Defending QKD Networks: Routing and Wavelength Assignment to Mitigate Physical-Layer Attacks

Mengyao Li, Qiaolun Zhang, Zheng Zhang, Zongshuai Yang, Stefano Bregni, Alberto Gatto, Vicente Martin, Raouf Boutaba, *Fellow, IEEE* and Massimo Tornatore, *Fellow, IEEE*

***Abstract*—Quantum key distribution (QKD) supports secret key exchange to enable data exchange with guaranteed security but remains vulnerable to *key exchange interruption* caused by physical-layer threats, such as high-power jamming attacks. In particular, in a fiber-based QKD network equipped with optical switching capabilities, a high-power jamming signal injected into *a single link* can be optically switched and propagated to other links, resulting in the *interruption of key exchange on multiple links*. To address this challenge, we introduce a novel metric, the *maximum number of affected requests* (maxNAR), which quantifies the worst-case impact of a single physical-layer attack. Based on this, we investigate a new problem: Routing and Wavelength Assignment with Minimal Attack Radius (RWA-MAR). The objective of RWA-MAR is to assign routes and wavelengths to QKD requests in such a way that the maximum number of requests disrupted by any single physical-layer attack is minimized. We first formulate the problem using an integer linear programming (ILP) model to minimize maxNAR. Due to computational limitations of ILP, we develop a scalable deep reinforcement learning (DRL) solution that dynamically balances security (minimizing maxNAR) and resource efficiency. The DRL solution is designed to utilize key caches (defined as QKP) to minimize the maxNAR, as the QKP can save the keys for future use. Extensive simulations across diverse topologies and workloads show that our approach achieves a significant reduction in maxNAR (over 50%) compared to a baseline that performs RWA without considering the reduction of an attack impact. Meanwhile, our approach maintains reasonable resource efficiency in key delivery, marking a step toward robust and attack-aware QKD network design.**

***Index Terms*—Integer linear programming, optical bypass, quantum key distribution, quantum key pool, trusted relay**

I. Introduction

QUANTUM key distribution (QKD) has emerged as a promising technology to support secure key exchange that is resistant to quantum attacks [1], [2]. Through the application of quantum mechanics, QKD guarantees information-theoretic security, making it resilient against adversaries equipped with quantum capabilities [1], [3]. Initially, QKD networks are limited to point-to-point configurations; recent innovations have paved the way for scalable multi-point QKD networks built on optical infrastructures [4], [5]. These scalable QKD networks, consisting of QKD nodes interconnected with quantum links, allow for the secure exchange of cryptographic keys between parties. Each quantum node is equipped with multiple QKD modules, which can be either a transmitter or a receiver. A QKD path, established between two end nodes, requires one quantum module at each end node to generate secret keys. Over the years, advancements in QKD protocols, devices, and systems have substantially improved performance metrics such as secret-key rate, transmission distance, and security [4].

Although QKD can exchange cryptographic keys with guaranteed information-theoretic security, it remains vulnerable to physical-layer attacks. Various threat vectors can disrupt key distribution, including out-of-band attacks that interfere with classical control channels or inject unauthorized signals to compromise quantum channels [6]–[8]. Specifically, institutions such as the National Institute of Standards and Technology (NIST) and the National Security Agency (NSA) have highlighted potential vulnerabilities about the physical-layer attacks on QKD in real-world scenarios [9], [10]. While classical optical networks have been extensively studied in terms of resilience to physical-layer attacks, most QKD-related research focuses on performance optimization through resource provisioning and routing strategies [11], [12], with comparatively little attention to physical layer attacks in QKD.

Threats such as out-of-band attacks, where adversaries interfere with classical control channels or inject jamming signals into channels, pose significant risks to the integrity of key distribution [7], [8]. An attacker may exploit high-power jamming attacks to disrupt a quantum link, potentially impacting not only the intended transmission but also other requests passing through the same fiber. If the attacker is constrained to compromise only a single location, the most strategic attack choice would be to target the fiber link that affects the largest number of concurrent requests, thereby maximizing disruption and significantly increasing the damage inflicted on the QKD network.

In this work, we formulate the routing and wavelength assignment with minimal attack radius (RWA-MAR) problem to investigate the potential impact of jamming attacks on QKD networks and propose strategies to mitigate these vulnerabilities. The concept of attack radius is defined as the set of requests that share at least one common physical link with

This work was supported in part by the Innovation for Defence Excellence and Security (IDEaS) program from the Department of National Defence (DND). This work was also partially supported by the project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU. (Corresponding author: Qiaolun Zhang.)

Mengyao li, Qiaolun Zhang, Zheng Zhang, Zongshuai Yang, Stefano Bregni, Alberto Gatto, and Massimo Tornatore are with Politecnico di Milano, Italy (e-mail: mengyao.li@polimi.it, qiaolun.zhang@polimi.it, zheng.zhang@polimi.it, zongshuai.yang@mail.polimi.it, stefano.bregni@polimi.it, alberto.gatto@polimi.it, massimo.tornatore@polimi.it)

Vicente Martin is with the Center for Computational Simulation and DLSIIS, ETSI Informaticos, Universidad Politécnica de Madrid, Madrid, Spain. (e-mail:vicente@fi.upm.es)

Raouf Boutaba is with the University of Waterloo, Canada (e-mail:rboutaba@uwaterloo.ca)

This article has been accepted for publication in IEEE Transactions on Quantum Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TQE.2026.3684081

a given lightpath, making them simultaneously vulnerable to a single attack. Our objective is to ensure that all requests are successfully served while minimizing the attack radius, thereby reducing the worst-case impact of a compromised link. We focus on QKD networks incorporating three critical technologies: (i) trusted relays (TRs), enabling key forwarding through intermediate trusted nodes using one-time pad encryption schemes [3]; TRs require additional quantum modules at each intermediate node and involve signal regeneration; (ii) optical bypasses (OBs), allowing direct key delivery between non-adjacent nodes using Reconfigurable Optical Add/Drop Multiplexers (ROADMs) [11]; OB only consumes two modules per connection, whereas the signal bypasses intermediate nodes entirely, enabling propagation without regeneration; (iii) quantum key pools (QKPs), acting as key caches at each node, storing pre-distributed keys for future use. These keys are generated in advance through quantum channels using either OBs or TRs or OBTR combination mechanisms and then stored in the QKP. Once generated and cached, their subsequent usage does not rely on real-time quantum transmission over optical links. Therefore, provided that the key generation process is not disrupted, keys stored in QKPs remain unaffected by later physical-layer attacks on fiber links, thereby increasing routing flexibility and reducing vulnerability to such attacks.

While both OBs and TRs enable efficient key distribution over non-adjacent nodes, their impact on the attack radius in QKD networks is different. Note that, in this work, we refer to OBs and TRs as logical capabilities rather than strictly physical components required to achieve such capabilities. OBs enable signals to traverse multiple nodes without regeneration, conserving quantum modules, which are valuable and limited resources at each node. However, this comes with a security trade-off: a physical-layer attack on any link within an OB path can propagate the signal, potentially disrupting multiple requests simultaneously and thus increasing the attack radius. In contrast, TRs regenerate the quantum signal at intermediate nodes, effectively limiting the spread of an attack to individual link segments, thereby reducing the maxNAR. However, this improved security comes at the cost of additional modules at each relay node. This trade-off between minimizing module usage (favoring OBs) and reducing the network's vulnerability to large-scale disruptions (favoring TR) is central to the RWA-MAR problem in QKD. Unlike classical optical networks [13], QKD networks must jointly consider physical-layer attack containment alongside resource allocation, factoring in the interplay between key caching, trusted relaying, and bypass routing. These dynamics make the problem both more complex and more security-critical.

To study the impact of jamming attacks in QKD networks, we use the number of affected requests (NAR) as a metric to measure the vulnerability of routing solutions. NAR refers to the number of requests that share at least one physical link and can therefore be disrupted by a single physical-layer attack. This idea is mutuated from the lightpath attack radius (LAR), which was introduced to study physical-layer attack mitigation in optical networks [6], [14]. Based on this idea, we define the maximum number of affected requests (maxNAR). The maxNAR value is the largest NAR among all physical links in the network. It reflects the worst-case disruption that can occur when a single link is attacked. In this work, maxNAR is used as the main metric to evaluate the vulnerability of QKD networks that employ OB and TR technologies.

An example of how OBs, TRs, and QKPs affect the maxNAR is illustrated in Fig. 1, which shows four key requests in the network. Request 1, between nodes (1,3), is served using QKD keys stored in QKPs (represented by a purple line). While for other requests, we assume that no pre-stored keys are available for these requests, so keys must be generated in real time. Request 2, between nodes (2,4), is served via TRs (orange lines). Requests 3 and 4, between nodes (1,3) and (2,4), respectively, are served via OB (green and blue lines). The figure also illustrates how a single attack on a physical link can affect multiple requests, including those routed over different paths. Specifically, a high-power jamming attack on link (1,2) disrupts Request 3. Since Request 3 utilizes optical bypass, the attack propagates to link (2,3) without signal regeneration at intermediate nodes, thereby disrupting Requests 2 and 4 as well. In contrast, Request 1 relies on pre-stored keys from the QKPs and therefore does not require real-time quantum transmission over the attacked links, remaining unaffected by the physical-layer attack. This example illustrates how optical bypass may increase the attack radius in QKD networks, highlighting the need for attack-aware routing strategies that minimize maxNAR.

