi(\mathcal{H}_{s,t})$ from interaction history, and $A_{t-1}$ records the previous assignment (required for computing reconfiguration penalties). The spoke-specific access OSNR values $\{O_s^{\mathrm{HS}}\}_{s\in\mathcal{S}}$ and metro-layer OSNR values $\{O_{\hat{d}(s),d}\}_{s,d}$ are precomputed from the network topology and treated as static parameters rather than state variables.

### B. Action Space

An action specifies a joint assignment configuration for all spokes:

$$A_t = \{(d_s, p_s, n_s)\}_{s\in\mathcal{S}}, \tag{14}$$

where spoke $s$ is assigned to hub $d_s \in \mathcal{D}_s^{\mathrm{feas}}$, port $p_s \in \mathcal{P}_{d_s}$, and allocated $n_s$ subcarriers. The assignment must satisfy port capacity constraints (7). The modulation format $m^\star_{s,d_s}$ is deterministically selected according to Eq. (4) based on the end-to-end OSNR from spoke $s$ to hub $d_s$, and is therefore not an explicit decision variable. The action space size grows combinatorially as $O(|\mathcal{D}|^{|\mathcal{S}|})$, making exact policy optimization intractable for realistic network sizes.

### C. Cost Function

The single-stage cost function balances multiple objectives: spectral efficiency, infrastructure costs, reconfiguration overhead, and intent satisfaction:

$$\begin{aligned} g(X_t, A_t) = &\sum_{d,p} c_{\mathrm{port}} \cdot \mathbb{I}\{\text{port } (d,p) \text{ active}\} \\ &+ \sum_s c_{\mathrm{SC}} \cdot n_s + c_{\mathrm{churn}} \cdot \mathbb{I}\{d_s \neq d_{s,t-1}\} \\ &+ \sum_{d\notin\mathcal{F}} c_{\mathrm{P2P}} \cdot \mathbb{I}\{\text{hub } d \text{ requires P2P forwarding}\} \\ &- \lambda \sum_s \Phi_{s,d_s,t}(z_{s,t}), \end{aligned} \tag{15}$$

where $c_{\mathrm{port}}$ penalizes port activation (transceiver cost), $c_{\mathrm{SC}}$ penalizes subcarrier usage (spectral cost), $c_{\mathrm{churn}}$ penalizes reconfiguration events (operational cost), $c_{\mathrm{P2P}}$ penalizes P2P reach-extension ports required when traffic aggregates at non-core hubs $d \notin \mathcal{F}$, and $\lambda > 0$ weights intent satisfaction via the preference score $\Phi_{s,d_s,t}(z_{s,t})$ defined in Eq. (12).

### D. Transition Dynamics

Demands evolve stochastically according to $L_{s,t+1} \sim \mathsf{P}_L(\cdot \mid L_{s,t})$ for all $s \in \mathcal{S}$. Intent vectors are updated by the encoder as defined in Eq. (8): $z_{s,t+1} = g_\phi(\mathcal{H}_{s,t+1})$, where the history $\mathcal{H}_{s,t+1}$ includes the current state-action pair $(X_t, A_t)$. The overall state transition is thus stochastic, inheriting randomness from the demand process.

### E. Objective

The control objective is to find a stationary policy $\mu : \mathcal{X} \to \mathcal{A}$ that minimizes the expected infinite-horizon discounted cost:

$$J_\mu(x) = \mathbb{E}^\mu\left[\sum_{t=0}^{\infty} \gamma^t g(X_t, \mu(X_t)) \,\middle|\, X_0 = x\right], \tag{16}$$

where the expectation is taken over trajectories generated by policy $\mu$ and the stochastic demand process.

## V. Approximate Dynamic Programming Solution

The combinatorial action space and coupling constraints make exact dynamic programming intractable. We develop an approximate solution based on rollout policy improvement that leverages a base heuristic [20].