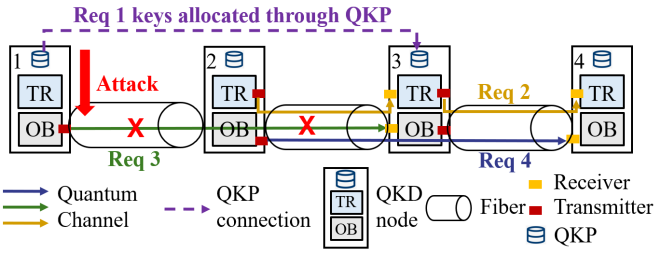


Fig. 1. Example of technologies specific to QKD networks and the impact of a physical attack on a single optical fiber link.

A core challenge in QKD network architecture is to ensure robust resilience to attacks as well as efficient utilization of limited resources. Minimizing the maxNAR generally promotes more distributed routing strategies, whereas efficient use of QKPs often favors more centralized routing, leading to higher resource concentration. Trusted relays can control attack impact, although they are dependent on more QKD modules, while optical bypass is concerned with resource conservation, along with increased attack impact. Moreover, our system also captures the time-dependent characteristics of QKD systems by incorporating the concept of timeslots, which corresponds to the time for QKP storage becomes available.

This article extends our previous conference paper published in Globecom [15]. In the conference version, we first introduced the routing and wavelength assignment with minimal attack radius (RWA-MAR) problem for QKD networks and proposed the maxNAR metric to quantify the worst-case impact of physical-layer attacks. The conference paper formulated the problem using an integer linear programming (ILP) model and proposed a heuristic solution based on

Tabu Search to efficiently minimize maxNAR. However, the heuristic approach proposed in [15] has limitations in terms of scalability and adaptability when applied to larger network topologies or more complex traffic scenarios. To address these limitations, this journal version extends the previous work by introducing a deep reinforcement learning (DRL)-based framework for solving the RWA-MAR problem. The main enhancements are summarized as follows:

- We develop a DRL-based framework that enhances solution quality, scalability, and adaptability for attack-aware routing in QKD networks. Compared with the heuristic algorithm in [15], the approach provides more consistent performance. It also yields lower maxNAR and average NAR (avgNAR) than both the baseline solution and the heuristic method.
- We conduct extensive numerical evaluations across different QKD network architectures and topologies. Compared with the conference version in [15], this paper provides a more comprehensive evaluation across different topologies evaluations and a deeper investigation of the trade-off between attack mitigation and resource.

The remainder of the paper is organized as follows. Section II discusses the related work of this paper. Section III describes the ILP and system model. Section V introduces the proposed DRL algorithm. Section VI presents the obtained results. Finally, Section VII concludes the paper.

II. Related Work and Contribution

Large-scale QKD networks have progressed from laboratory demonstrations to real-world deployments, with operational testbeds reported in multiple regions, including Switzerland, Italy, and China [3], [16]. In parallel, experimental studies have confirmed that quantum key distribution can coexist with classical optical traffic on the same fiber infrastructure, enabling hybrid optical–quantum communication systems and improving deployment practicality. In such networks, key management functions are centrally coordinated through a control entity that governs QKD nodes and modules while exploiting advanced transmission mechanisms such as trusted relays and optical bypasses [17], [18]. Trusted relay technology, widely adopted in existing QKD infrastructures, plays a critical role in overcoming distance limitations by securely regenerating keys at intermediate nodes [3], [19]. A notable example is the work of Chen et al. [1], which demonstrated a QKD network spanning 4,600 km using a chain of trusted relays. To further enhance routing flexibility, hybrid architectures combining trusted and semi-trusted nodes have also been explored [20]. In addition, Yu et al. [21] investigated mixed trusted–untrusted relay scenarios, analyzing the influence of initial key storage in QKP and varying traffic conditions.

Beyond trusted relays, OB has emerged as an effective complementary technique for enhancing routing efficiency and resource utilization in QKD networks [11], [22]–[24]. By enabling direct quantum transmission between non-adjacent nodes without signal termination at intermediate nodes, OB can significantly reduce hardware complexity and operational overhead. Dong et al. [22] introduced a quantum node architecture that supports optical bypass and modeled its impact using auxiliary graphs to reflect multi-level node adjacency determined by physical distances. Sun et al. [23] provided experimental evidence for bypass-enabled designs, demonstrating notable improvements in signal-to-noise performance. From an optimization perspective, Zhang et al. [11] investigated global routing strategies in QKD networks that jointly consider trusted relays and optical bypass, while Yu et al. [24] addressed integrated routing, wavelength, and timeslot allocation in short-reach QKD optical networks equipped with bypass capabilities.

TABLE I
Comparison of Related Works and This Paper

Work	Routing	Attack	QKP/OB/TR	Learning-based solution
[22]	✓	×	OB/TR	×
[11]	✓	×	QKP/OB/TR	×
[12]	✓	×	QKP/OB/TR	×
[24]	✓	×	OB/TR	×
[25]	✓	×	TR	DRL
[27]	✓	✓	–	DRL
This Work	✓	✓	QKP/OB/TR	DRL

As QKD networks continue to mature, there has been a growing number of works to investigate DRL-enabled adaptive decision-making in quantum communications. Sharma et al. [25] investigated employing DRL in improving routing efficiency and resource utilization in QKD lightpath establishment. Reiß et al. [26] analyzed learning-based optimization of secret key rates in long-distance quantum communication scenarios with quantum repeaters, showing the applicability of DRL to different regimes of transmission. Ding et al. [27] investigated the mitigation of quantum attacks in continuous-variable QKD optical networks to enhance the network resilience through dynamic reallocation of resources and adjustments of channel parameters. More recently, Seok et al. [28] proposed a DRL-driven framework enabling QKD networks to adaptively provision keys on a hop-by-hop basis against changing traffic demands and network conditions.

As QKD infrastructures continue to scale, protecting them against malicious disruptions has become an increasingly important concern. Much of the existing literature concentrates on security at the quantum device or protocol level [17], [29]–[31], whereas the robustness of QKD networks under physical-layer attacks has received comparatively less attention. Smith et al. [7] showed that even components commonly regarded as intrinsically secure—such as quantum random number generators—can be exposed to serious vulnerabilities through imperfections in their electronic implementations.

QKD networks are also vulnerable to physical-layer attacks, similar to classical optical networks, where such attacks can disrupt signal transmission and prevent key generation [6]. However, the impact of these attacks in QKD networks is fundamentally different due to the presence of QKPs, which have not been adequately considered in prior work. Unlike classical optical networks, where all services rely on real-time data transmission, QKPs enable keys to be generated in advance and stored for later use. This decouples key generation from key consumption, fundamentally altering the relationship between routing decisions and attack impact. As summarized

in Table I, most existing studies focus on routing optimization or learning-based management in QKD networks, without explicitly addressing network-scale attack mitigation under such QKD-specific mechanisms. In this work, we develop an attack-aware routing framework based on the maxNAR metric to capture the worst-case impact of physical-layer attacks. The proposed approach incorporates key caching through QKPs along with resource and transmission constraints, thereby improving the resilience of QKD networks against physical-layer threats.

III. Modeling and Statement of the Problem

In this section, we present the system model and define the RWA-MAR problem. First, we describe the QKD network model and the routing structure in the system. Then, we explain the key-rate model and the assumptions for QKP capacity. After that, we give the formal problem formulation and show how the maxNAR metric is calculated for different network architectures.

A. Network Model

We model a QKD network as a directional graph $G_p = (N_p, E_p)$, where N_p and E_p are the sets of nodes and links, respectively. We define maxNAR as the maximum number of requests that share at least one common physical link in the same direction. Link-sharing is a property that indicates whether two requests traverse at least one physical link in the same direction. We divide the time into discrete timeslots and construct a fully-connected auxiliary graph $G_p = (N_p, E_a)$ where each link denotes the opportunity for key distribution between adjacent and non-adjacent nodes. Key distribution between adjacent nodes can use a physical quantum channel or a logical auxiliary link enabled by QKP. A quantum channel is where qubits are transmitted on different wavelengths. Key distribution between non-adjacent nodes can use a quantum channel with OB/TR or an auxiliary link enabled by QKP key caching (i.e., stored keys in advance for future use). Each node in the network is equipped with a limited number of QKD modules, which include both transmitters and receivers. An OB connection requires a total of two modules, while a TR path consumes two modules (one QKD receiver and one QKD transmitter) at each intermediate node to enable key forwarding. Our evaluations are based on realistic QKD key rate models [11], incorporating both OB and TR mechanisms within the network architecture.

B. Achievable Key Rate and QKP Capacity

The achievable key rate model follows the approach in [11], which allows us to compute the maximum key rates for different transmission reaches, as summarized in Table II. When optical bypass is employed, the achievable key rate is reduced by approximately 11% for each traversed intermediate node due to accumulated optical losses. The required QKP capacity is estimated assuming AES-256 encryption in Cipher Block Chaining mode. A single AES key can securely encrypt up to 2^{48} AES blocks, corresponding to approximately 36,000 Tb of data. For a single-mode fiber carrying 100 channels at 1 Tb/s per channel, key rotation is therefore required every 360 seconds. Given a one-hour stage duration, the minimum QKP capacity needed per timeslot is 2560 bits [32].

TABLE II
Key rate for different reaches

Reaches	10km	20km	30km	40km	50km
Key rate	23 kb/s	13 kb/s	7 kb/s	3.5 kb/s	1.9 kb/s

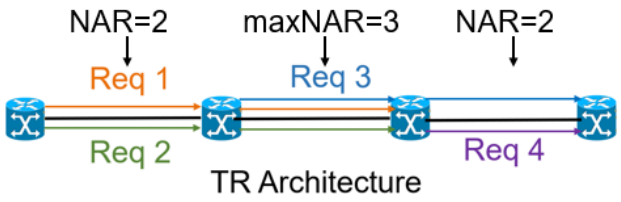
C. Problem Statement

The RWA-MAR problem is defined as follows. **Given** a QKD network topology, available QKD modules and quantum channels, a set of key requests, achievable key rates for different transmission reaches, and discrete timeslots, the goal is to **determine** the routing, wavelength assignment, and key-rate allocation for each request. The solution must satisfy constraints on path-dependent achievable key rates, as well as limits on quantum channels and QKD modules. The **objective** is to minimize the total maxNAR summed over all timeslots. We study the RWA-MAR problem under three network architectures, depending on the availability of optical bypass and trusted relay, following the definitions in Ref. [11]:

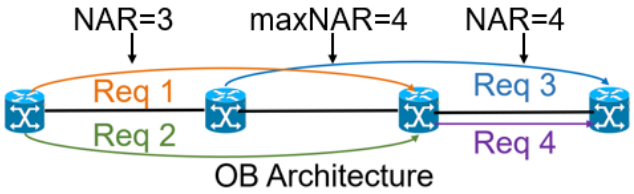
(1) *OBTR*: both trusted relays and optical bypasses permitted; (2) *OB*: only optical bypasses permitted; (3) *TR*: only trusted relays permitted. Note that the architecture supporting neither trusted relay nor optical bypass is not considered, as it cannot serve requests between non-adjacent nodes.

To illustrate the concept of maxNAR, Fig. 2 presents a simple example showing how the metric is determined under different routing architectures. We consider a 3 node network topology with 4 requests. In the figure, arrows indicate the location of a physical-layer attack and the resulting number of affected requests. The largest number of impacted requests among all attack positions is defined as maxNAR. In Fig. 2(a), all requests are routed through trusted relays. Under TR, the impact of an attack is confined to the requests directly traversing the compromised physical link. As a result, the NAR depends only on the number of requests sharing that link, yielding a relatively low maxNAR of 3 in this example. This improved resilience, however, is achieved at the expense of increased QKD module consumption. Note that quantum channels are directional; therefore, an attack in the opposite direction does not influence ongoing transmissions. Furthermore, since keys stored in QKPs are generated and cached in advance, they do not rely on real-time quantum transmission over optical links. As a result, they are not affected by physical-layer attacks on fiber links (as illustrated in Fig. 1). Therefore, scenarios where requests are served solely using cached QKP keys are not further considered in the attack impact analysis. In Fig. 2(b), Reqs 1, 2, and 3 are established using optical bypass, while Req 4 is served through a direct quantum link between adjacent nodes. Requests routed via OB are vulnerable to attack propagation, since optical signals traverse intermediate nodes without regeneration, allowing disturbances to spread across successive links. When an attack occurs on the central link, it directly disrupts Reqs 1, 2, and 3 that traverse this link. Because Req 1 is established

using optical bypass, the disturbance further propagates to the adjacent link without regeneration, thereby disrupting Req 4. Consequently, four requests are affected in total, yielding a maxNAR value of 4.



(a) NAR for TR, with maxNAR=3



(b) NAR for OB, with maxNAR=4

Fig. 2. Example of NAR calculation

IV. Integer Linear Programming Model

The sets, parameters, and decision variables used in the ILP formulation are summarized in Table III.

Objective function: minimize maxNAR

$$min \quad maxNAR = min \sum_{t \in T} maxNAR_t \tag{1}$$

1) Flow, link, and modules constraints: Let $S^-(i)$ and $S^+(i)$ denote the set of incoming and outgoing links from node i, respectively. Equations (2) and (3) represent the flow constraints for the QKD path, which can be either a quantum channel (OB/TR) or a QKP-enabled link. Equation (4) ensures that the number of modules used does not exceed the available modules at a given node. In a fully connected graph E_a, each link may correspond to several physical links $e \in E_p$ in the physical topology. Auxiliary links in the auxiliary graph represent the links utilizing OB, as these bypass the node. Equation (5) defines how these physical links combine to form a physical route $\phi \in \Phi$ for the auxiliary links $e \in E_a$. Equations (6) and (7) ensure that multiple QKD paths cannot share the same channel or route. This reflects the practical constraint that a quantum channel (i.e., a wavelength used for QKD transmission) cannot be simultaneously shared by multiple QKD paths on the same physical link due to the high sensitivity of quantum signals to interference.

$$\sum_{e \in S^+(i)} f_{e,w}^{p,t} - \sum_{e \in S^-(i)} f_{e,w}^{p,t} = \begin{cases} z_{p,w}^t \ if \ i = a(p) \\ -z_{p,w}^t \ if \ i = b(p) \\ 0 \ others \end{cases} \tag{2}$$
$$\forall p \in P, i \in N_p, t \in T, w \in W$$

TABLE III
Sets, Parameters, and Variables Description for ILP Model.

Sets	Description
G_p	Physical network topology
N_p	Set of physical-network nodes
E_p	Set of physical-network links
E_a	Set of links in the fully-connected graph
P	Set of node pairs of requests in the network
W	Set of QKD channels
T	Set of timeslots
Φ_e	Set of physical routes that use the same end nodes as auxiliary link $e \in E_a$
D	Set of requests
$S^+(i)$	Set of outgoing links from node i
$S^-(i)$	Set of incoming links for node i
Para	**Description**
O_n	Integer, number of QKD modules on node n
$h_{\phi,e}$	Binary, equals to 1 if link $e \in E_p$ in the route ϕ
k_d	Integer, required key rate of request $d \in D$
l_ϕ	Integer, key rate that can be supplied by route ϕ
Var	**Description**
$f_{e,w}^{p,t}$	Binary, equals to 1 if quantum channel w on link $e \in E_a$ is allocated for path between node pair p at timeslot t
$x_{e,w}^{p,t,\phi}$	Binary, equals to 1 if route ϕ is selected for connection between the end nodes of link $e \in E_a$ in QKD path between node pair p at channel w at timeslot t
$xx_{e'}^{p,t}$	Binary, at timeslot t, the routing of node pair p uses physical link e' in E_a
$p_{e,w}^{p,t}$	Binary, equals to 1 if link $e \in E_a$ used quantum channel in between node pair $p \in P$ on channel $w \in W$ at timeslot $t \in T$
$q_{e,w}^{p,t}$	Binary, equals to 1 if path p contains auxiliary link $e \in E_a$ based on QKP on channel w at timeslot t
$u_{p,w}^t$	Integer, key rate generated for path p on channel w at timeslot t
$z_{p,w}^t$	Binary, equals 1 if QKD path between node pair $p \in P$ uses quantum channel $w \in W$ at timeslot t
B_ϕ^t	Binary, at timeslot t, the routing ϕ has been used
$C_\phi^{p,t}$	Binary, at timeslot t, any attack on routing ϕ will affect the routing of node pair p
g_p^t	Integer, stored keys in QKP for path between node pair p at timeslot t
y_d^t	Binary equals to 1 if request d is served at timeslot t
$\gamma_{e,w}^{p,t}$	Integer, key rate provided from QKP for link e in QKD path between node pair p in channel w
$maxNAR_t$	Integer, the maxNAR at timeslot t

$$f_{e,w}^{p,t} = q_{e,w}^{p,t} \vee p_{e,w}^{p,t} \quad \forall e \in E_a, p \in P, t \in T, w \in W \tag{3}$$

$$\sum_{p \in P, e \in S^+(n), w \in W} p_{e,w}^{p,t} + \sum_{p \in P, e \in S^-(n), w \in W} p_{e,w}^{p,t} \leq O_n \quad \forall n \in N_p, t \in T \tag{4}$$

$$\sum_{\phi \in \Phi_e} x_{e,w}^{p,t,\phi} = q_{e,w}^{p,t} \quad \forall p \in P, e \in E_a, w \in W, t \in T \tag{5}$$

$$\sum_{p \in P} x_{e,w}^{p,t,\phi} \leq 1 \quad \forall e \in E_a, t \in T, w \in W, \phi \in \Phi_e \tag{6}$$

$$\sum_{p \in P, e' \in E_a, \phi \in \Phi_e} x_{e,w}^{p,t,\phi} * h_{\phi,e'} \leq 1 \forall e \in E_p, w \in W, t \in T \tag{7}$$