### A. Rollout Policy Improvement

We construct a policy $\mu_{\mathrm{ro}}$ by performing a one-step lookahead with respect to a base heuristic policy $\mu_0$. At state $x$, the rollout policy selects:

$$\mu_{\mathrm{ro}}(x) \in \arg\min_{A\in\mathcal{A}(x)} \left\{g(x, A) + \gamma\, \mathbb{E}\left[J_{\mu_0}(X') \mid x, A\right]\right\}, \tag{17}$$

where $J_{\mu_0}(X')$ is the cost-to-go from next state $X'$ under base policy $\mu_0$. The rollout property guarantees that $J_{\mu_{\mathrm{ro}}}(x) \leq J_{\mu_0}(x)$ for all $x$, i.e., the rollout policy is at least as good as the base policy. The key challenge is evaluating the minimization in (17) over the combinatorial action space $\mathcal{A}(x)$. We address this through a two-level approach. Firstly, a fast intent-aware greedy heuristic serves as the base policy $\mu_0$ and constructs candidate actions, and secondly the rollout evaluates a restricted set of high-quality candidates to select the best action.

The base heuristic $\mu_0$ constructs feasible spoke-to-hub assignments through a four-stage greedy procedure that exploits the learned intent structure. For each spoke $s$, we first form a ranked list of OSNR-feasible hub-port pairs:

$$\mathcal{C}_s(x) = \left\{(d,p) : d \in \mathcal{D}_s^{\mathrm{feas}}, p \in \mathcal{P}_d\right\}, \tag{18}$$

sorted by decreasing preference score $\Phi_{s,d,t}(z_{s,t})$ from Eq. (12), prioritizing hubs aligned with each spoke's inferred intent. Spokes are sorted by decreasing best-case subcarrier demand (first-fit decreasing):

$$\text{priority}(s) = \min_{d\in\mathcal{D}_s^{\mathrm{feas}}} n^{\min}_{s,d,t}. \tag{19}$$

Next, processing spokes in priority order, we assign spoke $s$ to its highest-preference candidate $(d,p) \in \mathcal{C}_s(x)$ satisfying:

$$n_{s,d,t} = \lceil \kappa_{s,t} \cdot n^{\min}_{s,d,t} \rceil \leq \mathrm{rem}_{d,p,t}, \tag{20}$$

**Algorithm 1** Sequential Rollout

**Input:** State $x_t$, encoder $g_\phi$, base heuristic $\mu_0$, discount $\gamma$, Spoke order $\pi$, candidate limit $R$, rollout evaluator $\widehat{J}_{\mu_0}(\cdot)$
**Output:** Assignment action $A_t$

Infer intents: $z_{s,t} \leftarrow g_\phi(\mathcal{H}_{s,t})\ \forall s \in \mathcal{S}$
$A \leftarrow \emptyset$
**for** $i = 1$ to $|\mathcal{S}|$ **do**
  $s \leftarrow \pi(i)$
  $\mathcal{U}_s \leftarrow$ feasible hub–port choices for $s$ given $A$ (keep top-$R$ by $\Phi_{s,d,t}(z_{s,t})$)
  **for all** $u \in \mathcal{U}_s$ **do**
    $\tilde{A}(u) \leftarrow \mu_0(x_t;\ \text{fixed } A \cup \{s \to u\})$
    $Q(u) \leftarrow g(x_t, \tilde{A}(u)) + \gamma\, \widehat{J}_{\mu_0}(x_t, \tilde{A}(u))$
  **end for**
  $u^\star \leftarrow \arg\min_{u \in \mathcal{U}_s} Q(u)$
  $A \leftarrow A \cup \{s \to u^\star\}$
**end for**
**return** $A_t \leftarrow A$

where $\mathrm{rem}_{d,p,t}$ is the remaining capacity on port $(d, p)$. The allocated subcarriers include a margin factor $\kappa_{s,t} \geq 1$ from Eq. (9) that buffers stability-oriented spokes against demand fluctuations. Finally, we perform bounded pairwise swaps that reduce the cost function (15) while preserving capacity constraints. The complete heuristic runs in $O(|\mathcal{S}| \log |\mathcal{S}| + |\mathcal{S}|C)$ time, where $C = \max_s |\mathcal{C}_s|$.