2) Key rate constraints: Eq. (8) ensures the key rate of QKP path p is less than the sum of the key rate provided by a quantum channel and from QKP. Eq. (9) ensures that keys are distributed only when the corresponding path p is active and available for use.

$$u_{p,w}^{t} \leq \sum_{\phi \in \Phi_e} (x_{e,w}^{p,t,\phi} * l[\phi]) + \gamma_{e,w}^{p,t} + M * (1 - f_{e,w}^{p,t}) \ \forall e \in E_a, p \in P, w \in W, t \in T \tag{8}$$

$$u_{p,w}^{t} \leq M * z_{p,w}^{t} \quad \forall p \in P, t \in T, w \in W \tag{9}$$

3) QKP storage constraints: Eq. (10) ensures that the stored keys in QKP are not less than 0. Eq. (11) expresses that the amount of keys stored in the QKP at stage t equals the remaining keys from the previous stage $t-1$, plus the newly generated keys supplied by the quantum channel, minus the keys consumed by all requests routed through path p. Eq. (12) ensures that the variable $\gamma_{e,w}^{p,t}$ equals 1 only when QKP is being used for path $p \in P$.

$$g_{p}^{t} \geq 0 \quad \forall p \in E_a, t \in T \tag{10}$$

$$g_{p}^{t} \leq g_{p}^{t-1} + \sum_{w \in W} u_{p,w}^{t} - \sum_{p' \in E_a} \sum_{w \in W} (\gamma_{p,w}^{p',t} + \gamma_{\bar{p},w}^{p',t}) - k_p * y_{p}^{t} \ \forall p \in P, t \in T \tag{11}$$

$$\gamma_{e,w}^{p,t} \leq M * q_{e,w}^{p,t} \quad \forall p \in P, e \in E_a, t \in T, w \in W \tag{12}$$

4) NAR constraints: Eq. (13) ensures that the variable $xx_{e'}^{p,t}$ equals to 1 when path p uses physical link e'. Eq. (14) defines if route ϕ is used on timeslot t. Eq.(15) ensures route ϕ has been used for request d at timeslot t. Eq.(16) calculates NAR at each timeslot t.

$$xx_{e'}^{p,t} \geq x_{e,w}^{p,t,\phi} \cdot h_{\phi,e'} \forall t \in T, w \in W, p \in P, e \in E_a, e' \in E_p, \phi \in \Phi_e \tag{13}$$

$$B_{\phi}^{t} \geq x_{e,w}^{p,t,\phi} \ \ \forall \ t \ \in \ T, w \ \in \ W, p \ \in \ P, e \ \in \ E_a, \phi \ \in \ \Phi_e \tag{14}$$

$$C_{\phi}^{d,t} \geq \left(xx_{e'}^{d,t} \cdot h_{\phi,e'} \right) \ \wedge \ B_{\phi}^{t} \ \forall \ t \in T, d \in D, e' \in E_p, e \in E_a, \phi \in \Phi_e \tag{15}$$

$$maxNAR_t \geq \sum_{d \ \in \ D} C_{\phi}^{d,t} \ \forall \ t \ \in \ T, e \ \in \ E_a, \phi \ \in \ \Phi_e \tag{16}$$

V. Deep Reinforcement Learning for RWA-MAR

To address the RWA-MAR problem in QKD networks, we propose a four-phases decision-making framework: Initialize data, prepare the possible routing, decision-making, and network update. This framework integrates network state information—including QKP availability, resource usage, and routing constraints—into a unified model capable of capturing the complex interplay between security and reasonable resource efficiency. The model jointly processes both topological and flow-related features to determine optimal routing decisions that minimize the maxNAR under physical-layer attack scenarios.

A. DRL Elements for RWA-MAR

In the following, we first define DRL elements for the RWA-MAR problem. Then, a proximal policy optimization (PPO)-based model is presented for feature extraction. In general, the goal of PPO is to find an optimal policy that maximizes the expected reward at timestep t corresponding to the final state of the trajectory. A timestep refers to a single unit of time during which an agent interacts with the environment, while an episode is a sequence of timesteps. PPO is a policy-gradient reinforcement learning algorithm that improves training stability by restricting large policy updates through a clipped surrogate objective [33]. PPO is adopted in this work because it provides a good balance between training stability, sample efficiency, and implementation simplicity, which makes it suitable for complex routing decision problems such as RWA-MAR.

1) State s_t: The state s_t represents the environment status at timestep t, where each timestep corresponds to processing an incoming request. The state representation combines two-dimensional matrices and a one-dimensional vector to capture the current QKD network conditions. The matrices represent the keys stored in the QKPs and the routing candidate path features, while the vector aggregates network resource and request information, including the source and destination nodes, candidate routing paths, request completion status, and the usage of quantum modules and wavelengths on nodes and links. The routing candidate path features matrix summarizes precomputed properties of feasible paths, such as hop count, expected maxNAR contribution, and estimated module consumption. The QKP matrix is structured as an $n \times n$ matrix, where each entry (i, j) indicates the amount of key material available for communication between nodes i and j. Finally, these matrices and vector are flattened into a unified representation suitable as input to the learning agent.

2) Action a_t: The action a_t represents the weight assigned to each physical link in the topology, indicating its relative importance in routing decisions. These weights are used as link costs in the network graph, and the routing path for each request is determined by applying a shortest-path algorithm over the weighted topology. In this way, the agent influences the routing outcome by adjusting the link weights according to the current network state.

3) Reward Function: The reward function provides real-time feedback to the PPO agent based on its action a_t, guiding the agent toward routing decisions that reduce the potential impact of physical-layer attacks while maintaining balanced resource utilization. In our model, the reward consists of two components: the negative value of $maxNAR$ and a normalized module-balance score. The $maxNAR$ term encourages the agent to minimize the worst-case attack impact in the network. $reward = -maxNAR + score$. The module-balance score reflects the distribution of QKD module utilization

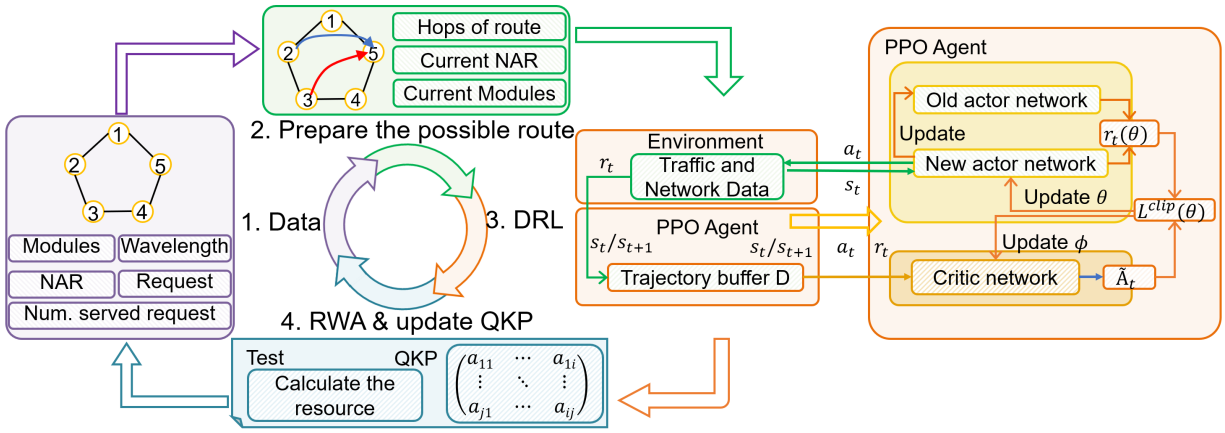


Fig. 3. Operating mechanism of our proposed model

across nodes and is computed from the normalized standard deviation of module usage. The first term drives the agent to minimize the worst-case attack impact, while the second term reflects the distribution of QKD module utilization across nodes, computed based on the normalized standard deviation of module usage. A higher balance score indicates a more even allocation of resources, preventing excessive concentration on specific nodes. Routing decisions are performed sequentially for incoming requests at each stage. After each decision for a single request, the network state and the corresponding reward are updated, enabling the agent to capture intermediate changes in resource utilization and routing distribution. These intermediate updates provide dense learning signals throughout the decision process, alleviating the reward sparsity issue associated with global worst-case metrics such as maxNAR and improving training efficiency. For the end of each episode, which all the requests are served well, the final reward will be calculated according to the situation of the network.

B. Four Phases of DRL Model

As illustrated in Fig. 3, the proposed PPO-based DRL model for QKD routing and resource allocation consists of four main phases.

1) Initialize data: The system initializes the network topology and associated resources, including node modules, wavelengths, current NAR, and the requests that need to be served, and the current number of served requests. These form the initial state for future decision-making.

2) Preparation of Candidate Routing Paths: For each unserved request, the algorithm first determines a set of feasible routing paths according to the current network state and available resources obtained from Phase 1. For each candidate path, key attributes such as hop count, NAR, and module consumption are pre-computed. These metrics convert raw network information into structured features that can be directly used by the learning-based decision process.