### B. Rollout Algorithm

We adopt a sequential rollout mechanism (Algorithm 1) that constructs the assignment spoke-by-spoke, using one-step lookahead with base-heuristic tail completion at each decision point. At decision epoch $t$, we first infer intent vectors $z_{s,t}$ for all spokes and compute overprovisioning factors $\kappa_{s,t} = \kappa(z_{s,t})$. We process spokes in first-fit decreasing order $\pi$ based on $\min_{d \in \mathcal{D}_s^{\text{feas}}} n_{s,d,t}^{\min}$ to reduce fragmentation. For spoke $s = \pi(i)$, we form the set $\mathcal{U}_s$ of feasible hub-port candidates satisfying remaining capacity constraints. For each candidate $u \in \mathcal{U}_s$, we fix the assignment $s \to u$ and invoke the base heuristic $\mu_0$ to complete assignments for all remaining spokes, yielding a full assignment $\tilde{A}(u)$. We evaluate each candidate via one-step lookahead:

$$Q(u) = g(x_t, \tilde{A}(u)) + \gamma\, \mathbb{E}[J_{\mu_0}(X_{t+1}) \mid x_t, \tilde{A}(u)], \qquad (21)$$

where the expectation over the next state $X_{t+1}$ is approximated by forward simulation under policy $\mu_0$. The candidate minimizing $Q(u)$ is selected and fixed, and the procedure advances to the next spoke.

## VI. Numerical Experiments

### A. Experimental Setup

We evaluate the proposed intent-aware ADP controller on a realistic metro-aggregation topology with $|\mathcal{M}| = 30$ candidate aggregation hubs and $|\mathcal{S}| = 100$ access spokes organized into 10 horseshoe segments. Spoke-specific access OSNR values $O_s^{\text{HS}}$ range from 17.5 to 23 dB, reflecting heterogeneous fiber quality and position-dependent path lengths within each ring. The modulation format set is $\mathcal{Q} = \{\text{QPSK}, \text{8-QAM}, \text{16-QAM}, \text{32-QAM}, \text{64-QAM}\}$ with spectral efficiencies $\{2, 3, 4, 5, 6\}$ bits/symbol and OSNR thresholds $\{5.5, 8.5, 12.5, 15.1\}$ dB. For each spoke-hub pair, the controller selects the highest feasible modulation format (Eq. (4)).

Traffic demands follow a stochastic process with mean load 40 Gbps per spoke and periodic fluctuations. Spokes are partitioned into three latent intent profiles: latency-sensitive (edge-compute traffic, prefer nearby hubs to minimize delay), cost-sensitive (backbone-bound traffic, prefer core hubs to avoid P2P ports), and stability-oriented (carrier-grade, prioritize assignment consistency). The encoder $g_\phi$ infers these profiles online and translates them into preference weights $w(z_{s,t})$ and overprovisioning factors $\kappa(z_{s,t})$ via two-layer ReLU MLPs. We benchmark against: (i) Greedy-Latency (nearest feasible hub), (ii) Greedy-Packing (first-fit decreasing, resource efficiency), and (iii) Perfect-Intent. Training completes in ≈8 min on an Intel Core i7-1260P (16 GB RAM); inference runs under 300 ms per epoch.

### B. Performance Analysis

Fig. 2 shows the cost-satisfaction trade-off. Greedy-Latency achieves high satisfaction (0.82) but at high cost (67.10), while Greedy-Packing reduces cost (42.94) at the expense of satisfaction (0.45). Intent-RL attains satisfaction 0.79 at cost 53.67, an 11% reduction over Greedy-Latency while remaining close to Perfect-Intent. Per-intent performance (Fig. 3) reveals that Greedy-Packing exhibits severe degradation for stability-oriented spokes (satisfaction 0.15) due to aggressive subcarrier reshuffling, while Intent-RL maintains balanced satisfaction (≈0.79) across all three intent classes. The cost reduction stems primarily from intelligent P2P port management. Intent-RL reduces backbone reach-extension ports from 24.0 (Greedy-Latency) to 15.11, a 37% improvement (Fig. 4), by routing cost-sensitive spokes directly to core hubs. This consolidation increases spectral utilization to 0.492, 13% higher than Greedy-Latency. Finally, Intent-RL achieves churn rate 0.218 versus 0.855 for Greedy-Packing, a 75% reduction enabled by intent-driven margin allocation $\kappa(z_{s,t})$ that buffers demand fluctuations.