3) Decision making: In this phase, we use DRL with a Proximal Policy Optimization (PPO) agent. Since PPO alternates between sampling data through interaction with the environment and optimizing a surrogate objective function. PPO offers a favorable trade-off between sample efficiency and implementation simplicity [33], [34]. The environment operates on preprocessed data from the previous phase, where the QKP matrix is flattened and merged into a feature vector. These features are fed into a PPO agent employing an actor–critic architecture. The actor network generates a probability distribution over possible actions, from which an action is sampled, while the critic network evaluates the chosen actions. PPO employs the clipped surrogate objective function $L^{clip}(\theta)$ to constrain policy updates [33], thereby preventing large, destabilizing changes. Interaction data s_t, s_{t+1}, r_t, a_t is stored in the buffer $\mathcal{D}$ for training. The policy parameters θ and ϕ are updated via gradient ascent to maximize $L^{clip}(\theta)$, enabling the agent to iteratively refine its policy and optimize routing decisions in the QKD network.

4) RWA and update QKP: Based on the selected path, the routing and wavelength assignment (RWA) procedure is executed, and the required resources are allocated to serve the requests. After all requests are served, the remaining resources (e.g., modules and wavelengths) are used by the QKP mechanism to generate additional keys for future use. The updated information on key availability, QKP storage, and resource status is then fed back into the system, and the process proceeds to the next iteration starting from Phase 1.

C. Proposed PPO-Based DRL Algorithm

Algorithm 1 summarizes the overall DRL-based routing procedure. The algorithm first initializes the network topology and available resources, as described in Phase 1. It then prepares candidate routing paths for each unserved request according to the procedure in Phase 2, which provides the necessary input for the subsequent decision-making process. Then initializing the parameters of both the policy network π_θ and the value function network V_ϕ, as well as the experience buffer $\mathcal{D}$, which stores the data collected during agent-environment interactions. Additionally, the environment $\mathcal{E}$ is initialized, which defines the current situation of the QKD network, including the topology, requests etc.. (Lines 1-7).

For each episode, the policy π_θ is saved as $\pi_{\theta_{\text{old}}}$ to allow for importance sampling during the policy update step. At

each timestep, the agent selects routing paths a_t for each request based on the current state of the network s_t. After the agent selects the actions, the environment responds with the corresponding reward r_t. The tuple (s_t, a_t, r_t, s_{t+1}) is stored in the experience buffer $\mathcal{D}$. Once a trajectory is determined, we compute the advantage estimates $\hat{A}_t$ (Lines 8-15).

After collecting sufficient data, the algorithm runs policy update. For every state-action pair, the probability ratio $r_\theta(s_t, a_t)$ between the old policy $\pi_{\theta_{\text{old}}}$ and the new policy π_θ is computed. The surrogate objective $L(\theta)$ is then estimated for updating the hyperparameter ϵ and the policy parameters θ. After running several iterations of policy updates, two major outputs are produced: the final reward list that represents the summation of rewards over episodes and the final state of the topology and the optimized routing coupled with the status of the network after training is performed. (line 16-25).

Algorithm 1 DRL Algorithm

1: **Input:** Environment $\mathcal{E}$, Policy $\pi_{\theta_{\text{old}}}$, Hyperparameters: ϵ, α, topology, requests
2: **Output:** Final Reward List, Final Topology State
3: Initialize data for topology, resources, etc. as phase 1
4: According to the requests, prepare the routing and information for the agent, as phase 2
5: Initialize policy parameters θ, value function parameters ϕ
6: Initialize experience buffer with empty trajectory $\mathcal{D}$
7: Initialize the environment $\mathcal{E}$
8: **for** each episode k **do**
9: Collect data from the current policy: (s_t, a_t, r_t, s_{t+1})
10: $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$, Save the current policy for importance sampling
11: $\pi_{\theta_{\text{old}}}$ and V_ϕ compute action probabilities and values at each episode.
12: **for** each timestep t in the trajectory **do**
13: Calculate and store the data (s_t, a_t, r_t, s_{t+1}) in the experience buffer $\mathcal{D}$.
14: Calculate advantages $\hat{A}_t$.
15: **end for**
16: Store trajectory in $\mathcal{D}$
17: Compute the probability ratio $r_\theta(s_t, a_t)$
18: Compute the surrogate objective: $L^{clip}(\theta) =$
19: $\hat{\mathbb{E}}_t \left[\min \left(r_\theta(s_t, a_t)\hat{A}_t, \text{clip}(r_\theta(s_t, a_t), 1-\epsilon, 1+\epsilon)\hat{A}_t \right) \right]$
20: Update the policy parameters θ by maximizing the objective
21: Optionally update value function parameters ϕ using the value loss
22: **end for**
23: Pick Routing and resource assignment for requests
24: Update QKP, and process QKP data as phase 4.
25: **Return:** Final Reward, Final topology situation.

VI. Illustrative Numerical Results

In this section, we first outline the simulation setup and benchmark methods. We then validate the performance of the proposed approach by comparing it to the ILP model on a small topology. We finally evaluate the performance of the proposed approach on two large topologies.

A. Simulation Setup

We evaluate the performance of the DRL algorithm under three different architectures (*OBTR*, *OB*, and *TR*).

1) Performance matrix: The performance of the proposed DRL algorithm is assessed using three metrics: the maxNAR, the average NAR (avgNAR), and the average number of QKD modules consumed per node. Results are compared against two benchmarks: (i) our previously proposed Tabu-search-based heuristic [15], [35], and (ii) a baseline method that applies depth-first search to compute shortest-path routing between node pairs. The heuristic incorporates a tunable priority parameter α [15], which controls the preference between optical bypass and trusted relay during initialization. Lower values of α favor OB to reduce module consumption at the cost of slightly higher maxNAR, while higher values emphasize TR to minimize maxNAR with increased resource usage. Note that α only affects the heuristic; subsequent routing optimization is performed independently by the Tabu-search procedure.

2) Training & Generalization: The DRL model is initially trained on the NSF topology under the OBTR architecture, followed by testing over 100 episodes. The ILP results were obtained on a machine equipped with an Intel Core i7-9700 CPU (8 cores, 3.0 GHz) with 32G RAM using the CPLEX solver. The same machine was used to measure the execution time of the trained DRL model, which requires approximately one to two second to produce routing decisions in PoliQi topology. The DRL training process was performed on a Linux workstation equipped with an Intel Core i9-14900KF processor ((16 cores, up to 6.0 GHz)) and 64 GB RAM running Ubuntu Linux. Training the DRL model for 8,000 iterations required approximately 11.5 hours. Once trained, the model can generate routing decisions within about one second. After training, the model is evaluated to assess its adaptability across different QKD network architectures, topologies, varying request loads, and timeslot conditions. The model is trained on the NSF topology [36] and then tested on multiple topologies, including the NSF topology itself, the larger Madrid topology, and the smaller PoliQi topology, in order to demonstrate the generalization capability of the proposed approach.

3) Topologies and Simulated Scenarios: (i) We first compare the DRL algorithm with ILP on a five-node ring topology (similar to the PoliQi QKD testbed in Milan [11]). In this topology, all nodes are equipped with 10 QKD modules, and each link supports 5 quantum channels. A total of 7 requests of 10 kb/s per timeslot are generated. (ii) We then evaluate DRL on the NSF topology [36], which consists of 14 nodes and 21 links, where each node is equipped with 70 QKD modules. In this topology, each link supports 100 quantum channels. Link distances are scaled to [5, 15] km to satisfy QKD reach constraints. ILP is not applied in this case due to scalability limitations. The number of requests varies from 10 to 145, corresponding to up to 80% of all node pairs. Among these requests, 80% demand 5–10 kb/s, while 20% require 15–25 kb/s. Each scenario is tested over six instances, and the results are averaged. (iii) Finally, we evaluate scalability using a real-world QKD topology, namely the Madrid quantum network topology [37]. In this work, we use the topology of its most recent iteration [38]. To ensure feasible QKD transmission distances, the topology is reduced to 20 nodes (Mqn-20) by removing five long-reach nodes and scaling link distances so that the maximum request length remains within 80 km. The Mqn-20 topology was provided by our collaborators at Universidad Politécnica de Madrid (UPM). While part of this topology has been reported in Ref. [38], the complete topology remains undisclosed. In this topology, each node is equipped with 150 QKD modules, and each link

supports 100 quantum channels. We evaluate 10–200 requests under the same demand distribution used for the NSF topology, and results are averaged over six instances.

B. Evaluations on PoliQi Topology

We first discuss results on the PoliQi topology. As shown in Fig. 4, the proposed DRL approach achieves the same (optimal) maxNAR as the ILP and the heuristic under OBTR and TR architectures, with a maxNAR of 2 (note that TR consumes more QKD modules than OBTR, and here $\alpha = 0$). As expected, the OB configuration results in the highest maxNAR. Notably, the ILP requires over ten hours to converge, whereas the DRL achieves the same performance in approximately one second (tested over 10 episodes), while the heuristic requires about five seconds.

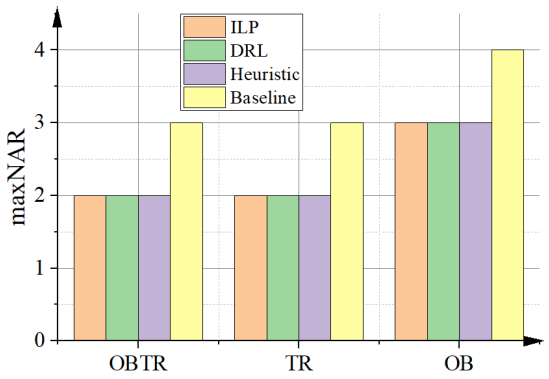


Fig. 4. Evaluation for three QKD architectures in PoliQi topology.

C. Evaluations on NSF Topology

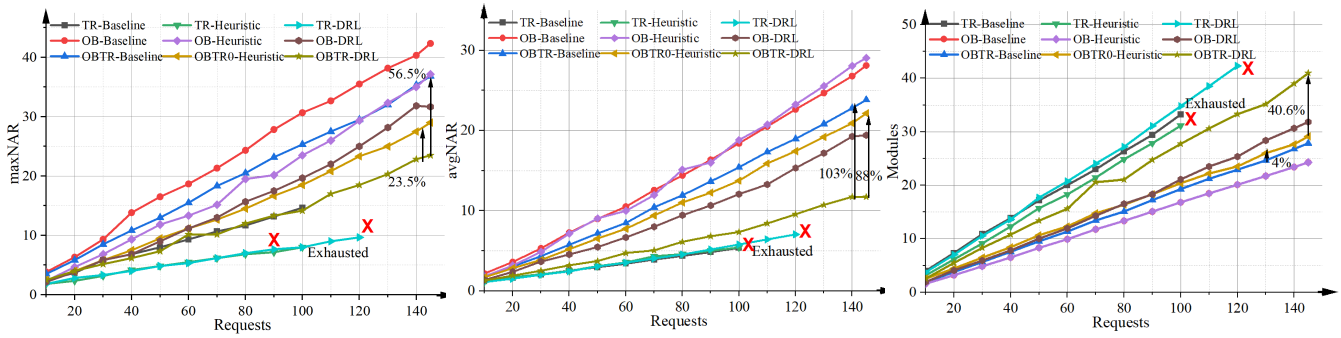
1) Evaluation of Attack Mitigation Performance for Three QKD Architectures: We next evaluate the DRL algorithm on the NSF topology, benchmarking it against a baseline that ignores maxNAR minimization. In Fig. 5(a), the DRL achieves a maxNAR reduction of 23.5% and 56.6% compared to the heuristic ($\alpha = 0$) and the baseline, respectively, under the OBTR configuration. The heuristic itself reduces maxNAR by approximately 27% relative to the baseline. Notably, our DRL method significantly enhances the performance of the TR architecture, sustaining up to 120 requests before resource exhaustion. In contrast, the heuristic and the baseline exhaust resources after only 100 requests due to higher module consumption. Besides, TR can achieve an 83% gap compared to the OBTR. The result of the OBTR architecture for DRL can achieve a similar result to the TR architecture of the baseline. As expected, OB yields the highest maxNAR, while TR achieves the lowest.

Fig. 5(b) presents the avgNAR, representing the average NAR across all paths and requests in the network. Compared to the maxNAR results, the gap between the DRL and the baseline increases significantly, reaching 103%, while the gap between the DRL and the heuristic is 88%. This demonstrates that the DRL approach is highly effective in distributing routing across the topology (lower the average NAR), thereby minimizing the concentration of traffic on vulnerable links and reducing the overall network exposure to attacks. It outperforms not only methods that ignore vulnerability reduction, but also heuristic-based approaches. This improvement in avgNAR arises because DRL attempts to decrease all the possible NAR across the whole topology, but even one single OB connection can considerably increase the maxNAR, as the signal propagates through shared optical channels. Moreover, the results demonstrate that DRL significantly decreases vulnerability across the entire network. The avgNAR of the OB architecture under DRL is lower than that of both the OB and OBTR architectures under the heuristic and baseline approaches. The gap between the OB and OBTR configurations under DRL is 65.8%. For all three architectures, the heuristic yields a slightly higher avgNAR than the baseline. This is because, although the heuristic reduces maxNAR by distributing requests, it leads to a slight increase in the average number of affected requests. Among all architectures, TR consistently yields the lowest avgNAR, while OB produces the highest, as expected.

Finally, Fig. 5(c) shows the average module utilization under different network architectures. As expected, the TR uses more modules than all other architectures, while OB uses the least number of modules, which is two modules per lightpath. Additionally, it is expected that both heuristics and baselines use equal modules in the case of OB, as they exhibit similar routing performances. The DRL approach, however, uses 40.6% and 47.4% more modules than the heuristic and baseline approaches, respectively. This is because the DRL has a tendency to set up additional lightpaths and more TR path in order to come up with more flexible and optimized routing decisions. Under the TR architecture, DRL consumes over 60% of the available modules on average, and some nodes reach full module utilization. This higher resource usage enables DRL to accommodate more requests than the heuristic method in TR architecture. In the OBTR architecture, the heuristic consumes 4% more modules than the baseline. Overall, these results highlight the trade-off between module usage and routing performance, where increased resource consumption allows for improved maxNAR and avgNAR through greater routing flexibility.

2) Evaluation on Multiple Stages: Then, we evaluate the impact of QKPs over five timeslots using the DRL model, the heuristic with OBTR0 configuration, and the baseline in the NSF topology with 145 requests. As shown in Fig. 6, all three methods begin with relatively high maxNAR values in the first timeslot; however, DRL achieves a reduction compared to the baseline value from 37 to 27, while the heuristic achieves a maxNAR of 30. In terms of avgNAR, DRL achieves reductions of 59% and 72.4% compared to the heuristic and baseline, respectively. In contrast, the difference in avgNAR between the heuristic and baseline is relatively minor. The effect of key caching becomes evident in the second timeslot, where NAR values sharply decline. In the third and fourth timeslots, maxNAR stabilizes around 1–2, as the QKP reserves become sufficient to meet demand. In the fifth timeslot, a slight increase in maxNAR is observed due to near depletion of stored keys. However, DRL continues to outperform the other approaches, demonstrating its effectiveness in dynamic key management and allocation under varying demand conditions.

This article has been accepted for publication in IEEE Transactions on Quantum Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TQE.2026.3684081



(a) maxNAR (b) avgNAR (c) average number of used QKD modules per node

Fig. 5. Evaluation of attack mitigation performance for three QKD architectures in NSF topology.

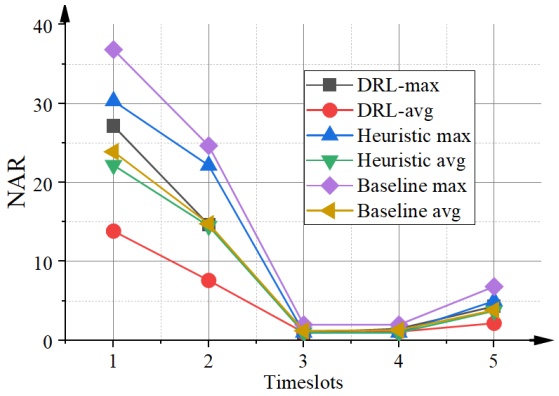


Fig. 6. Evaluation of attack mitigation performance for five stages.

3) Attack mitigation performance of DRL vs. heuristic for varying α: Finally, we analyze the impact of the high-priority parameter α on maxNAR, avgNAR, and module utilization, as illustrated in Fig. 7. In the heuristic approach, OBTR80 corresponds to $\alpha = 80$, indicating an 80% preference for initially selecting TR paths, while OBTR0 fully prioritizes OB paths. We compare our DRL model against the OBTR80 heuristic configuration and observe that DRL achieves superior results even when compared to the TR-biased heuristic. OBTR80 represents the best heuristic configuration identified in our experiments, as larger values of α (e.g., $\alpha = 90$) fails to produce feasible solutions for all requests.

As illustrated in Fig. 7(a), DRL under the OBTR setting achieves maxNAR reductions of 9.7%, 23%, and 56.6% compared to OBTR80, OBTR0, and the baseline, respectively. Within the heuristic results, OBTR80 reduces maxNAR by 23% compared to OBTR0. While our DRL is much faster than the heuristic, DRL method requires less than two minutes to compute routing decisions, while the heuristic costs more than 10 minutes to serve 145 requests.

For avgNAR, shown in Fig. 7(b), DRL achieves improvements of 47%, 88%, and 103% compared to OBTR80, OBTR0, and the baseline, respectively. OBTR80 also outperforms OBTR0 by 36.6%. These results demonstrate that DRL not only achieves lower maxNAR but also delivers substantially better avgNAR performance, with the performance gap in avgNAR exceeding that of maxNAR in many cases.