## VII. Conclusion

This paper presented an intent-aware reinforcement learning mechanism for autonomous P2MP light-tree configuration in metro-aggregation networks. The approach integrates a latent encoder within an ADP architecture to infer implicit service preferences and translate them into hub selection and subcarrier allocation decisions under physical-layer constraints. The sequential rollout mechanism provides a computationally tractable solution to the combinatorial assignment problem while preserving the cost improvement guarantees of policy rollout. Numerical evaluation on a metro topology with 30 candidate hubs and 100 access spokes demonstrates that Intent-RL reduces P2P reach-extension ports by 37% over Greedy-Latency, cuts reconfiguration churn by 75% over Greedy-Packing, and maintains balanced intent satisfaction (≈0.79) across all classes, confirming that latent intent inference is a viable alternative to explicit operator-specified objectives.

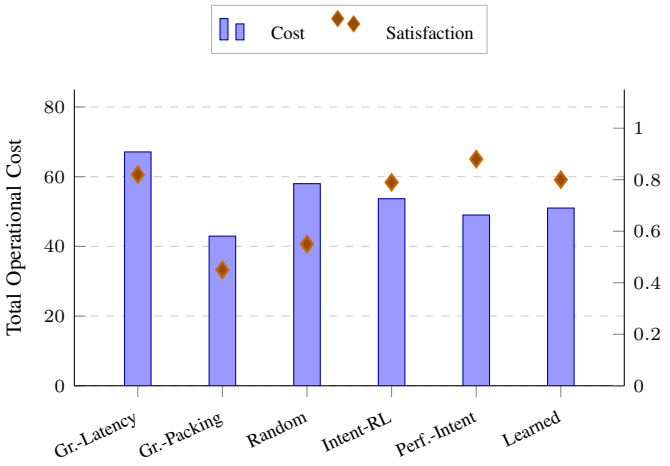


Fig. 2. Total operational cost (bars) and average intent satisfaction score (diamonds) for Greedy-Latency, Greedy-Packing, Random, Intent-RL, Perfect-Intent, and Learned. Lower cost and higher satisfaction are preferable.

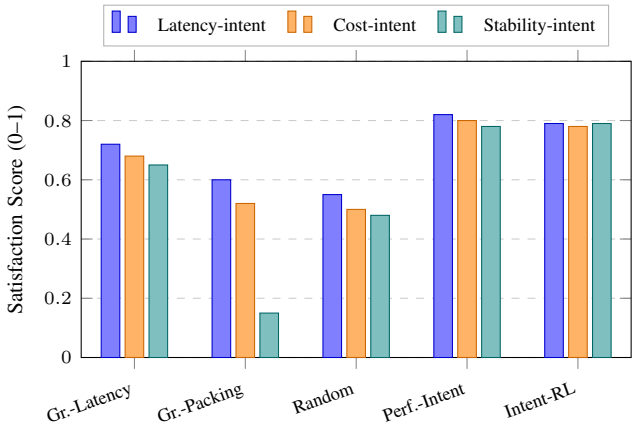


Fig. 3. Per-intent satisfaction for Greedy-Latency, Greedy-Packing, Random, Perfect-Intent, and Intent-RL across latency-sensitive, cost-sensitive, and stability-oriented spoke classes. Intent-RL achieves balanced satisfaction (≈0.79) across all three classes, whereas Greedy-Packing collapses for stability-oriented spokes (0.15).

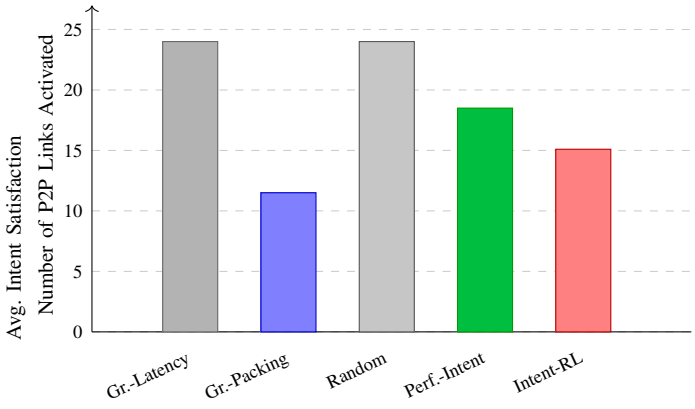


Fig. 4. P2P reach-extension port count for Greedy-Latency, Greedy-Packing, Random, Perfect-Intent, and Intent-RL. Fewer ports indicate better cost-sensitive hub selection. Intent-RL (15.1) reduces P2P port usage by 37% compared to Greedy-Latency (24.0).