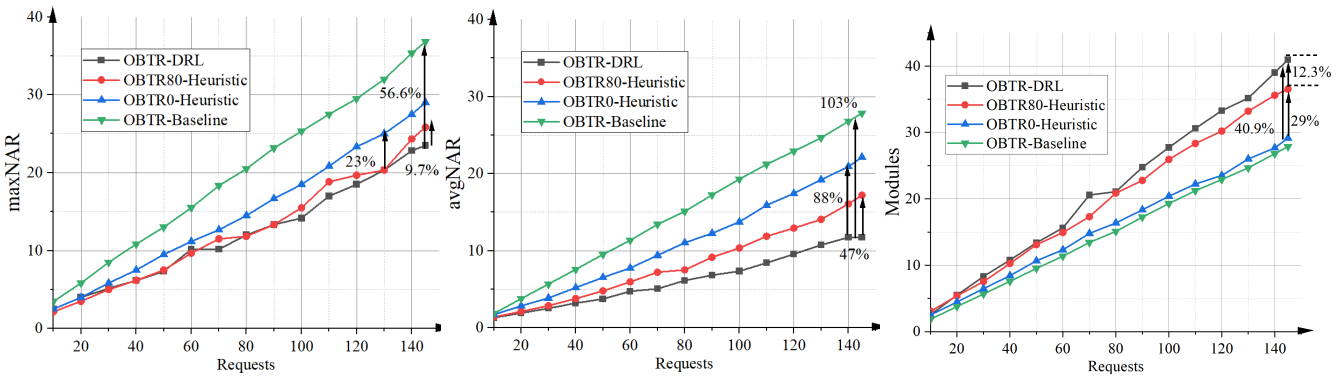
Module utilization results are shown in Fig. 7(c). As expected, DRL consumes more modules due to its more aggressive lightpath allocation strategy, highlighting the trade-off between minimizing maxNAR and resource consumption. Specifically, DRL uses 12.3%, 40.9%, and 47.5% more modules than OBTR80, OBTR0, and the baseline, respectively. Meanwhile, OBTR80 consumes 29% more modules than OBTR0, while OBTR0 increases module utilization by only 7% compared with the baseline, yet still achieves noticeable improvements in both maxNAR and avgNAR. Overall, these results indicate that the proposed DRL-based approach consistently outperforms the heuristic, even when compared with the high-α OBTR80 configuration, demonstrating its ability to achieve a better balance between attack mitigation and resource efficiency through adaptive learning.

D. Evaluations on Mqn-20 topology

The Mqn-20 topology introduces additional challenges compared to the NSF topology due to its highly non-uniform structure. It has fiber links with diverse lengths and attenuation levels, leading to a heterogeneous network environment. The topology combines ring-based local clusters with long-haul connections, where a small number of central nodes function as important transit routers. These nodes create structural bottlenecks with restricted routing options and are particularly susceptible to resource depletion, especially in terms of available QKD modules. Even when the module capacity per node is increased, the network remains vulnerable to request blocking under heavy traffic conditions. In terms of computing performance, the heuristic algorithm takes around 40–60 minutes to perform the entire set of requests on the given topology, whereas the proposed algorithm is capable of computing the reliable DRL solution in around 4 minutes, indicating a significant improvement in the computing performance of the algorithm.

To ensure clarity and focus in visual comparisons, we include only the OBTR results for the heuristic and baseline methods in the figures, as the trends under OB and TR architectures are consistent with those observed in the NSF topology. Our analysis here prioritizes the OBTR setting, which represents a more challenging and realistic deployment

This article has been accepted for publication in IEEE Transactions on Quantum Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TQE.2026.3684081



(a) maxNAR (b) avgNAR (c) average number of used QKD modules per node

Fig. 7. Attack mitigation performance of the proposed solution vs. the heuristic for different α values in NSF topology

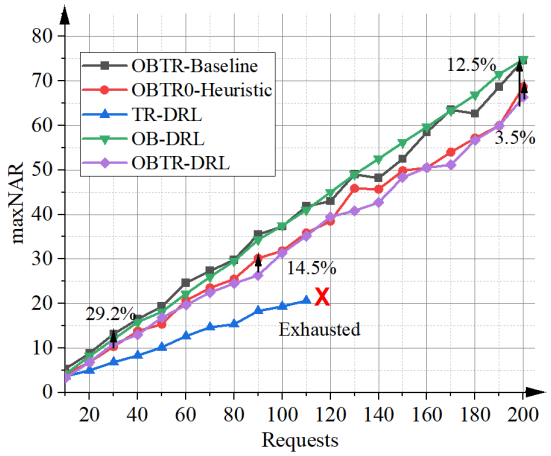


Fig. 8. Evaluation of maxNAR for Mqn-20 topology

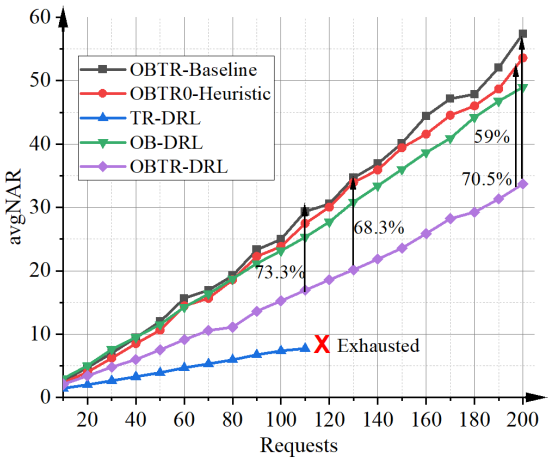


Fig. 9. Evaluation of avgNAR for Mqn-20 topology

scenario. Figure 8 presents the results for the maxNAR metric on the Mqn-20 topology. Due to the existence of bottlenecks, reducing the maxNAR proves especially difficult—many requests are forced to traverse the same critical node, leaving little opportunity for path diversity. While the gap between our DRL approach and the heuristic is narrower than in the NSF case, DRL still outperforms both the baseline and the heuristic. Specifically, DRL achieves up to a 29.2% improvement over the baseline and maintains a 12.5% gap at the 200-request level. Compared to the heuristic, DRL achieves a maximum 14.5% improvement and a 3.5% gain at the 200-request point. Notably, the TR-based architecture fails after 110 requests due to module exhaustion on bottleneck nodes. OB-only routing, and yields similar maxNAR values to the baseline.

Figure 9 shows the avgNAR performance for the same topology. While the improvements of maxNAR are constrained due to inherent bottlenecks in topological constraints, the metric avgNAR demonstrates evident advantages. Because avgNAR is less sensitive to some bottleneck links and better reflects the overall distribution of traffic, it can provide a clearer view of load balancing across the network. Based on the metric, the proposed model of DRL consistently works well and improves up to 73.3% compared to the baseline and achieves a 70.5% reduction when the number of requests reaches 200. Comparing it with the heuristic approach, the DRL model can achieve up to 68.3%, with a 59.0% reduction at 200 requests. These results give evidence that the proposed DRL approach is still effective in more complex, metro-scale QKD topologies. While structural bottlenecks bound the maximum reduction of maxNAR, the DRL method remains superior compared to both the baseline and heuristic solutions. The further reductions of avgNAR emphasize the robustness and scalability of the proposed approach in larger networks.

VII. Conclusion

This paper investigates the vulnerability of QKD networks to physical-layer attacks and introduces the RWA-MAR problem. To capture the worst-case impact of a single physical-layer attack, we develop an ILP formulation together with the maxNAR metric. In addition, we propose a DRL-based approach that can operate under three network architectures: OB, TR, and OB-TR. Simulation results demonstrate that the proposed approach reduces maxNAR by up to 56.6% compared with a baseline method that does not consider attack mitigation. To the best of our knowledge, this work is the

first to explicitly model and optimize maxNAR as a design objective for improving the resilience of QKD networks.

REFERENCES

[1] Y.-A. Chen, Q. Zhang, T.-Y. Chen, W.-Q. Cai, S.-K. Liao, J. Zhang, K. Chen, J. Yin, J.-G. Ren, Z. Chen, S.-L. Han, Q. Yu, K. Liang, F. Zhou, X. Yuan, M.-S. Zhao, T.-Y. Wang, X. Jiang, L. Zhang, W.-Y. Liu, Y. Li, Q. Shen, Y. Cao, C.-Y. Lu, R. Shu, J.-Y. Wang, L. Li, N.-L. Liu, F. Xu, X.-B. Wang, C.-Z. Peng, and J.-W. Pan, "An integrated space-to-ground quantum communication network over 4,600 kilometres," *Nature*, vol. 589, no. 7841, pp. 214–219, 2021.

[2] J. Li, P. Zheng, Z. Li, Y. Yang, N. Yu, Q. Sun, and J. Lu, "Decentralized key management and service in quantum key distribution networks: An experimental implementation," *IEEE Journal on Selected Areas in Communications*, 2025.

[3] Y. Cao, Y. Zhao, Q. Wang, J. Zhang, S. X. Ng, and L. Hanzo, "The evolution of quantum key distribution networks: On the road to the qinternet," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 839–894, 2022.

[4] W. Kozlowski, F. A. Kuipers, R. Smets, and B. Turkovic, "Quip: A p4 quantum internet protocol prototyping framework," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 7, pp. 1936–1949, 2024.

[5] L. Ruiz and J. C. Garcia-Escartin, "Routing and wavelength assignment in hybrid networks with classical and quantum signals," *IEEE Journal on Selected Areas in Communications*, vol. 43, no. 2, pp. 412–421, 2025.

[6] M. Furdek and N. Skorin-Kapov, "Physical-layer attacks in all-optical wdm networks," in *2011 Proceedings of the 34th International Convention MIPRO*. IEEE, 2011, pp. 446–451.

[7] P. Smith, D. Marangon, M. Lucamarini, Z. Yuan, and A. J. Shields, "Out-of-band electromagnetic injection attack on a quantum random number generator," *Physical Review Applied*, vol. 15, no. 4, p. 044044, 2021.

[8] A. Alomari and S. A. Kumar, "Securing iot systems in a post-quantum environment: Vulnerabilities, attacks, and possible solutions," *Internet of Things*, vol. 25, p. 101132, 2024.

[9] R. Renner and R. Wolf, "The debate over qkd: A rebuttal to the nsa's objections," 2023. [Online]. Available: https://arxiv.org/abs/2307.15116

[10] W. Beullens, "Breaking rainbow takes a weekend on a laptop," in *Annual International Cryptology Conference*. Springer, 2022, pp. 464–479.

[11] Q. Zhang, O. Ayoub, A. Gatto, J. Wu, F. Musumeci, and M. Tornatore, "Routing, channel, key-rate, and time-slot assignment for qkd in optical networks," *IEEE Transactions on Network and Service Management*, vol. 21, no. 1, pp. 148–160, 2023.

[12] Q. Zhang, N. Di Cicco, M. Ibrahimi, R. C. Almeida, A. Gatto, R. Boutaba, and M. Tornatore, "Link configuration for fidelity-constrained entanglement routing in quantum networks," in *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM)*. Vancouver, Canada: IEEE, May 2025, pp. 1–10.

[13] O. Amer, W. O. Krawec, and B. Wang, "Efficient routing for quantum key distribution networks," in *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)*. IEEE, 2020, pp. 137–147.

[14] N. Skorin-Kapov, J. Chen, and L. Wosinska, "A new approach to optical networks security: Attack-aware routing and wavelength assignment," *IEEE/ACM transactions on networking*, vol. 18, no. 3, pp. 750–760, 2009.

[15] M. Li, Q. Zhang, Z. Yang, S. Bregni, A. Gatto, R. Boutaba, and M. Tornatore, "Routing and wavelength assignment with minimal attack radius for qkd networks," in *Proc. IEEE GLOBECOM*, 2025, accepted for publication arxiv:https://arxiv.org/abs/2508.10613.

[16] W. Li, L. Zhang, H. Tan, Y. Lu, S.-K. Liao, J. Huang, H. Li, Z. Wang, H.-K. Mao, B. Yan, Q. Li, Y. Liu, Q. Zhang, C.-Z. Peng, L. You, F. Xu, and J.-W. Pan, "High-rate quantum key distribution exceeding 110 mb s–1," *Nature photonics*, vol. 17, no. 5, pp. 416–421, 2023.

[17] M. Zhao, R. Yuan, C. Feng, S. Han, and J. Cheng, "Security of coherent-state quantum key distribution using displacement receiver," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 7, pp. 1871–1884, 2024.

[18] M. Pereira, G. Currás-Lorenzo, Á. Navarrete, A. Mizutani, G. Kato, M. Curty, and K. Tamaki, "Modified bb84 quantum key distribution protocol robust to source imperfections," *Physical Review Research*, vol. 5, no. 2, p. 023065, 2023.

[19] B. Huttner, R. Alléaume, E. Diamanti, F. Fröwis, P. Grangier, H. Hübel, V. Martin, A. Poppe, J. A. Slater, T. Spiller, W. Tittel, B. Tranier, A. Wonfor, and H. Zbinden, "Long-range qkd without trusted nodes is not possible with current technology," *npj Quantum Information*, vol. 8, no. 1, p. 108, 2022.

[20] L.-Q. Chen, J.-Q. Chen, Q.-Y. Chen, and Y.-L. Zhao, "A quantum key distribution routing scheme for hybrid-trusted qkd network system," *Quantum Information Processing*, vol. 22, no. 1, p. 75, 2023.

[21] X. Yu, Y. Liu, X. Zou, Y. Cao, Y. Zhao, A. Nag, and J. Zhang, "Secret-key provisioning with collaborative routing in partially-trusted-relay-based quantum-key-distribution-secured optical networks," *Journal of Lightwave Technology*, vol. 40, no. 12, pp. 3530–3545, 2022.

[22] K. Dong, Y. Zhao, X. Yu, A. Nag, and J. Zhang, "Auxiliary graph based routing, wavelength, and time-slot assignment in metro quantum optical networks with a novel node structure," *Optics express*, vol. 28, no. 5, pp. 5936–5952, 2020.

[23] W. Sun, L.-J. Wang, X.-X. Sun, Y. Mao, H.-L. Yin, B.-X. Wang, T.-Y. Chen, and J.-W. Pan, "Experimental integration of quantum key distribution and gigabit-capable passive optical network," *Journal of Applied Physics*, vol. 123, no. 4, 2018.

[24] X. Yu, X. Ning, Q. Zhu, J. Lv, Y. Zhao, H. Zhang, and J. Zhang, "Multi-dimensional routing, wavelength, and timeslot allocation (rwta) in quantum key distribution optical networks (qkd-on)," *Applied Sciences*, vol. 11, no. 1, p. 348, 2020.

[25] P. Sharma, S. Gupta, V. Bhatia, and S. Prakash, "Deep reinforcement learning-based routing and resource assignment in quantum key distribution-secured optical networks," *IET Quantum Communication*, vol. 4, no. 3, pp. 136–145, 2023.

[26] S. D. Reiß and P. van Loock, "Deep reinforcement learning for key distribution based on quantum repeaters," *Physical Review A*, vol. 108, no. 1, p. 012406, 2023.

[27] S. Ding, Y. Cheng, and C. C.-K. Chan, "Drl-assisted quantum attack mitigation in resource allocation of cv-qkd over optical networks," *Journal of Optical Communications and Networking*, vol. 17, no. 4, pp. 262–274, 2025.

[28] Y. Seok, J.-B. Kim, Y.-H. Han, H.-K. Lim, C. Lee, and W. Lee, "Deep reinforcement learning-driven optimization of end-to-end key provision in qkd systems," *Journal of Network and Systems Management*, vol. 33, no. 2, pp. 1–32, 2025.

[29] V. Mani, "Security challenges to iot and cloud-based systems in the era of quantum attacks," in *Communication Technologies and Security Challenges in IoT: Present and Future*. Springer, 2024, pp. 227–239.

[30] P. Sharma, V. Bhatia, and S. Prakash, "Routing based on deep reinforcement learning in quantum key distribution-secured optical networks," in *Proceedings of the IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*. New Delhi, India: IEEE, Dec 2023, pp. 1–5.

[31] Z. Yang, A. Ghubaish, R. Jain, A. AlFuqaha, A. Erbad, R. Kompella, H. Shapourian, and R. Nejabati, "Layer-wise security framework and analysis for the quantum internet," *IEEE Journal on Selected Areas in Communications*, pp. 1–1, 2025.

[32] M. Li, Q. Zhang, A. Gatto, S. Bregni, G. Verticale, and M. Tornatore, "Drl-based progressive recovery for quantum-key-distribution networks," *Journal of Optical Communications and Networking*, vol. 16, no. 9, pp. E36–E47, 2024.

[33] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[34] Y. Gu, Y. Cheng, C. P. Chen, and X. Wang, "Proximal policy optimization with policy feedback," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 7, pp. 4600–4610, 2021.

[35] F. Glover, "Tabu search: A tutorial," *Interfaces*, vol. 20, no. 4, pp. 74–94, 1990.

[36] X. Dong, T. E. El-Gorashi, and J. M. Elmirghani, "On the energy efficiency of physical topology design for ip over wdm networks," *Journal of Lightwave Technology*, vol. 30, no. 12, pp. 1931–1942, 2012.

[37] V. Martin, J. P. Brito, L. Ortíz, R. B. Méndez, J. S. Buruaga, R. J. Vicente, A. Sebastián-Lombraña, D. Rincón, F. Pérez, C. Sánchez, M. Peev, H. H. Brunner, F. Fung, A. Poppe, F. Fröwis, A. J. Shields, R. I. Woodward, H. Griesser, S. Roehrich, F. de la Iglesia, C. Abellán, M. Hentschel, J. M. Rivas-Moscoso, A. Pastor-Perales, J. Folgueira, and D. López, "Madqci: a heterogeneous and scalable sdn-qkd network deployed in production facilities," *npj Quantum Information*, vol. 10, no. 1, p. 80, 2024.

[38] A. Sebastián-Lombraña, L. Ortiz, J. P. Brito, J. Faba, R. B. Méndez, J. S. de Buruaga, R. J. Vicente, J. Setien, J. J. Romero, C. Escribano, P. Salas, J. L. Bejarano, and V. Martin, "Advancing the future of quantum communication networks: The new madqci," in *Proceedings of the International Conference on Transparent Optical Networks (ICTON)*. Bari, Italy: IEEE, Jul 2025.