## References


[1] L. S. de Sousa and A. C. Drummond, "Metropolitan optical networks: A survey on single-layer architectures," *Opt. Switch. Netw.*, vol. 47, no. C, Feb. 2023.

[2] P. Soumplis *et al.*, "Enabling extended access aggregation with light-trees and point-to-multipoint coherent transceivers," *Journal of Optical Communications and Networking*, vol. 17, no. 4, p. 249, 2025.

[3] S. Din *et al.*, "Dynamic subcarrier allocation in point-to-multipoint coherent metro-aggregation networks," *Optical Fiber Technology*, vol. 84, p. 103768, 2024.

[4] ETSI, "Zero-touch network and service management (zsm); intent-driven autonomous networks; generic aspects," European Telecommunications Standards Institute (ETSI), ETSI GR ZSM 011, 2024.

[5] Y. Zhang *et al.*, "Overview on routing and resource allocation based on machine learning in optical networks," *Optical Fiber Technology*, vol. 60, p. 102355, 2020.

[6] P. Pavon-Marino *et al.*, "Benefits of point-to-multipoint coherent pluggables in metro-aggregation networks," *Journal of Optical Communications and Networking*, vol. 14, no. 5, pp. B30–B39, 2022.

[7] N. Skorin-Kapov *et al.*, "Re-thinking optical transport for 5g optical metro networks: A case study," in *International Conference on Optical Network Design and Modeling (ONDM)*, 2021.

[8] S. Hosseini *et al.*, "Multi-period planning in metro-aggregation networks exploiting point-to-multipoint coherent transceivers," *Journal of Optical Communications and Networking*, vol. 15, no. 3, pp. 155–164, 2023.

[9] P. Soumplis *et al.*, "Qos-aware resource allocation in point to multipoint (p2mp) metro aggregation networks," in *Optical Fiber Communication Conference (OFC) 2025, Technical Digest Series*. Optica Publishing Group, 2025, paper Th3D.2.

[10] P. Soumplis, K. Christodoulopoulos, A. Napoli, K. Yiannopoulos, and E. Varvarigos, "Dynamic risk-aware reconfiguration in coherent p2mp extended access networks under time-varying demands," in *European Conference on Optical Communication (ECOC) 2025*, Copenhagen, Denmark, 2025.

[11] E. Li *et al.*, "Invited paper: P2mp coherent transceivers for wavelength-switched optical networks with hub-and-spoke traffic," Conference invited paper (camera-ready PDF), 2023.

[12] M. Liyanage *et al.*, "A survey on zero touch network and service management (zsm) for 5g and beyond networks," *Journal of Network and Computer Applications*, p. 103362, 2022.

[13] K. Mehmood, K. Kralevska, and D. Palma, "Intent-driven autonomous network and service management in future cellular networks: A structured literature review," 2021.

[14] Yu *et al.*, "Building a comprehensive intent-based networking framework: A practical approach from design concepts to implementation," Preprint, 2024.

[15] J. Asensio *et al.*, "Zsm framework for autonomous security service level agreement life-cycle management in b5g networks," *Future Internet*, vol. 17, no. 2, p. 86, 2025.

[16] J. Suárez-Varela *et al.*, "Routing in optical transport networks with deep reinforcement learning," *Journal of Optical Communications and Networking*, vol. 11, no. 11, pp. 547–558, 2019.

[17] Y. Pointurier and K. Heidari, "Reinforcement learning based routing in all-optical networks with physical impairments," in *Fourth International Conference on Broadband Communications, Networks and Systems (BROADNETS 2007)*, 2007.

[18] K. Rakelly, A. Zhou, D. Quillen, C. Finn, and S. Levine, "Efficient off-policy meta-reinforcement learning via probabilistic context variables," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 97, 2019, pp. 5331–5340.

[19] L. Zintgraf, K. Shiarlis, M. Igl, S. Schulze, Y. Gal, K. Hofmann, and S. Whiteson, "Varibad: A very good method for bayes-adaptive deep rl via meta-learning," in *International Conference on Learning Representations (ICLR)*, 2020.

[20] D. P. Bertsekas, *Reinforcement Learning and Optimal Control*. Belmont, MA: Athena Scientific, 2019.